SESSION

PERFORMANCE ANALYSIS AND EVALUATION

Chair(s)

TBA

Computing Performance Bounds for Analysis of Indoor Wireless Fading Channels for RFID-Based Localization

F. Schwaner[1], **D. Blackmer**[1] **and A. Abedi**[1]

[1]Department of Electrical Engineering, University of Maine, Orono, ME 04473

Abstract—*This article considers the varying K-factor of Rician channels and attempts to derive an upper bound on the K-factor of an indoor Radio Frequency Identification Tag channel based solely on the physical shape of the area in which experiments are conducted. The K-factor is found to be highly dependent upon the physical size and shape of the area in which the channel exists. Ray-tracing is used in an empty hallway to estimate the K-factor for distances ranging from 1 to 10 feet. It is assumed that the channel is flat-fading with additive white Gaussian noise. Experimental data from radio frequency identification tags are collected to verify the theoretical model.*

Keywords: Radio Frequency Identification Tag, Rayleigh Fading, Rician Fading, Channel Model, Indoor Channel Model

1. Introduction

Wireless communications have become an integral piece of everyday living. They are used primarily for the convenience of not having to be tethered to the entity with which you are trying to communicate. However, this characteristic creates challenges and sources of error not seen in the wired communications. One of the most notable phenomena in wireless channels is multipath fading, where the signal being sent travels not only along the line-of-sight (LOS) path, but also reflects off objects and travels to the destination. There is much work being done trying to accurately model Rician Channels as they are the most common model for fading. Channels have been modeled in both static systems [1, 2, 3] and mobile systems [4, 5]. With such systems, the multipath fading affects the channel by degrading signal quality in a random fashion. When this occurs, the conventional methods of sending data are not optimal. However data transmission can be optimized based on characteristics of the channel. One of the parameters that is considered for optimization of a channel is the K-factor, which is the ratio of energy delivered along direct paths, E_D, to the energy delivered along scattered paths, E_S [6, 7]. Some of these schemes will be explained soon.

$$K = \frac{E_d}{E_s} \tag{1}$$

Because of the potential for optimization, there have been several research works estimating the K-factor of a channel [8, 9, 10]. Most of these works estimating the K-factor have been done by collecting empirical data and using that to find K [8, 9]. This tactic works well if collecting data from the channel is possible before we need to characterize it. Large outdoor systems K-factors may be too difficult or complex to estimate without data. However, the use of reverberation chambers to emulate Rician channels with controllable K-factors is a recent advancement in channel testing [10, 11, 12]. A model for predicting the K-factor given the variance of the scattered paths is presented in [10].

There are many advantages that come from being able to estimate the K-factor of a wireless channel. In a multiple-input, multiple output (MIMO) orthogonal frequency division multiplexing (OFDM) channel, the K-factor can be exploited to maximize data throughput [13]. By considering both the K-factor and the signal-to-noise ratio (SNR) of the MIMO-OFDM channel, data throughput was increased to 80 Mbps, twice the throughput of the scheme that only considered SNR. Another method of exploiting the K-factor for channel optimization is to use an adaptive transmission scheme that changes modulation type based on the channels current K-factor and SNR [14]. In such a system, data transmission could be increased by 58% in an enhanced general-purpose radio service (EGPRS) system.

In this paper, we attempt to establish an upper bound for the K-factor in a hallway using only the physical shape of the hallway. Ray-tracing is used to determine the energy delivered along the 10 most dominant paths. Several assumptions are made about the channel, including the assumption that the walls are perfectly reflective non scattering surfaces, meaning the angle of reflection is the same as the angle of incidence and no energy is lost in the reflection of the wall.

This project was motivated as part of an indoor localization system designed for tracking and navigation in a close-quartered indoor environment. There have been many advancements in Radio Frequency Identification tag (RFID) technology that have tailored it for use in indoor applications such as: navigation, inventory tracking, and structural or environmental monitoring. However, due to reflection issues and attenuation problems, accurate determination of a RFID tag position indoors can be a challenging procedure. One tactic that can help is to characterize the channel by combining experimentally obtained data with a theoretical model. For this project, the number of successful RFID reads in a given period of time is used to approximate the SNR of the channel and the data is then matched to a Rayleigh and Rician fading models.

This article is structured as follows. First, a system model is described and the derivation of the bit error rate (BER) of a Rayleigh channel and Rician channel in additive white Gaussian noise (AWGN) is derived. Then, the determination of the probability of a successful RFID tag interrogation is explained and plotted. Finally, the experimental data and the theoretical models are fitted together and compared followed by a conclusion in Section 5.

2. System Models

In this section, a theoretical model for channel SNR is created in the following fashion. A generic system model is defined first. A Rayleigh variable is substituted and evaluated as detailed in [15]. Finally, a Rician model is used and its parameter is estimated.

2.1 Generic System Model

The simplest model to consider is the free space model. Coincidentally, this model applies to channels in an open-source environment where there are no surfaces to reflect signals. In this model, the energy received by the tag, E_R, decreases exponentially with distance D.

$$E_R = \frac{E_T}{4\pi D^2} \tag{2}$$

where E_T is the transmitted energy.

This works as one channel model, but one can also build other channel models by more properly defining the system in the form of:

$$\mathbf{y} = h\mathbf{x} + \mathbf{n} \tag{3}$$

where $\mathbf{n}$ is a vector of AWGN, $\mathbf{x}$ is a vector of the amplitude received through free-space loss, h is the channel charcteristic and $\mathbf{y}$ is the output vector of the channel. In terms of power, the SNR, λ can be thus represented as

$$\lambda = \frac{E_T|h|^2}{N_0} \tag{4}$$

where h can be represented by a random variable which can support various distributions. In the next two sections, Rayleigh and Rician fading models are presented in the context of the proposed modeling approach. This is accomplished by substituting a Rayleigh distributed and Rician distributed variable into h.

2.2 Rayleigh Fading Model

Rayleigh channels are typically a fair representation when there is no LOS component, while there is an abundance of indirect reflected paths. When a Rayleigh variable is substituted into h, the probability of bit error, when the phase can be coherently detected, is [15]

$$P_B = \frac{1-\sqrt{\frac{E_R}{E_R+N_0}}}{2} = \frac{1-\sqrt{\frac{\lambda}{\lambda+1}}}{2} \tag{5}$$

2.3 Rician Fading Model

A simple closed-form solution of BER in a Rician channel has yet to be found. However if the system, as described by (3) is experiencing Rician fading, $H{\sim}Rice(p|v,\sigma)$, which has a probability density function of

$$P_H(p) = \frac{p}{\sigma^2}e^{-\frac{p^2+v^2}{2\sigma^2}}I_0(\frac{vp}{\sigma^2}) \tag{6}$$

where v is the distance between the reference point and the center of the distribution, σ is a scaling factor and I_0 is a modified Bessel function of the first kind and zero-order.

$$I_0(\eta) = \frac{1}{\pi}\int_0^\pi e^{\eta cos(\theta)}d\theta \tag{7}$$

With this representation of the distribution, the K-factor is given by

$$K = \frac{v^2}{2\sigma^2}$$
$$K(in\ dB) = 10log_{10}\frac{v^2}{2\sigma^2}dB \qquad (8)$$

By setting the scaling factor σ to 1, the variable v can be described in terms of the K-factor by

$$v = \sqrt{2K} \qquad (9)$$

Energy in the system will then be a random variable of type $Rice^2(p|\sqrt{2K}, 1)$ which is of non-central χ^2 distribution with two degrees of freedom and non-centrality parameter $2K$. The energy in the system can then be represented as:

$$Y = \gamma X + N \qquad (10)$$

where γ is the non-central χ^2 variable representing the channels fading effect on the input energy X and N is the energy added by the Gaussian noise. The distribution of Y can then be represented as the convolution of both γX and N. A MATLAB script is then used to approximate the BER of the channel using

$$P_B = \int_{-\infty}^{0} \gamma(p)X * N(p)dp \qquad (11)$$

where $*$ denotes convolution. To enhance accuracy of (11), we need to find an upper limit for the Rician K-factor.

3. Estimating the Rician K-factor

In this section, we attempt to derive an upper-bound such that curves representing Rayleigh and Rician Channels completely sandwich the experimental data. The reason for this being that a Rayleigh Channel is a special case of Rician Channel, where the K-factor is 0. Thus providing a natural lower-bound. In determining the K-factor of the channel, we assume the signals are bouncing off a non-scattering surface, where the signal reflects off an object at the same angle as its angle of incidence. Also, we assume that no energy is lost during reflection. These assumptions can be relaxed later to enhance the approximation accuracy, but it is out of the scope of this work.

3.1 Experimental Setup

Tests are performed in an empty hallway which has a width of 6 feet (1.83 m). The tag is placed at the end of the hallway at the same height as the reader. The reader is placed 1 foot away from the tag and attempts to read the tag 100 times. The number of successful reads is noted. The reader is then moved back a foot and the process is reapeated. This is done for distances 1-10 feet. This entire sequence was then repeated several times to confirm consistency between the sets. The walls in this hallway were bare for this experiment. This was done to eliminate any extraneous sources of reflection. Special care was taken to ensure that no other devices that operate in the 915 MHz spectrum were present. This stipulation aids in supporting our assumption that the noise appears to be white and gaussian.

The hardware used in this project is: one ALN-9534 ultra-high frequency (UHF) RFID tag. This particular tag model was chosen for its consistent readability regardless of orientation and azimuthal angle relative to the reader. The interrogator was chosen to be a Motorola MC9090 as it met the requirements for the indoor tracking system under consideration in this project. This reader is capable of 1 Watt power transmission during its interrogation pulse eight times a second, making it capable of reading a group of tags eight times every second.

3.2 K-factor Derivation

Figure 1 displays a situation in which the tag is placed a distance d away from the reader and interrogated. The signals dominant path is the LOS path on which the signal travels directly to the tag and back again without reflecting off any walls or objects. The second most dominant path is the path in which the signal reflects off of one wall halfway between the reader and tag. This path is mirrored in our situation because of the fact that we are in a hallway and the signal can reflect off either the wall on the right or the wall on the left. This pattern continues as the signal reflects more frequently, the energy delivered to the reader along that path decreases. The amount of energy delivered to the reader along each path decides whether a path is assigned as a dominant path or a scattered path. A diagram depicting paths with 0, 1 and 2 reflections may be seen in Figure 1.

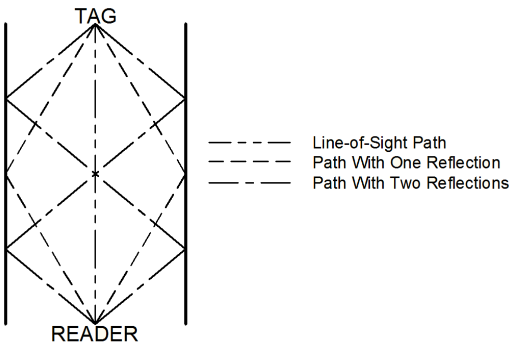


Fig. 1: Diagram of reflecting paths in a typical hallway.

The energy is computed for all paths with less than 10 reflections. Paths are assigned to be dominant paths if the power delivered by them is greater than or equal to 80% of the energy delivered by the LOS path. Paths delivering less power are assigned to be scattered paths and the quotient of these two sums is the K-factor of the channel. This value is a theoretical upper bound for the Rician K-factor as it does not include all the spurious signals that result from typical reflective surfaces. A plot of the derived values for the K-factor as distance increases can be seen in Figure 2.

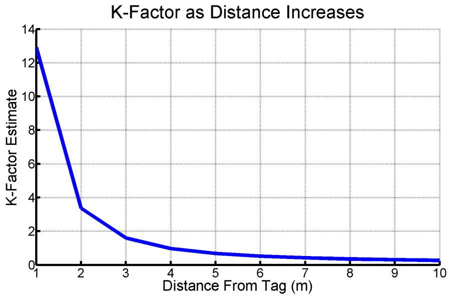


Fig. 2: Ratio of energy in dominant paths to scattered paths.

Computing the energy along a path with m reflections is performed using (2) where D is represented as (12).

$$D = \sqrt{d^2 + 1.83^2 m} \tag{12}$$

4. Probability of Packet Success As a Function of SNR

The probability of packet success from the SNR is easily determined using probability theory. The tags used transmit 96 bits of Binary-phase shift keying (BPSK) modulated signals. It is assumed that since there is no error-correcting codes that all 96 bits must be received correctly in order to read the tag. The probability of receiving one bit correctly, P_b can be computed using the Q-function as follows.

$$P_b = Q(\sqrt{\lambda}) \tag{13}$$

The probability of packet success (P) is just the probability of receiving all 96 bits correctly.

$$P = {P_b}^{96} \tag{14}$$

The experimental data obtained in this was then compared to the theoretical models derived in Section 2.

5. Concluding Remarks

Comparing the experimental data with the proposed theoretical models for Rayleigh and Rician Fading (Figure 3), verifies the appropriateness of the bounds and models for indoor scenarios.

The Rayleigh Channel provided a lower bound for the channel's probability of packet success whereas the Rician model provided an upper bound for the same performance metric. One can see by comparing Figures 2 and 3 that as the K-factor decreases, the Rician model approaches the Rayleigh model. This model can be used to establish an upper bound for the Rician K-factor of an environment. Using this bound as an approximation, one can more accurately estimate the K-factor of the channel. This allows the user of the channel to exploit the K-factor for maximum data transmission possible.

The results of this article demonstrates the potential for accurate channel estimation based solely on the physical parameters of the space used for transmission. This is a new result compared to the prior research on K-factor estimation in that it does not rely on collecting data from the channel. This can be very useful in cases where the channel cannot be tested prior to it being used. Tightening of the upper bound is a subject of future work as well as is the inclusion of scattered light from signal reflections.

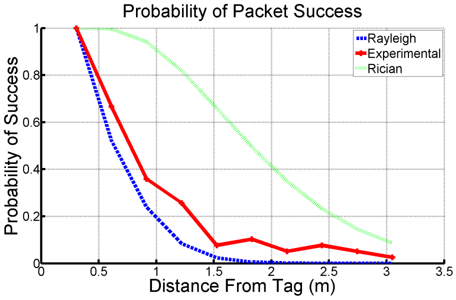


Fig. 3: Experimental data with upper and lower bound

References

[1] C. Xiao, Y. R. Zheng, and N. Beaulieu, "Novel sum-of-sinusoids simulation models for rayleigh and rician fading channels," *Wireless Communications, IEEE Transactions on*, vol. 5, no. 12, pp. 3667 –3679, December 2006.

[2] C. Xiao, Y. Zheng, and N. Beaulieu, "Statistical simulation models for rayleigh and rician fading," *Communications, 2003. ICC '03. IEEE International Conference on*, vol. 5, pp. 3524 – 3529, May 2003.

[3] J. Roberts and J. Abeysinghe, "A two-state rician model for predicting indoor wireless communication performance," *Communications, 1995. ICC '95 Seattle, 'Gateway to Globalization', 1995 IEEE International Conference on*, vol. 1, pp. 40 –43, June 1995.

[4] L.-C. Wang and Y.-H. Cheng, "A statistical mobile-to-mobile rician fading channel model," *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st*, vol. 1, pp. 63 – 67, May 2005.

[5] L.-C. Wang, W.-C. Liu, and Y.-H. Cheng, "Statistical analysis of a mobile-to-mobile rician fading channel model," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 1, pp. 32 –38, January 2009.

[6] H. Che, L. Ligthart, and R. Prasad, "Working conditions for an influencing factor on the performance of adaptive ofdm system under a varying channel," *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, vol. 2, pp. 1939 – 1943, September 2003.

[7] M. McKay and I. Collings, "Capacity bounds for correlated rician mimo channels," *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 2, pp. 772 – 776, May 2005.

[8] N. Shroff and K. Giridhar, "Biased estimation of rician k factor," *Information, Communications Signal Processing, 2007 6th International Conference on*, pp. 1 –5, December 2007.

[9] A. Doukas and G. Kalivas, "Rician k factor estimation for wireless communication systems," *Wireless and Mobile Communications, 2006. ICWMC '06. International Conference on*, pp. 69 –69, July 2006.

[10] C. Lemoine, E. Amador, and P. Besnier, "Improved estimation of the k-factor for rician channels emulation in a reverberation chamber," *Antennas and Propagation (EuCAP), 2010 Proceedings of the Fourth European Conference on*, pp. 1 –5, April 2010.

[11] A. Sorrentino, G. Ferrara, and M. Migliaccio, "The reverberating chamber as a line-of-sight wireless channel emulator," *Antennas and Propagation, IEEE Transactions on*, vol. 56, no. 6, pp. 1825 –1830, June 2008.

[12] C. Holloway, D. Hill, J. Ladbury, P. Wilson, G. Koepke, and J. Coder, "On the use of reverberation chambers to simulate a rician radio environment for the testing of wireless devices," *Antennas and Propagation, IEEE Transactions on*, vol. 54, no. 11, pp. 3167 –3177, November 2006.

[13] K.-Y. Lin, R.-T. Juang, H.-P. Lin, W.-J. Shyu, and P. Ting, "Link adaptation of mimo-ofdm transmission exploiting the rician channel k-factor," pp. 184 –188, July 2009.

[14] R.-T. Juang, H.-P. Lin, and M.-J. Tseng, "Adaptive transmission scheme based on rician channel k-factor for egprs systems," *Autonomic and Autonomous Systems and International Conference on Networking and Services, 2005. ICAS-ICNS 2005. Joint International Conference on*, pp. 10 –10, October 2005.

[15] A. Chandra, D. Biswas, and C. Bose, "Ber performance of coherent psk in rayleigh fading channel with imperfect phase estimation," pp. 130 –134, December 2010.

A study of the performance of vector based forwarding in underwater acoustic sensor network

Chetan Ragpot, Nassirah Laloo and Mohammad Sameer Sunhaloo
School of Innovative Technologies and Engineering
University of Technology Mauritius
La Tour Koenig, Mauritius.
E-mail: chetan.aea@gmail.com, n.laloo@utm.intnet.mu, sameer.s@utm.intnet.mu

Raja K. Subramanian
E-mail: rksuom@yahoo.co.in

Abstract — In this paper, we investigate the performance of the vector based forwarding protocol in shallow and deep water in terms of propagation delay and signal to noise ratio. We determine whether water is shallow or deep by varying the underwater propagation speed at different depth. We evaluate vector based forwarding in shallow and deep water through simulation using Aqua-Sim on ns-2. We have observed that vector based forwarding performs better in deep water than in shallow water.

Keywords: Underwater Acoustic Sensor Network, Vector Based Forwarding, Propagation Speed, Propagation Delay.

1.0 Introduction

Wireless Sensor Network (WSN) in aqueous medium also known as Underwater Acoustic Sensor Network (UASN) is distinctive due to its surrounding environment. This area of study is attracting the interest of many researchers and has enabled a broad range of applications including information collection, assisted monitoring, mine reconnaissance, equipment monitoring, disaster prevention, under ocean exploration and environmental monitoring [6].

WSN in aqueous medium has the ability to explore the underwater environment in details [6]. To ensure maximum efficiency, a good communication system as well as an effective routing protocol is needed. This will enable the underwater devices to communicate precisely. Underwater propagation speed varies with temperature, salinity and depth. By varying the underwater propagation speed at different depths, two scenarios may be examined accurately namely: shallow and deep water. In both shallow and deep water, different ambient noise levels and different spreading factors may be applied and analyzed to determine the efficiency of specific routing protocols.

This paper is organized as follows. In Section 2, we present a general 3D Architecture and a 3D architecture with Autonomous Underwater Vehicles (AUV). Vector based forwarding (VBF), which is an efficient and robust underwater routing protocol is described in Section 3 and in the same section we introduce our concept. Finally in Sections 4 to 6 we describe the implementation and results of the simulation performed followed by a conclusion in section 7.

2.0 Three-Dimensional Architecture

We consider a general 3D model and a 3D model with AUV [1]. In the general 3D model, the sensor nodes are anchored at different depths and are equipped with a floating buoy, which can be inflated by a pump. The sensors can be regulated by adjusting the length of the wire that connects the sensor to the anchor. In the other 3D model, which we have considered, some nodes act as AUV, which consist of underwater sensors.

3.0 Vector Based Forwarding

The VBF as proposed by Xie et al. [8] is considered as the base routing protocol for robust, scalable, and energy efficient routing in underwater acoustic network [5, 7].

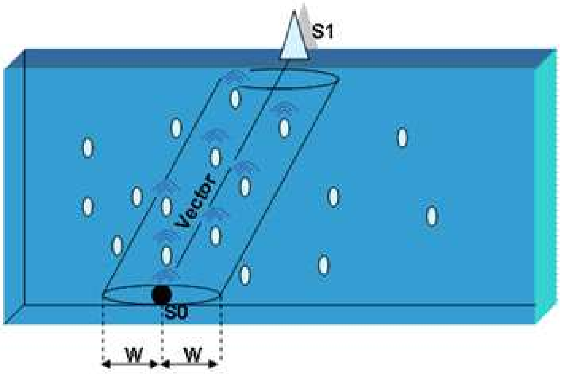


Figure 1: Vector based forwarding model.

In Figure 1, we illustrate the vector based forwarding protocol model. The sensor nodes are distributed in 3D. The nodes are equipped with devices that enable them to measure the distance and signal's angle of arrival.

Node S_0 is considered as the source and node S_1 is the sink. When S_0 wants to send data packets to sink S_1, it first establishes a routing vector $(\overrightarrow{S_0S_1})$ as shown in the figure. W is the threshold distance from the routing vector, which makes a cylinder pipe with central axis S_1S_0 and radius W. Node S_0 broadcasts the packet with S_1 as target. Upon receiving the packet, the nodes calculate their corresponding distance from the routing vector. If the node is within the range W from the routing vector, then the packet is forwarded to the next node, otherwise the packet is discarded.

In VBF, participating nodes are mainly those which are within the routing pipe depicted in Figure 1. However when the sensor nodes are densely deployed, VBF may involve too many nodes in data forwarding thus increasing energy consumption.

Effective node selection in VBF may be achieved by considering the self adaptation algorithm proposed by Xie in [7]. The most desirable nodes are selected as forwarders based on the value of $T_{\text{adaptation}}$ given by

$$T_{\text{adaptation}} = \sqrt{\alpha} \times T_{\text{delay}} + \frac{R-d}{v_0} \tag{1}$$

where α is the desirableness factor, R is the transmission range, d is the distance between the selected forwarder node and the next forwarder node, and v_0 is the propagation speed of acoustic signals in water. The purpose of the delay in (1) is to distinguish the importance of the nodes in the transmission range of a forwarder. In VBF, T_{delay} is set large enough.

VBF has been proposed as an efficient routing protocol for aqueous medium [8] and it has been used for protocols efficiency comparison [5, 7]. To the best of our knowledge, no study has been done to investigate the performance of VBF in shallow and deep waters. Also, as far as we are aware, VBF has been considered with an average speed of 1500 ms^{-1}.

In this paper we consider varying propagation speed based on the assumption that it will provide better accuracy [2]. The VBF routing protocol is considered with varying propagation speed in order to judge its performance in shallow and deep water. We consider the work of Coppens [3], where the ocean speed c is considered with varying pressure (depth), salinity and temperature according to the following equation:

$$\begin{aligned} c = {} & 1449.05 + 45.7t - 5.21t^2 + 0.23t^3 + (1.333 - 0.126t + 0.009t^2)(s-35) \\ & + (16.23 + 0.253t)z + (0.213 - 0.1t)z^2 + (0.016 + 0.0002(s-35))(s-35)tz, \end{aligned} \tag{2}$$

where z is the depth in meters, s is the salinity in parts per thousand and $t = T/10°C$. We note that (2) is valid for $0°C \leq T \leq 35°C$, $0 \leq S \leq 45$ parts per thousand, and $0 \leq z \leq 4000\,m$ where T is given by

$$T = \begin{cases} 2.0, & z > 4500; \\ 2.0 + 0.00057142(4500 - z), & 1000 < z < 4500; \\ 4.0 + 0.016(1000 - z), & z > 750; \\ 8.0 + 0.028(750 - z), & z > 250; \\ 22.0, & \text{otherwise.} \end{cases}$$

For further details about VBF, please refer to [7, 8].

4.0 Simulation

The simulations were built using the underwater package Aqua-Sim of ns-2. While supporting 3D deployment, Aqua-Sim allows effective simulation of acoustic signal attenuation and packet collisions in underwater sensor networks. Basic information about the ocean environment and the underwater channel are provided through an Otcl script. Modules uw-common, uw-mac and uw-routing, found in Aqua-Sim are modified to support the simulation.

The simulations described in this paper make use of the 3D network architecture with randomly deployed sensor nodes. The VBF routing protocol is used with one data source and one sink. The LinkQuest UWM 2000 [9] is taken as a reference for the sensor nodes with the parameters given in Table 1:

Bit Rate	10 $kbps$
Energy consumption for sending mode	2 W
Energy consumption for receiving mode	0.8 mW
Energy consumption for idle mode	0.2 mW

Table 1: Parameters for LinkQuest UWM 2000

The size of the data packet and large control packet for VBF is set to 50 Bytes. The size of the small control packet for VBF is set to 50 Bytes. The pipe radius in VBF is set to 20 m.

Performance Metrics

We now describe the performance metrics [4], which have been used evaluate the performance of VBF. Propagation delay is taken to be the total time delay in second to send a number of packets from the source to the destination through VBF routing protocol. Signal to Noise Ratio (SNR) is taken as the ratio of the total power transmitted and the total noise in the network to send a number of packets from the source to the destination through VBF.

5.0 Implementation

For our simulation, shallow water consists of depth less than 200 m and cylinder spreading. Deep water consists of depth greater or equal to 200 m and spherical spreading.

Shallow Water

For the case of shallow water, we consider a cube of length 100 m. Using Coppens equation with a temperature of 22°C, a salinity of 36.5 parts per thousand and a depth of 10-100 m, a range of 1526.99 to 1528.33 is obtained for the underwater propagation speed.

Now, we consider a source node, transmitting 50 packets to a AUV sink. Both the source and the sink are placed at the same level as shown in Figure 2 and they are tested by varying the depth by changing the locations of the source and the sink simultaneously. We increased the depth from 10 m to 100 m, thus causing transmission at different propagation speeds.

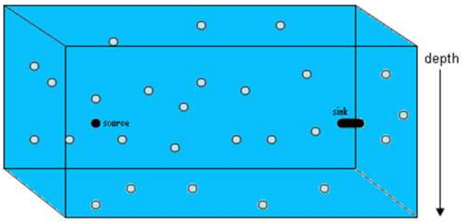


Figure 2: Source and sink are placed at the same level.

Next, we consider the case where the source is placed at the bottom and the sink is considered as a boat floating at the sea surface as shown in Figure 3.

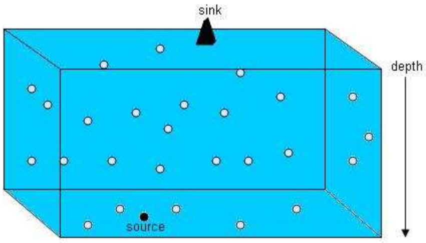


Figure 3: Source is placed at the bottom and the sink is at the top.

The depth considered is less than 100 m. The source node transmits 50 packets at a varying frequency of 500 Hz to 25000 Hz. SNR is calculated with varying frequency in both shallow and deep water. Within the frequency range 500 Hz − 25000 Hz, shipping noise, caused by water vehicle, is not considered because only a frequency level of below 500 Hz affects the shipping noise. Within the range 500 Hz − 25000 Hz only wind speed is considered as it affects the ambient noise. The wind speed is set to $2ms^{-1}$. We note that in shallow water the spreading attenuation is cylindrical.

Deep Water

For deep water, we consider the cuboid of dimension 500 × 500 × 1000 m^3. Using the Coppens equation with varying temperature in the range $22.0°C - 7.2°C)$, salinity in the range $36.5 - 34.8$ parts per thousand and depth in the range $200 - 1000$ m, a range of 1531.84 ms^{-1} to 1482.75 ms^{-1} is obtained for the underwater propagation speed.

Again, we consider the case where the source and the sink are at the same level. We set the transmission range to 100 m. The source is set at (10, 10, z) and the sink at (450, 450, z), where z is varied according to depth, thus making the value of the propagation speed to vary.

Next, we consider a source sensor node and a AUV. Both of them are placed in deep sea as shown in Figure 4. In the case of deep water the propagation speed is considered as 1507.83 ms^{-1}. The frequency varies from 500 Hz to 25000 Hz.

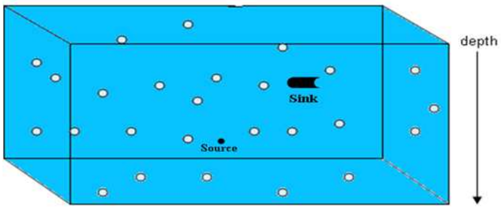


Figure 4: Source and sink are placed deep in the sea.

6.0 Results

In this section, we present the results which we have obtained from the simulations carried out. It follows from Coppens equation that in shallow water when the depth increases from 10 m to 100 m, the propagation speed also increases. Thus, keeping the same routing path and the same distance travel by the 50 packets from the source to the destination, the propagation delay decreases with increase in depth in shallow water as shown in Figure 5.

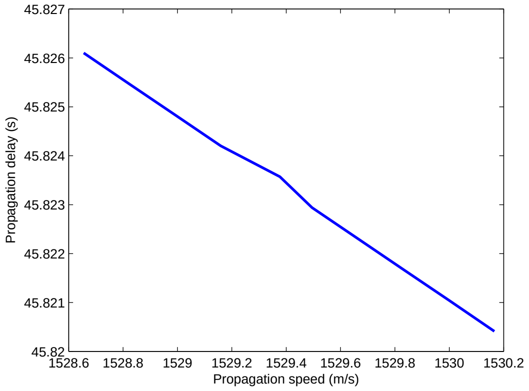


Figure 5: Propagation delay in shallow water.

Again from Coppens equation, it is clear that underwater propagation speed decreases with the increase of depth from 200 m to 1000 m. Hence in deep water environment, Figure 6 shows that the propagation delay increases when the depth is increased.

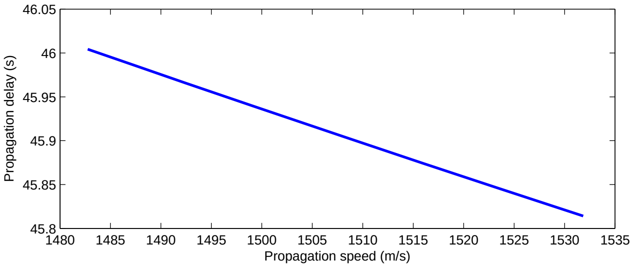


Figure 6: Propagation delay in deep water.

Attenuation is the decrease of the signal strength. It depends on the distance and the spreading factor. Figure 7 shows the attenuation loss using VBF routing in deep and shallow water. The routing path distance and the spreading factor have been used to obtain the attenuation. The figure also shows that with an increase in frequency, the attenuation increases in both shallow and deep water. But the attenuation loss in deep water is much higher than in shallow water by about 27 dB.

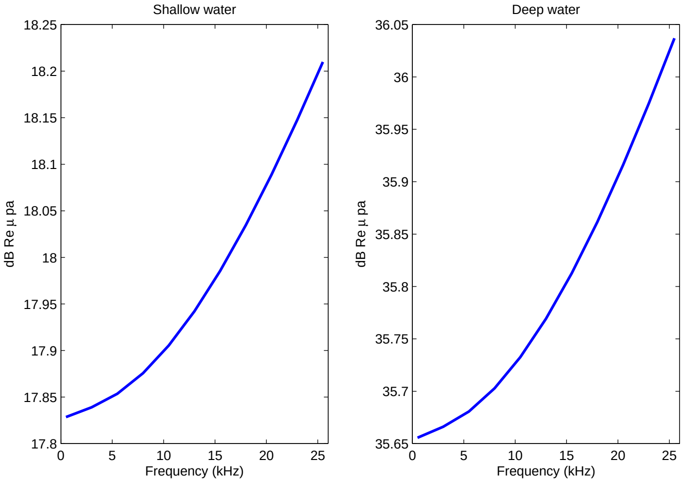


Figure 7: Attenuation in shallow and deep water.

Ambient noise also known as background noise is the loss due to its environment. In Figure 8, we show the ambient noise level in deep and shallow water.

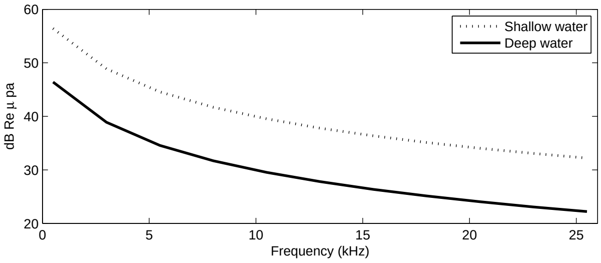


Figure 8: Ambient noise in shallow and deep water.

Total attenuation is the combined loss of the ambient noise and the attenuation due to path loss. Despite the increase of the path loss with the increase of frequency, the total attenuation of the signal decreases with the increase of frequency in both shallow and deep water as shown in Figure 9. However total attenuation in deep water is much higher than in shallow water when shipping noise is ignore and wind speed is 2 ms^{-1}.

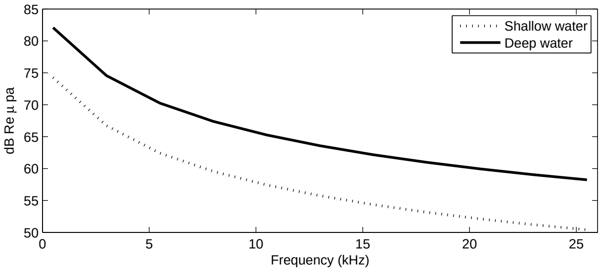


Figure 9: Total attenuation in shallow and deep water.

From Figure 10, we observe that SNR in shallow water is much higher than in deep water. This is due to the higher attenuation loss in deep water than in shallow water. With an increase in frequency, the SNR also increases in both shallow and deep water.

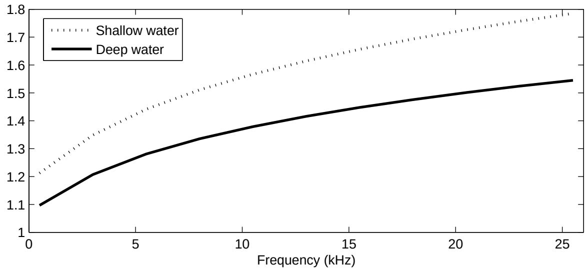


Figure 10: Signal to noise ratio is shallow and deep water.

7.0 Conclusion

Based on our results, we conclude that although ambient noise in shallow water is higher than in deep water, the vector based forwarding routing protocol performs better in shallow water than in deep water. This is due to the attenuation of the signal which is much higher in deep water than in shallow water. Also pressure is higher in deep water than shallow water and this causes a rapid decrease in signal strength in deep water as compared to shallow water.

References

[1] I. F. Akyildiz, D. Pompili D, and T. Melodia T. State of the art in protocol research for underwater acoustic sensor networks. *Mobile Computing and Communications Review*, 11:11–22, 2007.

[2] V. Chandrasekhar, W. K. Seah, Y. S. Choo, and H. V. Ee. Localization in underwater sensor networks - survey and challenges. In *Proceedings of the 1st ACM international workshop on Underwater networks*, pages 33–40, 2006.

[3] A. B. Coppens. Simple equations for the speed of sound in neptunian waters. *Journal of the Acoustical Society of America*, 69:862–863, 1981.

[4] M. Domingo and R. Prior. Energy analysis of routing protocols for underwater wireless sensor networks. *Computer Communications*, 31:1227–1238, 2008.

[5] Z. Guo, G. Colombi, B. Wan, J.-H. Cui, D. Maggiorini, and G. P. Rossi. Adaptive routing in underwater delay/disruption tolerant sensor networks. In *2008 Fifth Annual Conference on Wireless on Demand Network Systems and Services*, pages 31–39, 2008.

[6] Z. Jiang. Underwater acoustic networks - issues and solutions. *International journal of intelligent control and systems*, 13:152–161, 2008.

[7] N. Nicolaou A. See P. Xie P, Z. Zhou and J.-H. Cui amd Z. Shi. Efficient vector-based forwarding for underwater sensor networks. *Journal on Wireless Communications and Networking*, pages 1–14, 2010.

[8] P. Xie and J.-H. Cui amd L. Lao. VBF : Vector-based forwarding protocol for underwater sensor networks. In *Proceedings of the IFIP Working Conference on Networking*, 2009.

[9] P. Xie, Z. Zhou, Z. Peng, H. Yan, T. Hu, J.-H. Cui, Z. Shi, Y. Fei, and S. Zhou. Aqua-sim: A ns-2 based simulator for underwater sensor networks. *IEEE/MTS OCEANS*, 2009.

Uplink Performance Analysis in Multiple MIMO Rayleigh Interference Channel for WCDMA

E. K. Affum Ampoma[1] , Reynolds Okai [2], and Stanley Moffatt[3]

[1.] Department of Computer Science, Koforidua Polytechnic, Koforidua, Ghana

[2.] School of Engineering, Koforidua Polytechnic, Koforidua, Ghana

[3.] School of Informatics and Engineering, Regent University, Accra, Ghana

eaffume@gmail.com, ssmoffatt@gmail.com

Abstract- *WCDMA has emerged as the most widely adopted air interface technology for Third Generation Systems as defined by the 3GPP. Evolved from CDMA and providing high spectrum efficiency, the percentage of lost call per cell represents the percentage of established calls that were lost as a result of either pilot pollution or any other reason, most importantly interference. This paper evaluates the Bit Error Rate (BER) of MIMO uplink WCDMA system in multiple interference with Maximum Mobile Transmitted Power and Equivalent Isotropic Radiated Power (EIRP) of 21dBm and 18dBm respectively. The paper further presents simulation results to support the theoretical analysis on reverse link capacity analysis in terms of cell load factor.*

Keywords: WCDMA, MIMO, Interference, Uplink load Factor

1 Introduction

The Code Division Multiple Access (CDMA) system is an interference limited system in which link performance depends on the ability of the receiver to detect a signal in the presence of interference. Therefore, the key issue in a CDMA network design is to minimize multiple access interference that can be achieved by critical power control. Interference on the voice channels causes cross talk where the subscriber hears interference in the background due to an undesired transmission on control channels, leading to missed and blocked calls due to errors in the digital signaling. Interference is more severe in the urban areas, due to the greater RF noise floor and the large number of base stations and mobiles [1] and therefore, has been recognized as a major bottleneck in increasing capacity responsible for dropped calls [2]. Sources of interference include another mobile in the same cell, a call in progress in a neighboring cell, other base stations operating in the same frequency band, or any cellular system which inadvertently leaks energy into the cellular frequency band. In [3], Heiska et al analyzed capacity reduction of WCDMA downlink in the presence of interference from adjacent narrow-band system by taking into account different downlink interference mechanisms such as wide-band noise from the transmitter as well as adjacent channel interference, intermodulation, and cross-modulation originating in the mobile receiver, and concluded that capacity per cell is sensitive to the cell size, and therefore, very careful network planning is needed in order to operate the WCDMA networks efficiently. The Interference Performances, when WCDMA and HSDPA coexist as analyzed by Pei Li [4], provided simulation results indicating that the system performance in the hybrid cells is better than the pure macro cell for WCDMA and HSDPA. After investigation of WCDMA inter-operator adjacent channel interference , Joyce et al [5], proposed a number of measures which both operators and vendors should take to avoid deadzones in an operational WCDMA network.

Extensive studies have been done by Gao Peng et al [6] where they analyzed the interference between WCDMA and WIMAX systems to evaluate the impact of inter-system interference produced by coexistence of systems in the same geographical area in adjacent frequency and concluded that WCDMA and WiMax systems could coexist, and gave the proposals of interference mitigation method in the case of coexistence of two systems. Results on the other-cell to own-cell interference values and traffic capacity for dedicated indoor WCDMA systems were presented in [7]. Also, Kiiskila et al [8] discussed receiver complexity and presented optimal and suboptimal spatial maximum, a posteriori receivers in a concatenation of Linear Minimum Mean Square Error (LMMSE) equalizer structure for Multiple-Input-Multiple-Output (MIMO) Wideband Code

Division Multiple Access (WCDMA) systems, and further, proposed that in frequency selective fading channels where LMMSE part mitigates the Multiple Access Interference (MAI) and Inter-Antenna Interference (IAI) is by the spatial MAP or its approximation. Potential GPRS 900/180-MHz and WCDMA 1900-MHz Interference to Medical Devices were also investigated by Iskra et al and compared the potential for interference to medical devices from Radio Frequency (RF) fields radiated by GSM 900/1800-MHz, General Packet Radio Service (GPRS) 900/1800-MHz, and Wideband Code Division Multiple Access (WCDMA) 1900-MHz handsets. Performance analysis of MQAM for MIMO WCDMA systems in fading channels has been extensively studied [10] with authors developing an analytical framework that could handle an arbitrary number of transmit and receive antennas in both open-loop and closed-loop systems with numerical results showing that the system could achieve significant performance improvement by using the combined transmit and receive antenna diversity.

The above analyses and importance notwithstanding, WCDMA needs further investigation especially in the context of its performance in coexistence with other systems. In this context the uplink WCDMA was analyzed in Raleigh fading channel in the present study. Specifically we present an accurate BER analysis of a MIMO uplink WCDMA system in multiple interference employing Maximum Mobile Transmitted Power and Equivalent Isotropic radiated power (EIRP) of 21dBm and 18dBm respectively with emphasis on the interference from adjacent cells. This paper further presents simulation results to support the theoretical analysis on reverse link capacity analysis in terms of cell loading factor.

Organization: In section two, the uplink load factors and efficiency of multiuser receiver of CDMA system are analyzed. Section three focuses on the MIMO system. Intercell interference and the reverse link capacity in single-cell and multi systems are analyzed in section four. In section five, numerical and simulation results are shown and discussed, while section six concludes this paper.

2 CDMA Uplink Load Factors

2.1 Uplink Load Factors

The (E_b/I_t) for the ith user is expressed as

$$\left(\frac{E_b}{I_t}\right)_i = \frac{\beta_c}{(\beta_i v_i)} \cdot \frac{R_i}{(I_{total} - R_i)} \qquad (1)$$

Where β_c is the chip rate, R_i is the received signal power from ith user, the channel activity factor of ith user is represented by v_i. β_i is the bit rate of the ith and I_{total} as total received power including thermal noise power at the base station. Let $R_i = \varphi_i . I_{total}$ where φ_i is the load factor of the ith connection [1], then

$$\varphi_i = \frac{1}{1+\frac{\beta_c}{(E_b/I_t)_i \cdot \beta_i v_i}} \qquad (2)$$

The total received interference [1], without the thermal noise N_T, can be expressed as the sum of the total received powers from all M user in the same cell $I_{total} - N_T = \sum_{i=1}^{M} R_i = \sum_{i=1}^{M} \varphi_i \cdot I_{total}$. If *noise rise* of the entire system is expressed as $noise\ rise = I_{total}/N_T = 1/(1 - \sum_{i=1}^{M} \varphi_i) = 1/(1 - \delta_{ul})$, where δ_{ul} is the uplink load factor and is expressed as $\delta_{ul} = \sum_{i=1}^{M} \varphi_i$. When δ_{ul} approaches 1, the corresponding noise-rise approaches infinity and the system reaches its pole capacity. If γ is the interference factor due to other cells, in terms of interferences from other cell, the interference factor can be expressed as

$$\gamma = \frac{other\ cell\ interference}{own\ cell\ interference}$$

Then uplink load factor can then be written as

$$\delta_{ul} = (1+\gamma). \sum_{i=1}^{M} \varphi_i \qquad (3)$$

$$= (1+\gamma). \sum_{i=1}^{M} \left(\frac{1}{1+\frac{\beta_c}{(E_b/I_t)_i \cdot \beta_i v_i}} \right) \qquad (4)$$

The load equation at (4) predicts the amount of *noise-rise* over *thermal-noise* due to interference, and also could be used to make semi analytical predictions of the average capacity of CDMA cell and finally could also be employed in predicting cell capacity and planning noise-rise for dimension purposes [1]. The noise-rise is equal to $-10\log(1 - \delta_{ul})$. The interference margin in the link budget must be equal to the maximum planned *noise-rise*

2.2 Multiuser Receiver Efficiency

The interference caused by the presence of other users in the cell is called Multiple Access Interference [MAI]. Conventional signal detectors detect only single user's signal. When there are multiple users in the same environment, the conventional detectors treat other users' signals as noise or interference. MAI affects system capacity and system performance. When there are more users, the MAI is high [1]. The system performance is also affected by the near-far problem. Mitigation of the MAI is possible by, good cross-correlation code waveform design, open-loop power control for mobiles and closed-loop power control for the base station, forward error correction code and sectorized or adaptive antennas that focus reception over a narrow desired angle range. The use of multiuser detection techniques has also been suggested in the WCDMA UMTS system. Multiuser detection (MUD) and interference cancellation (IC) technique improve the system performance by canceling the intercell interference. MUD also known as co channel interference suppression or multiuser demodulation exploits the considerable structure of the multiuser interference in order to increase the efficiency with which channel resources are employed [11].

Since MUD efficiency varies in different radio environment, the capacity improvement attainable by MUD is not fixed. The impact of MUD on coverage introduces a new variable to the network planning process, since MUD efficiency needs to be taken into account in the coverage design. The efficiency of MUD is estimated from the load that can be supported with a specified E_b/I_t value with a multiuser received. In the analysis, the number of users with a RAKE receiver is represented by M_{RAKE} and those with a MUD receiver by M_{MUD}. The efficiency of MUD receiver also denoted by η at a give E_b/I_t is [12] $M_{RAKE} = (1-\eta)M_{MUD}$. The capacity of the network MUD receiver in base transceiver station (BTS) in terms received signal power R_{sp}, power control efficiency γ_c is expressed as

$$\frac{E_b}{I_t} = \frac{R_{sp} P_G \gamma_c}{(1-\eta)\psi_{Intra} + \psi_{Inter} + N_o} \qquad (5)$$

Where N_o is thermal noise, ψ_{Intra} is the intracell interference from own cell mobiles, ψ_{Inter} is the interference from the mobiles not connected to this particular base station, and P_G is the processing again. But $\mu = \psi_{Intra} / \psi_{Intra} + \psi_{Inter}$ and hence $\psi_{Inter} = ((1-\mu)/\mu)).\psi_{Intra}$. Therefore, substituting it into Equation (5) and neglecting the effect of thermal noise. Equation (5) becomes

$$\frac{E_b}{I_t} = \frac{R_{sp} P_G \gamma_c}{(1-\eta)(M-1)R_{sp} + \left(\frac{1-\mu}{\mu}\right) M.R_{sp}} \qquad (6)$$

Where M is the number of users associated with the BTS. Further solving equation (6) for M, will result in

$$M = \frac{\mu[\gamma_c P_G (E_b/I_t)^{-1} + (1-\eta)]}{(1-\mu\eta)} \qquad (7)$$

In an unloaded network, the uplink limits the achievable range and coverage, as the maximum transmission power of the mobile station is lower compared with the maximum transmission power of the base station in the downlink. In a loaded network, the downlink may limit the range if there is more load and thus more interference in the downlink than the uplink. The received signal-to-interference ratio at the base station is given as

$$\frac{E_b}{I_t} = \frac{E_{b,loaded}}{\psi_{Intra} + \psi_{Inter} + N_o} \qquad (8)$$

Where E_b is the received energy per bit, ψ_{Intra} is the intracell interference from own cell mobiles, ψ_{Inter} is the interference from the mobiles not connected to this particular base station, and N_o is the thermal noise. In case of an unloaded network $\psi_{Intra} = 0, \psi_{Inter} = 0$, and the required E_b/N_o for range calculations is equal to E_b/I_t. In the loaded network, the fraction of own-cell interference from total interference is defined as

$$\varpi = \frac{\psi_{Intra} + S}{\psi_{Intra} + S + \psi_{Inter}} \qquad (9)$$

Where $S = E_b/P_G$ the received signal is power from one user and P_G is the processing gain. ϖ depends upon propagation environment. The higher the path-loss attenuation factor, the higher the ϖ. ψ_{Inter} can be expressed in term of ψ_{Intra} as

$$\psi_{Inter} = \psi_{Intra}\left(\frac{1}{\varpi} - 1\right) + \frac{E_{b,loaded}}{P_G}\left(\frac{1}{\varpi} - 1\right) \qquad (10)$$

but

$$\psi_{Intra} = (M-1)\frac{E_{b,loaded}}{P_G} \qquad (11)$$

therefore,

$$(\psi_{Inter} + \psi_{Intra}) = \left(\frac{M}{\varpi} - 1\right)\frac{E_{b,loaded}}{P_G} \qquad (12)$$

$$\frac{E_{b,loaded}}{I_t} = \frac{E_{b,loaded}}{\left(\frac{M}{\varpi} - 1\right)\frac{E_{b,loaded}}{P_G} + N_o} = \left(\frac{E_b}{N_o}\right)_{unloaded} \qquad (13)$$

Solving the required E_b/N_o in the loaded case gives

$$\left(\frac{E_b}{N_o}\right)_{loaded} = \frac{1}{\left(\frac{E_b}{N_o}\right)^{-1}_{unloaded} - \left(\frac{M}{\varpi} - 1\right)\frac{1}{P_G}} \qquad (14)$$

The effect of the MUD receiver can be taken into account by using the efficiency of the MUD η as a measure of performance of the MUD receiver. With MUD receiver, the intracell interference $\psi_{Intra,MUD}$ can be written as

$$\psi_{Intra,MUD} = (1-\eta)\psi_{Intra}$$

$$= (1-\eta)(M-1)\frac{E_b}{P_G} \qquad (15)$$

and

$$\psi_{inter} =$$

$$\left(\frac{1}{\varpi}\quad 1\right)(1\quad \eta)(M\quad 1)\frac{E_b}{R_a} \mid \frac{E_b}{R_a}\left(\frac{1}{\varpi}\quad 1\right) \qquad (16)$$

the total interference will be

$$\psi_{inter} + \psi_{intra,MUD}$$

$$= \frac{E_{b,loaded}}{R_a}\left[\frac{M(1-\eta)+\eta}{\varpi} - 1\right] \qquad (17)$$

The required E_b/I_t in the loaded network with MUD receiver becomes

$$\left(\frac{E_b}{N_o}\right)_{loaded,MUD} \qquad (18)$$

$$= \frac{1}{\left(\frac{E_b}{N_o}\right)^{-1}_{unloaded} - \left[\frac{M(1-\eta)+\eta}{\varpi} - 1\right]\frac{1}{R_c}}$$

The transmitted power from a mobile is given as

$$S_{TX,MS}$$

$$= \frac{E_b}{N_o} + R_b + N_f + kT - G_{HO} - G_{MS} - G_{BS} \qquad (19)$$

In Equation (19) above all the terms are the same except for E_b/N_o, regardless of the base station receiver algorithm. $S_{TX,MS}$ is determined only from the E_b/N_o requirement. The decrease in the required transmission power with MUD receiver is thus given as

$$\frac{S_{TX,MS}}{S_{TX,MS,MUD}} = \frac{\left(\frac{E_b}{N_o}\right)_{loaded}}{\left(\frac{E_b}{N_o}\right)_{unloaded,MUD}}$$

$$= \frac{\left(\frac{E_b}{N_o}\right)^{-1}_{loaded} - \left[\frac{M(1-\eta)+\eta}{\varpi} - 1\right]\frac{1}{R_c}}{\left(\frac{E_b}{N_o}\right)^{-1}_{unloaded} - \left[\frac{M}{\varpi} - 1\right]\frac{1}{R_c}} \qquad (20)$$

3 Frequency Selective MIMO Channel

The general expression of frequency-selective MIMO channel indicates N_I signals $x_\mu[k], 1 \leq \mu \leq N_I$ from the input of the system at each time instant k and we obtain N_O output. Therefore, the v^{th} output at time instant k can be expressed as [13]

$$y_v[k] = \sum_{\mu=0}^{N_I}\sum_{\kappa=0}^{L_r-1} h_{v,\mu}[K,\kappa] \,.\, x_\mu[K-\kappa] + n[K] \qquad (21)$$

where L_t denotes the largest number of taps among all the contributing channels. The channel matrix has the form [14].

$$H[K,\kappa] = \begin{bmatrix} h_{1,1}[K,\kappa] & \cdots & h_{1,N_I}[K,\kappa] \\ \vdots & \ddots & \vdots \\ h_{N_O,1}[K,\kappa] & \cdots & h_{N_O,N_I}[K,\kappa] \end{bmatrix} \qquad (22)$$

3.1 Receiver Processing

If coherent single-user matched filter is used where the receiver is assumed to know the fading coefficients of the user of interest and the transmitted signal from each antenna $K = 1$ [15], then an antenna will receive

$$\begin{aligned} y_1 &= A_{11}bs_{11}(t) + \sigma n_1(t) \\ &\vdots \\ y_D &= A_{D1}bs_{D1}(t) + \sigma n_D(t) \end{aligned} \qquad (23)$$

Optimum decision rule selects $b \in \{-1,1\}$ that minimizes

$$\int_0^T \sum_{d=1}^{D} |y_d(t) - A_{d1}bs(t)_{d1}|^2\, dt \qquad (24)$$

According to the optimum decision rule [10] the inner product of the $y(t)$ and $s(t)$ is the sufficient statistic [14]. This means that the optimum rule decision for a single –user case is expressed as

$$\hat{b} = sgn\left(\Re\left\{A\sum_{d=1}^{D} A_{d1}\hat{y}_{d1}^{*}\right\}\right) \qquad (25)$$

Therefore, the probability of error of a MIMO system could be expressed as [14].

$$P_k^{De}(\sigma) = E\left[Q\left(\frac{\sum_{d=1}^{D}|A_{dk}|^2}{\sum_{d=1}^{D}|A_{dk}|^2\left(\sigma^2+\sigma^2_{I_{UWB}^{(k)}}\right)}\right)\right] \qquad (26)$$

3.2 Channel Capacity of MIMO Systems

The Space Division Multiplexing (SDM) over MIMO channels using multiple transmitting and receiving antennas is one of the most promising technologies for improving bits per Hertz (bit/s/Hz). In an Additive White Gaussian Noise (AWGN) channel the channel capacity C is given by

$$C = log_2(1+\beta)$$

Where β is the signal-to-noise ratio. The MIMO channel capacity is given by [6].

$$C = log_2\left[det\left(I_{mn} + \frac{\beta}{N_T} H.H^H\right)\right]. \quad (27)$$

As the parallel channel capacity, where I is n by n identity matrix, H is a channel matrix, N_T and N_R denote the number of transmitting and receiving antennas, and $(\ )^H$ denotes the complex conjugate transpose. This equation indicates that the channel capacity can be increased in proportion to the number of antennas if $N_T = N_R$. This possible increase in terms of bits per Hertz is why SDM/MIMO is attracting a significant amount of attention these days.

4 Intercell Interference

Considering an omnidirection cell site serving a given set of mobiles, if mobiles are divided into two groups which are mobiles that are powered up and mobiles that are not powered up, the mobiles that are powered up are further, divided into four subgroups: Active and transmitting mobiles, Active but not transmitting mobiles (mobiles in non conversational mode), Idle and transmitting (mobiles in access mode) and Idle and not transmitting (mobiles in non access mode) [1]. Assume that interference at the cell site by mobiles and the access mode is typically small and neglected. This may be accounted for as a source of some degradation in system quality and capacity. We focused only on the active mobiles in our analysis.

Assume there are M mobiles transmitting at a given time in a cell. In a CDMA environment for each mobile, there are (M – 1) interferers. At the cell site, the average signal power received from the *ith* mobile is S_{ri}. This signal power provides bit energy equal $E_b = S_{ri}/R$ where, R is the mobile transmission rate in bps. The thermal noise power is $N_0 B_w$ where N_0 is the thermal noise power spectral density (psd), and B_w is the spreading bandwidth. The average interference (psd) at the base station is expressed as

$$I_o = \frac{1}{B_W}\sum_{i=1}^{M-1} V_f . S_{ri} \quad (28)$$

Where, V_f = channel activity factor.

In equation (28), assuming a perfect control in the reverse link and that the signals transmitted from all the mobiles arrived at the base station with the same received power. i.e. $S_{ri} = S$ for all values of *i (i.e.* $(1 \leq i \leq M - 1)$. The total interference andthermal noise (psd) will be

#

$$I_t = I_0 + N_0 = \frac{1}{B_W}\sum_{i=1}^{M-1} V_f . S_{ri} + N_0 . \quad (29)$$

#

Recognizing that $S_{ri} = S$, I_t then becomes#

$$I_t = \frac{(M-1).V_f.S}{B_W} + N_0 \quad (30)$$

The E_b/I_0 will be given as,

$$\frac{E_b}{I_t} = \left(\frac{B_W}{r}\right) \cdot \frac{s}{[N_0 B_w + (M-1).V_f.S]} \quad (31)$$

$$= G_p \cdot \frac{S}{[N_0 B_W + (M-1).V_f.S]}$$

Where G_p = processing gain = B_W/r. The signal strength, S in dB as,

$$S = P_m + G_m + G_b + G_{dv} + G_{sho} + L_p + M_{fade}$$
$$+ L_{body} + L_{pent} + L_{cable}$$

Where

G_m = *transmit antenna gain of the mobile (dB)*

G_b = *receive antenna gain of base station (dB)*

G_{dv} = *base station antenna diversity (dB)*

From (31)

$$M = 1 + G_p . \left[\frac{1}{(E_b/I_t).v_f}\right] - \frac{N_0 . B_W}{S.v_f} \quad (32)$$

also,

$$S = \frac{(E_b/I_0).N_0}{\frac{1}{R} - \frac{(M-1)v_f(E_b/I_0)}{B_w}} \quad (33)$$

Let $\beth$ represent the interference factor from other cells (31) can be expressed as

$$G_p . \frac{S}{[N_0 B_W + (M-1).V_f.S(1+\beth)]} \# \quad (34)$$

Include an imperfect power factor, φ and rewrite equation (34) as,

$$G_p \cdot \frac{S}{B_W \cdot N_0 + (M-1) \cdot V_f \cdot \left(\frac{S}{\varphi}\right) \cdot (1+\Xi)} \qquad (35)$$

Solving equation (35) for M, we get,

$$M = 1 + G_p \cdot \left[\frac{\varphi}{\left(\frac{E_b}{I_t}\right) \cdot V_f \cdot (1+\Xi)}\right] - \frac{N_0 \cdot B_W \cdot \varphi}{S \cdot V_f \cdot (1+\Xi)} \qquad (36)$$

Solving equation (35) for S, we get

$$S = \frac{\frac{E_b}{I_t} \cdot N_0}{\frac{1}{R} - \frac{(M-1) \cdot V_f \cdot (1+\Xi)\frac{E_b}{I_t}}{B_W \cdot \varphi}} \qquad (37)$$

From equation (36), the maximum value of M is,

$$M_{max} = 1 + G_p \cdot \left[\frac{\varphi}{\left(\frac{E_b}{I_t}\right) \cdot V_f \cdot (1+\Xi)}\right] \qquad (38)$$

M_{max} is called the pole point or asymptotic cell capacity that is achieved as $S \to \infty$. For simplification, neglecting 1 and rewriting equation (38) gives,

$$M_{max} \approx G_p \cdot \left[\frac{\varphi}{\left(\frac{E_b}{I_t}\right) \cdot V_f \cdot (1+\Xi)}\right] \qquad (39)$$

equation (33) can further be expressed as,

$$\frac{S/\varphi}{N_0 B_W} = \frac{1}{M_{max} \cdot V_f \cdot (1+\Xi) \cdot (1-\rho)} \qquad (40)$$

where, $\rho \blacksquare \frac{M}{M_{max}} \blacksquare$ cell loading factor.

5 Results & Discussions

In this study, a maximum cell loading factor of 18% and signal to noise ratio of $30dB$ were used for the capacity analyses. With M mobiles of 20 and received antenna gain of the base station of $9dB$ and with maximum mobile transmitted power and equivalent isotropic radiated power of $18dBm$ and $21dBm$ respectively, the bits per second performance for the 4*4 MIMO System shown in figure 1 for $30dB$ is around 35bps that overwhelmed the other systems. Figure 2 also provides the BER analysis of the various systems in multiple interferers. With cell loading factor of 0% and signal to noise ratio of $10dB$ the 4*4 MIMO systems performed better. With cell load of 50% and a cell range of 1.02km the allowable path loss without MUD was around $98.48dB$. With path loss with MUD at a cell range of 1.357km was around $110.67dB$. It was also realized that as the cell range and allowable path loss in dB decreases the cell load increases dramatically. Interestingly, similar results were obtained in [1] where they further observed that base station multiuser detection (MUD) receiver can provide good coverage even with high system load after initial deployment and, finally, concluded that the effect of MUD on cell range depends on propagation environment.

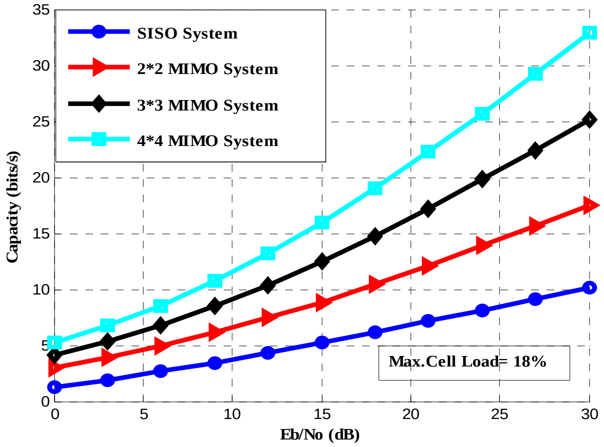


Fig. 1 Capacity performance of WCDMA with 18% Cell Load factor

6 Conclusion & Future Work

Analyses of coverage of a loaded and unloaded WCDMA network conducted in this paper revealed that the propagation environment affects the cell range with a given cell loading. Furthermore, for efficient CDMA operation the spectrum must be cleared in a sufficient guard band and guar zone. Also, spectrum monitoring is highly recommended as early as possible in the CDMA system since it is tedious to identify the source of external interference. Intermodulation interference, adjacent and co-channel interference could be considerd in further studies.

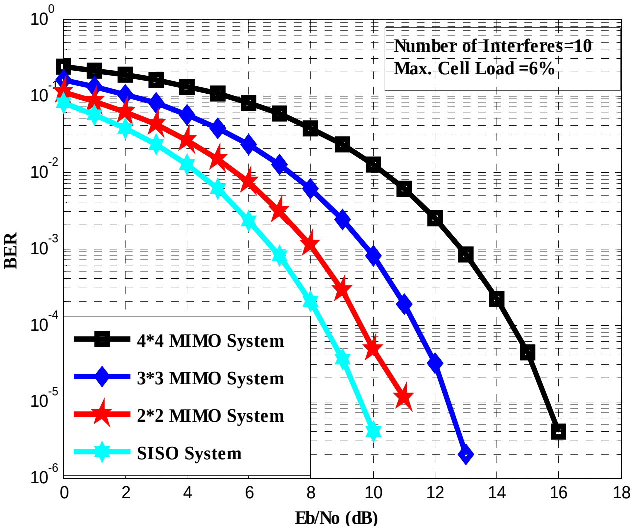


Fig. 2 BER performance of WCDMA in Multiple

References

[1] Vijay Garg, K. *"Wireless Networks Evolution 2G&3G"* Prentice Hall PTR, New Jersey, 2002.

[2] Theodore Rappaport, S. "*Wireless Communications*" Second Edition, Pearson Education, Inc, New Jersey, 2002.

[3] Heiska, K., Posti, H., Muszynski, P., Aikio, P., Numminen, J., Hamalainen, M."Capacity Reduction Of WCDMA Downlink in the Presence of Interference from Adjacent Narrow-Band System" *Vehicular Technology, IEEE Transactions*, vol. 51, pp. 37, Jan 2002.

[4] Pei Li, "The Interference Performances when WCDMA and HSDPA Coexist", *Information and Communication Technologies, IEEE,* vol. 2, pp. 2450, Damascus, Oct. 2006.

[5] Joyce, R.M., Graves, B.D., Osborne, I.J., Griparis, T., Conroy, G.R."An investigation of WCDMA inter-operator adjacent channel interference", *3G Mobile Communication Technologies, 4th International Conference, IEEE***,** pp. 149, 25-27 June 2003.

[6] Gao Peng, Tu Guofang, Fang Yuan, Liang Shuangchun "The analysis of the interference between WCDMA and WIMAX systems", *Communications Technology and Applications, ICCTA '09. IEEE International Conference,* pp. 180, Beijing, Dec. 2009.

[7] Schuh, R.E., Andersson, R., Stranne, A., Sommer, M., Karlsson, P. "Interference analysis for dedicated indoor WCDMA systems", *Vehicular Technology Conference, IEEE* , vol.2, pp. 992, 6-9 Oct. 2003.

[8] Kiiskila, K., Hooli, K., Ylioinas, J., Juntti, M., "Interference resistant receivers for WCDMA MIMO downlink", *Vehicular Technology Conference, VTC 2005-spring, IEEE*, vol. 2, pp. 836, May-June 2005.

[9] Iskra, S., Thomas, B.W., McKenzie, R., Rowley, J. "Potential GPRS 900/180-MHz and WCDMA 1900-MHz Interference to Medical Devices", *Biomedical Engineering, IEEE Transactions*, vol. 54, issues 10, pp. 1858, Oct. 2007.

[10] Fuyong Xu, Guangqiu Li "Performance analysis of MQAM for MIMO WCDMA systems in fading channels", *Communications, Circuits and Systems International Conference IEEE*, vol. 1, pp. 207, 27-30 May 2005.

[11] Sergio, V. "*Multiuser Detection*" The Press Syndicate of the University of Cambridge, Pit Building Trumpington Street, Cambridge, CB2, 1RP, United Kingdom.

[12] Ojanpera, T., and Prasad, R. "*Widband CDMA for Third Generation Mobile Communication*" Artech House, Boston, 1998.

[13] Foerster J. "*Channel modeling sub-committee report final*", IEEE 802.15-02/490 (see http://ieee802.org/15/)

[14] Ampoma, E.K.A., Rao, T.R. Labay, V.A. "Capacity & performance issues in a MIMO based MB-OFDM ultrawide band communication system", *Adaptive Science & Technology, 2nd International Conference, IEEE*, pp. 432, Accra, Jan. 2009.

[15] Liu, H., Qiu, R. C., Tian, Z., "Error performance of pulse-based ultra wideband MIMO systems over indoor wireless channels," *IEEE Wireless Communication conference.* vol.4, pp. 2939–2944, Nov. 2005

[16] Proakis G. *"Digital Communication," Fourth Edition*, Irwin/McGraw-Hill, an imprint of the Mcgraw-Hill Companies, Inc, New York, USA, 2000.

SESSION

MOBILE COMPUTING AND AD HOC NETWORKS

Chair(s)

TBA

Location Cluster with Nearest Neighbors in Signal Space: An Implementation in Mobile Service Discovery and Tracking

Jon Tong-Seng Quah [1], and Luo-Ren Lim [1]

[1]School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

Abstract - *A rapid development in mobile telecommunication is the use of location information for context aware service. To amalgamate multiple location sensor technologies, we developed an Intelligent Service Area and Region Identification System (iSARIS) to allocate system resource for mobile service discovery and tracking. Primary location service is provided by cell identification which maps a wide region of physical space. A switching module toggles a secondary localization service within the interest space. This service provides additional localization performance through Wi-Fi pattern mapping to further segment smart spaces where independent providers can offer their mobile services. The technique uses location cluster to minimize search boundary through the use of an enhanced Nearest Neighbor in Signal Space algorithm.*

Keywords: Location Service, Mobile Service Discovery

1 Introduction

The primary function of communication in a mobile phone today has rapidly expanded to include many computing features. Shrinking of computing capabilities into more portable form is becoming mainstream. Smartphone brings an evolutionary change to how people view computing power. Technologies such as GSM, UMTS, GPS, Wi-Fi, and Bluetooth are included in a single mobile platform that can support many computing activities. By creating intelligent user centric services through the extraction of meaningful contexts in mobile devices, developers greatly improve user experience. Many of today's popular mobile applications utilize wireless localization. Location information is therefore emerging as a key context for mobile applications.

Techniques for wireless localization can be generalized into deterministic and probabilistic methods. Deterministic methods perform well in environment with clear line-of-sight (LOS) propagation paths but are not as applicable in enclosed spaces where probabilistic methods [1] are more commonly used. This is due to problems like multipath propagation, fading and signal dragging effects. Mobile phones on the move are likely to visit locations which encompass both outdoor and indoor situations. As optimal localization techniques differ for both situations, location systems require a rethink of how we can offer seamless user experience in an environment transition. Getting user to manually switch applications is inefficient.

2 Mobile Localization

Location is the most utilized aspect of context awareness and can be expressed as a spatial form or textual descriptions. Spatial location is expressed in a coordinate system involving latitude, longitude and altitude using standard like the World Geodetic System (WGS 84) datum. Textual description can either be an address or a landmark and often more ambiguous than a spatial expression. Many existing applications exploit location information to search for products, create tour guide, offer direction guidance, match neighboring devices, send mobile advertising and social networking [2]. The main classes of applications are listed in Table I.

TABLE I
WIRELESS LOCALIZATION APPLICATIONS

Applications	*Uses*	*Examples*
Navigation	Handling routes between locations	Road directions, Journey planning, Tour guide
Positioning	Identifying a spatial fix	Emergency services
Tracking	Monitoring movement of objects	People locator, Assets tracking, Social networking, Mobile advertising
Mapping	Creating graphical representation	Terrain mapping

2.1 Network Localization

The Cell Global Identity (CGI) is a globally unique identity made up of Cell Identification (CID), Location Area Code (LAC), Mobile Network Code (MNC), and Mobile Country Code (MCC) to address mobile cells worldwide. The concept works on the fact that mobile phone is always connected to the closest cell towers. By identifying which towers the phone communicates with, we will know that the phone is near to a particular Base Transceiver Station (BTS) and hence estimate its location. In practice, this process is complicated and involves parameters that network can optimize such as signal quality and variability. Mobile phone usually locks onto the strongest signal but not necessarily the closest as mobile networks are optimized for capacity and call handling rather than localization. Trevisani E. and Vitaletti A. [3] estimate an average accuracy to about 500-800 meters and suggest its use in resource discovery service. CGI depends on distribution of base stations for location precision. Network knowledge of the phone within the controlling cell site together with sector information enables a rough estimate of the location of the caller, regardless of phone type. E-CID [4] enables accuracy improvement over conventional Cell ID by using additional Timing Advance and Network Management Records from GSM networks to fine tune measurements by adding measured time between the start of a radio frame and a data burst to CGI.

Other forms of network based techniques process localization results at a network server. In GSM networks, information flows are managed through free-to-access control channel. This allows manufacturers to implement features to monitor neighboring cells and its corresponding RSS value. Accuracy is dependent on the concentration of BTS cells. Uplink Time Difference of Arrival (U-TDOA) is one network based solution that determines location through cellular signal. It compares the times a signal reaches multiple Location Measurement Units (LMUs) installed at the operator's base transceiver stations. U-TDOA is well suited for indoor and urban environments relying on multilateration to get a location fix. Enhanced Observed Time Difference (EOTD) systems utilize cellular characteristics of asynchronous GSM network. Each GSM based BTS emits a synchronization burst to all mobile subscribers in its vicinity regularly. Mobiles monitor the synchronization bursts of the service and neighboring BTSs to maintain connection. EOTD extends on this GSM functionality. Location is determined by comparing signal arrival times from three or more BTSs at the mobile phone. The mobile phone records burst arrival times and deduce a position using the coordinates of the BTSs, arrival time of synchronization bursts from each BTS and the timing differences between BTSs. The use of an external reference point eliminates the need for transceivers to remain time-synchronized, but in contrast it requires enhancement to phone software and additional network equipment.

2.2 Satellite Localization

The global positioning system (GPS) [5] is a wide area outdoor radio positioning system that employs orbiting satellites for location fix. The system consists of 24 satellites in 6 circular orbits at an altitude of approximately 20200km. Each orbit contains 4 equally spaced satellites inclined at 55 degrees. Once a receiver locked-on to a satellite, the receiver recognizes and time shifts its internal clock through a unique Coarse/Acquisition code. The time to lock-on is known as the time-to-first-fix (TTF) and can take up to 15 mins from a cold start. When the clock synchronized with the satellite's atomic clocks, the distance to each satellite could be determined by subtracting the known transmission time from the calculated receiver time. Two signals in the L1 (1575.42MHz) and L2 (1227.60MHz) bands are broadcast by the satellites but only L1 is for public use. Receivers make pseudorange or carrier phase measurement on the L1/L2 signals to generate a location reading. A lock-on would not be achieved from fragmentary signal until a clear signal can be received continuously. Ignoring ground effects, the worse case horizontal positioning accuracy based on line of sight signals is ≤ 22m at 95% confidence interval although accuracy up to ≤10m is typically achieved. Upgrades under a GPS Modernization program [6] improved acquisition codes to better account for ionospheric errors, radio frequency and multipath interferences.

Assisted Global Positioning System (A-GPS) [7] improves startup performance of GPS system by using network data. Some A-GPS implementations reduced the amount of processing required by offloading the processing work onto the network's server. An assistance server have better satellite signal due to static placement and has higher computational performance. It can supply the GPS almanac to a receiver, thus allowing the GPS module to lock on to the satellites more rapidly. It can also provide precise timing information. Cell towers with assistance functions have accurate coordinates which account for various local factors affecting the GPS signal. The tower can also compare fragmentary signals received from GPS receivers with its own reading to obtain a faster location fix.

2.3 Short Range Radio Localization

Short range radio technologies such as Wi-Fi and Bluetooth are commonly found in smartphones. Wi-Fi allows mobile user access to internet with WLAN access points. Bluetooth is a short range PAN standard used in simple file transfer and headset connectivity. These network protocols are often adapted to implement location schemes. Such technique uniquely identifies locations by comparing and deducing RSS signal patterns. However, they are certain limitations. Despite its association with power, RSSI value is arbitrarily decided by the equipment manufacturers. Different manufacturers provide their own accuracy, granularity, and range for the actual power and RSSI values. Thus, different equipment or software would exhibit different sensitivity to a single transmission despite being at the same location. This affects the accuracy of signal fingerprinting based techniques as radio maps generated is often only generic to the device where it was created and not easily ported to other devices. Nevertheless, this method has proven to be a fairly efficient way to conduct positioning in multipath and obstructions dominated spaces compared to conventional deterministic methods. Short range radio localization using *k*-Nearest Neighbor (*k*NN) algorithm has been frequently applied for Wi-Fi location tracking [8-10, 13]. Jayaraman et. al. [14] demonstrated 3 dimensional space localization processes using mobile data collectors in wireless sensor network.

3 iSARIS Platform

3.1 Purpose

The primary objective of iSARIS is to integrate sensing technologies onto a single layer where the strengths of each can be maximized. Localization services are migrated to a single platform where mobile entity switches smartly to the correct technologies for the relevant environment. Our approach is to implement a location based solution that allows mobile device to passively observe surrounding radio signals and recognize Service Region (SR) and Service Area (SA) space for smart services.

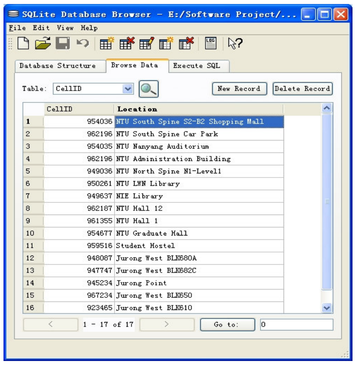
3.2 Service Discovery Background Service

A mobile service discovery service was developed using Microsoft Visual Studio 2008 and Windows Mobile SDK 6. This background service runs on the mobile phone to retrieve the active CID parameters. Retrieving cell tower information from Windows Mobile devices requires access to the Radio Interface Layer (RIL). RIL functions are implemented via a device driver through two distinct layers. The lower level Platform Dependent Driver (PDD) layer of the RIL is radio stack dependent and its implementation differs with manufacturers. The Model Device Driver (MDD) layer on the other hand is radio stack independent and contains code to communicate with the RIL Proxy as well as code that implements any radio stack independent features in the RIL driver.

Cell tower information is returned to the caller of **RIL_GetCellTowerInfo** through a callback function.

```
public class CellTowerInfo
    {
        private static AutoResetEvent dataReceived =
new AutoResetEvent(false);
        private static RIL.RILCELLTOWERINFO
towerInfo;
        public static CellTower GetCellTowerInfo()
        {
            IntPtr hRIL = IntPtr.Zero;
            IntPtr hResult = IntPtr.Zero;

            hResult = RIL.RIL_Initialize(1,
                new
RIL.RILRESULTCALLBACK(CellTowerData),
                null, 0, 0, out hRIL);
            if (hResult != IntPtr.Zero)
                return null;
            hResult =
RIL.RIL_GetCellTowerInfo(hRIL);
            dataReceived.WaitOne();
            RIL.RIL_Deinitialize(hRIL);
            CellTower tower = new CellTower();
            tower.LAC =
(int)towerInfo.dwLocationAreaCode;
            tower.MCC =
(int)towerInfo.dwMobileCountryCode;
            tower.MNC =
(int)towerInfo.dwMobileNetworkCode;
            tower.CID = (int)towerInfo.dwCellID;
            return tower;
        }

        private static void CellTowerData(uint
dwCode,
            IntPtr hrCmdID, IntPtr lpData, uint
cbData, uint dwParam)
        {
            towerInfo = new RIL.RILCELLTOWERINFO();
            Marshal.PtrToStructure(lpData,
(object)towerInfo);
            dataReceived.Set();
        }
    }
```

To retrieve the location of the cell tower that the mobile is connected to, the background service queries a SQLite database. SQLite was used to store known SR CID information due to its lightweight design for embedded system. Its low memory cost is ideal for mobile platforms and can support software development on mainstream operating system such as Windows, Linux and Unix as well as programming languages, like PHP, java and C. Simulations are conducted on the Windows Mobile 6 Emulator to test the prototype. When the service scanned for CID from the GSM network, it compares the current active CID with the dataset in the SQLite database to determine if the location is a known SR. A CID match indicates the mobile phone has entered into the vicinity of the shopping complex. If the CID does not match any known SR, the process maintains its monitoring mode. The database is shown in Fig 1



	CellID	Location
1	954036	NTU South Spine S2-B2 Shopping Mall
2	962196	NTU South Spine Car Park
3	954035	NTU Nanyang Auditorium
4	962196	NTU Administration Building
5	949036	NTU North Spine N1-Level1
6	950261	NTU LWN Library
7	949637	NIE Library
8	962187	NTU Hall 12
9	961355	NTU Hall 1
10	954677	NTU Graduate Hall
11	959516	Student Hostel
12	948087	Jurong West BLK680A
13	947747	Jurong West BLK682C
14	945234	Jurong Point
15	967234	Jurong West BLK650
16	923465	Jurong West BLK610

Fig. 1. CID table in the SQLite Database. The table shows the search space for the background service and is stored locally on the mobile device.

3.3 Service Region and Area

BTS towers are located at the corners where three hexagonal coverage cells converge in a cellular map. Each tower has three sets of directional antennas aimed in three different directions and receiving/transmitting into three different cells at different frequencies. This provides a minimum of three channels for each cell. Large cells are subdivided into smaller cells for higher volume areas. A SR is made up of the mobile coverage cells servicing the physical location of the complex and its immediate vicinity. The background service observes telecommunication exchanges between the device and its associated BTS. Any additional power used to monitor signals while on the move is kept minimal as the process runs parallel to normal telecommunication operation.

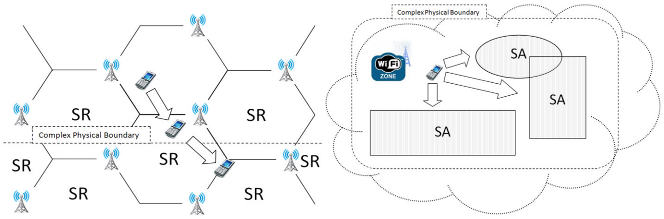


Fig 2 A mobile enters a SR from the public area as it nears the complex (left) and accessing a SA defined under Wi-Fi coverage (right).

A rapid coarse estimate of the location is obtained through the background service discovery service. An alert is flagged when an operating SR is identified which in turn triggers a secondary positioning method suitable for indoor environment. Many complexes today are equipped with WLAN for internet access purposes. Access points placed at strategic locations create Wi-Fi hotspots for consumer to access via Wi-Fi enabled devices. Users can access data services through these wireless hotspots. Signal pattern matching forms the basis for SA creation. These SAs marked a clear definition of smart space where a particular service is valid. A single SR may contain multiple SAs and overlapping areas between SAs are also possible. SA allows individual service space to be built for smaller players to offer individualized services on a smaller physical scale.

4 Mobile Application Design

4.1 Overview

A shopping mall services system was built based on the proposed iSARIS platform. Customers can interact with smart services which include those by the mall operator and shop vendors housed within the same complex. Our shopping complex implementation defined the physical shopping complex and its immediate surroundings as a SR. Upon entering the SR, customer receives an automated alert from the complex service server listing the offerings available. While roaming in the SR, the mobile device can access shopping mall services such as directory information, special events, mall promotions and helpdesk chat service. This is delivered through the shopping mall WLAN infrastructure.

Once within the SR, the mobile device enables the Wi-Fi module to improve positioning accuracy and identifies available SAs. SAs are created for vendors to define smaller physical spaces to offer their own unique services. The mall Wi-Fi infrastructure handles vendor content and decides which service to push to customer when they enter a SA. A typical service in a SA can include discount notifications, pricelists, shop advertisements and vendor promotions.

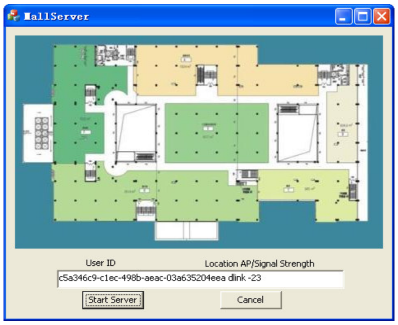
4.2 System Deployment

Mobile devices connect to the position and data server via the wireless access points (AP) installed in the shopping mall. RSSI & Media Access Control/ Service Set Identifier (MAC/SSID) information from the devices are sent to the server for processing. The devices also communicate with the server for mall and vendor information. The position server is tasked with the collection of (MAC/SSID, signal strength) value pairs gathered by customers' mobile devices to determine each device's position. Vendor websites can be access through the mall server. The mall server provides vendors corresponding authorities in the shopping mall server to manage their websites. A position server tracks mobile positions with respect to SAs in the vicinity and determines if related information is forwarded to the mobile.

Toggling between primary to secondary localization services during the detection of a SR is managed by a switching service on the device. The process of switching between positioning technology is automated with no active intervention from user. In the shopping mall implementation, the mobile automatically diverts location recognition from telecommunication cell identification to a Wi-Fi based pattern search when the mobile device nears the shopping mall. The interface module contains browsing and selection services. Functions include accessing information from mall and vendors, interactive service selection and service registration. A mobile client user interface is implemented to deliver these functions to the user. The location module handles mobile client location related events. It monitors the setup and dismantling of vendor-designated SAs and alerts mobile that enters into the physical locality of a defined SA.

4.3 Location Service

1) Position & Data Server Design: Position and data handling applications are MFC programs hosted on a mall server. The server handles tasks such as receiving signal signature from devices, running location algorithm and determining relevant vendor data to send to customer. Position server receives mobile device signal strength parameters via TCP connections. A location algorithm is used to calculate an estimated position fix for each client.



User ID Location AP/Signal Strength

c5a346c9-c1ec-498b-aeac-03a635204eea dlink -23

Fig 3 Allocation of SA on the position server. The position of the device determines whether it is in a vendor-defined SA. Data server sends corresponding vendor information to the device across a UDP connection.

2) Signal Strength Acquisition: The implementation code contains two main parts. A static ScanThreadFunc() function create a scanning thread responsible for wireless access points scan in the shopping mall. Scans for Wi-Fi access points are handled as a loop task in this thread. The thread is created when SR is identified and terminated whenever the Wi-Fi connection is disabled. The acquisition of access points' signal strength is handled in the WifiScanConnect class by a ScanWifi() member function.

```
DWORD   __stdcall   WifiScanConnect::ScanThreadFunc(
void* pParam )
{ THREAD_PARAM    *pThreadParam    =    (THREAD_PARAM
*)pParam;
  HANDLE hEvent = pThreadParam->hEvent;
  WifiScanConnect          *pWifiScanConnect         =
(WifiScanConnect *)pThreadParam->lpObject;
  while(WaitForSingleObject(hEvent,500)          ==
WAIT_TIMEOUT)
  {
    pWifiScanConnect->ScanWifi();
  }
  return 0;
}

void WifiScanConnect::ScanWifi(void)
{
  ULONG oidcode;
  ULONG bytesreturned;
  static BOOL bFlag = TRUE;
  if(bFlag)
  {
   m_pBSSIDList       =       (NDIS_802_11_BSSID_LIST
*)VirtualAlloc(NULL,
         sizeof(      NDIS_802_11_BSSID_LIST)       *
NUMBEROF_BSSIDS,
           MEM_RESERVE | MEM_COMMIT,PAGE_READWRITE);
   bFlag = FALSE;
  }
  if( m_pBSSIDList == NULL)
  {
   return;
  }
  else
  {
   memset(m_pBSSIDList,          0,          sizeof(
NDIS_802_11_BSSID_LIST) * NUMBEROF_BSSIDS);
   oidcode = OID_802_11_BSSID_LIST_SCAN;
   DeviceIoControl(m_handle,
           IOCTL_NDIS_QUERY_GLOBAL_STATS,
           &oidcode,sizeof( oidcode),
             (ULONG *) NULL,0,&bytesreturned,NULL);
   memset(      m_pBSSIDList,       0,       sizeof(
NDIS_802_11_BSSID_LIST) *NUMBEROF_BSSIDS);
   oidcode = OID_802_11_BSSID_LIST ;
   if(DeviceIoControl(m_handle,
             IOCTL_NDIS_QUERY_GLOBAL_STATS,
             &oidcode,
             sizeof( oidcode),
             (ULONG *) m_pBSSIDList,
             sizeof(    NDIS_802_11_BSSID_LIST)    *
NUMBEROF_BSSIDS,
             &bytesreturned,
             NULL))
    {
     if (m_pBSSIDList != NULL)
     {
      EnterCriticalSection(&cs);
      for(unsigned  int  i  =0;  i  <  m_pBSSIDList-
>NumberOfItems; i++)
      {
       SSIDSignal.clear();
       SSID_SIGNAL_VALUE stSsidSignalValue;
       int temp = i;
       PNDIS_WLAN_BSSID  cpBssid  =  m_pBSSIDList-
>Bssid;
       while(temp != 0 )
       {
             cpBssid                              =
(PNDIS_WLAN_BSSID)((char*)cpBssid     +     cpBssid-
>Length);
        temp--;
          }
          strcpy_s(stSsidSignalValue.cWifiSsid,
NDIS_802_11_LENGTH_SSID,
      (const char*)cpBssid->Ssid.Ssid);
      stSsidSignalValue.lSignalValue   =   cpBssid-
>Rssi;
      SSIDSignal.push_back(stSsidSignalValue);

      }
      LeaveCriticalSection(&cs);
     }
    }
```

3) Location Algorithm: The location server collects signal strength signature observed by client devices to decide their approximate position. To enable better efficiency, factors such as computation time, accuracy and power consumption are considered for the localization processes. The design reduces computation time by only extracting useful information required and optimizing location data storage. The method is based on signal pattern matching using a radio map. Its implementation is divided into two phases: Calibration Phase and Online Positioning Phase.

During the calibration phase, the RSS of APs are collected at different locations to construct radio maps. Certain factors are predetermined during this phase. Due to signal fluctuations or larger distances, some APs are not always visible throughout the scan. Signals from these rogue APs are often very weak or appear intermittently. A noise filter is applied to remove the irrelevant rogue APs from the input dataset to reduce unnecessary calculation. The noise filter module sets a time *t* to record all RSS values and APs detected. APs present most frequently within the time interval are chosen. Time *t* has to be as short as possible to reduce computation time while ensuring that sufficient information and good RSS measurements are captured for location estimation. Tests revealed optimal *t* to be at least 17 seconds.

The next criterion is to determine the number of relevant APs *n* such that there are sufficient distinguishable APs to determine the location. The condition for *n* is such that all the locations are covered by at least *n* number of APs most of the time. This ensures a minimum set of APs to distinguish positions with reasonable accuracy. To minimize data search complexity, a radio map clustering module is created to reduce the search space required to return a result. It does so by categorizing locations that shared the same set of APs during calibration phase thereby cutting down on computation time. Each group of locations is called a cluster. During the online positioning phase, the module only searches the relevant cluster which contains the detected APs instead of probing the entire radio map search space for a match. Each location cluster has at least *q* APs detected.

The matching of signal signature uses a method similar to Nearest Neighbors in Signal Space (NNSS) pattern recognition adopted in RADAR [8-10]. In general, *k*-Nearest Neighbors (*k*NN) networks tessellate the input space by setting weights and thresholds of first layer hidden neurons. This is done by calculating the boundaries of space containing points nearest the training pattern. Subsequent processing layer handles the classification for the unknown input pattern [11]. Three hidden layers of processing are applied. The first layer is an input layer that distributes all input patterns. Second layer provides feedback control input back to the first layer. The final layer forms variable threshold neurons for output classification. The *k*NN algorithm is adapted to calculate the degree of match between calibrated RSS dataset and online RSS signature. Euclidean distances are applied to measure the similarity between two points $M = (m_1, m_2, \dots, m_n)$ and $L = (l_1, l_2, \dots, l_n)$:

$$d = \sqrt{(m_1 - l_1)^2 + (m_2 - l_2)^2 + \cdots + (m_n - l_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (m_i - l_i)^2}$$

The calibration data are compared with the signal strength measurements in the mobile device during the online positioning phase using:

$$d_j = \sqrt{\sum_{i=1}^{n} (c_j^{AP_i} - s^{AP_i})^2}$$

where $c_j^{AP_i}$: RSS values of AP_i at point j in the signal strength map,
s^{AP_i}: is the measured RSS value of AP_i ,
n: total number of APs.

Small d_j distance value between a target location and an identified location may signify close proximity between two locations. The pattern matching implementation is as follows:

```
int PatternRec::LeastDistance(int ReNum)
{
  double min = 10000;
  for(int d=0; d<10; d++)
  {
    if(MatchData(type[d].Feature, InputNum)<min)
    {
      min = MatchData(type[d].Feature, InputNum);
      ReNum = d;
    }
  }
  LocNumber = ReNum;
  return ReNum;
}
```

However, conventional NNSS method put together RSS values from all known APs without considering RSS variation from individual AP. Variations in RSS can influence d. Thus, k lowest d values are considered to estimate true deviation. The choice of k depends on the size of the dataset. Larger k reduces the effect of noise on the classification but decrease the ability to distinguish boundaries between classes. An enhancement to NNSS [12] algorithm is integrated to take into account the threshold for RSS variation around its mean, θ and the maximum number of APs from which RSS varies beyond that threshold, τ.

```
Input: sorted neighbor list of size S, where $S \geq k$
begin
for m = 1 to k
  begin
    count=0
    for i = 1 to n
    begin
      if  ($|c_j^{AP_i} - s^{AP_i}| > \theta$)
        count = count + 1
    endfor
    if (count $\geq$ τ)
      remove neighbor m from neighbor list
  endfor
if (less than k entries in neighbor list)
  return old list
else
  return updated list
end
```

This enhanced NNSS technique is performed on the first k entries of the input neighbor list sorted according to Euclidean distance. Using cross-validation techniques, k = 3 was determined to be the optimal distinguishing parameter for boundary determination purpose.

5 Performance Analysis

The influence of the clustering module on computation time was investigated using a test area divided into a 10x10 grid under Wi-Fi coverage of 6 randomly placed APs shown in fig 4. Signal signature at each grid is recorded. By varying cluster size q, a series of location cluster were formed to reduce the search space.

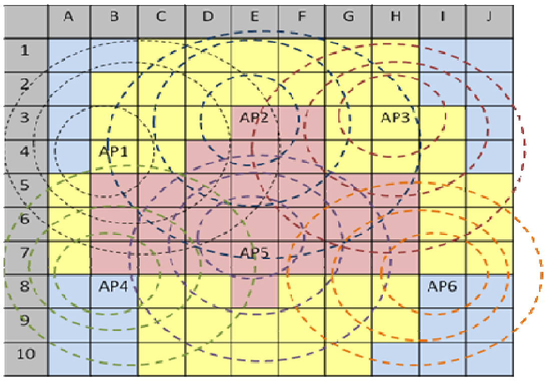


Fig 4 Wi-fi coverage of 6 randomly placed APs in test area.

TABLE II
COMPARISON OF q

	q=1	q=2	q=3	q=4	q=5
No. of clusters formed	100	12	9	3	0
Largest cluster coverage	N.A.	20%	13%	14%	N.A.
Smallest cluster coverage	N.A.	6%	6%	9%	N.A.
Average coverage	N.A.	11%	8.40%	11.3%	N.A.

From Table II, q = 3 is most effective at reducing the search boundary of the matching algorithm. This is, however, not enough to draw conclusion about the computational performance. Manual code instrumentation was applied to establish the runtime for each execution to evaluate the effect of cluster size and number of clusters. The software was executed repeatedly to determine the computation performance for various clusters. The program was adjusted such that only the tested cluster is executed.

Table III shows the average time taken for different type of clusters to complete the computation. Against the various possible q values, q=3 offers the best performance across different type of clusters except when cluster size is smallest.

TABLE III
AVERAGE TIME TAKEN TO COMPLETE CLUSTER COMPUTATION (ms)

	First Cluster	Last Cluster	Smallest Cluster	Largest Cluster
q=2	17.5	19.9	18.6	16.9
q=3	16.6	19.3	22.7	16.9
q=4	20.5	21.2	20.5	19.4

q is set at 3 to achieve the optimal performance and an average of 7% decrease in computation time is observed with the use of the clustering module. The performance gain is likely to be even more substantial in congested Wi-Fi environment due to the added benefit from clustering large number of APs.

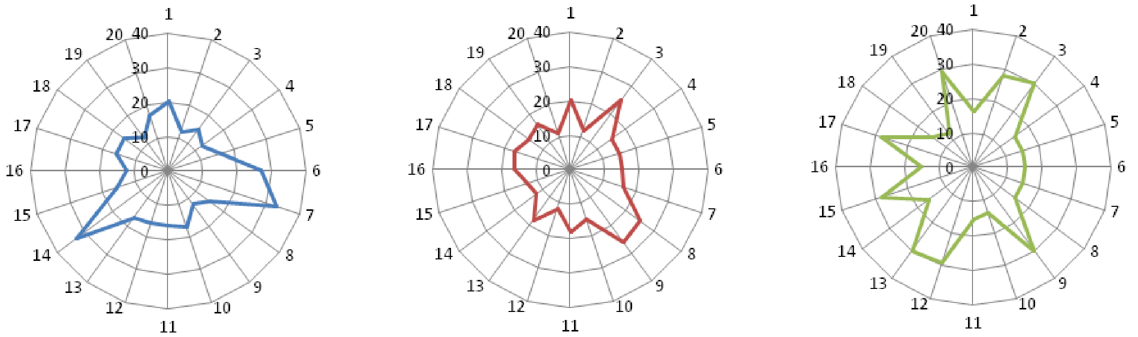


Fig 5 Influence of *q=2,3,4* (left to right) on computation time. The vertical axis represents the computation time in milliseconds (ms). Horizontal axis represents the amount of executed runs.

6 Conclusion and Future Works

The concept and architecture of the iSARIS platform is presented in this paper and its use is illustrated in a shopping mall scenario. The optimal localization techniques in different situations are determined by how radio wave transmissions are affected in both indoor and outdoor environment. A switching module toggle between primary and secondary localization services during the detection of a service region. Primary location service provided by cell identification based technique can quickly identify a coarse region of physical space. This space could be a building, a group of buildings or other entities where cell identification level of positioning performance is no longer acceptable. Secondary positioning service through the Wi-Fi infrastructural cluster enhanced *k*NN classifier provides additional positioning performance and is used sparingly to avoid resource draining in this shopping mall implementation. Further smart space segmentation in the form of service areas are created where independent service provider offers related online services. By automating the transition process using the iSARIS platform, the mobile offer a better and more seamless user experience. This location based service provisioning design offers a simple and practical approach to reduce service engagement time and is suitable for mobile environment where quick and efficient means of switching localization control is required.

Current software distribution model requires user to download the relevant applications in order to use a service. Majority of applications downloaded to a mobile device are often not used on a regular basis. To avoid this inefficiency, it is also our intention to further develop the iSARIS platform to implement a different application distribution model where software applications are only sent to user when they are needed. We hope to utilize the SA feature to determine a smart area where any application of interest to the user will be make available on the mobile when it enters the SA and removed the instant the mobile leaves the relevant SA. The possible benefit of this system is less storage as less applications need to be installed in the mobile and indirectly reducing the interface clutter experienced by user when too many applications are downloaded. A further area of work we are working on is to implement application recommender for each smart space using user participation tags.

7 Acknowledgement

This research work is supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research grant NRF2007IDM-IDM002-080.

8 References

[1] Castro P, Chiu P, Kremenek T and Muntz R, A Probabilistic Room Location Service for Wireless Networked Environments, presented at the *Ubicomp 2001: Ubiquitous Computing*, 2001, pp. 18-34.

[2] Barkhuus L, Brown B, Bell M, Sherwood S, Hall M and Chalmers M, From awareness to repartee: sharing location within social groups, presented at the *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, Florence, Italy, 2008, pp. 497-506.

[3] Trevisani E and Vitaletti A, Cell-ID Location Technique, Limits and Benefits: An Experimental Study, presented at the *Proceedings of the Sixth IEEE Workshop on Mobile Computing Systems and Applications*, 2004, pp. 51-60.

[4] Kunczier H and Anegg H, Enhanced cell ID based terminal location for urban area location based applications, *Consumer Communications and Networking Conference, 2004. CCNC 2004. First IEEE*, 2004, pp. 595-599.

[5] Hofmann-Wellenhof B, Lichtenegger H and Collins J, *Global positioning System. Theory and Practice*. 1993.

[6] McDonald K, The modernization of GPS: plans, new capabilities and the future relationship to Galileo, *Journal of Global Positioning Systems*, **2002; 1(1)**, pp. 1-17.

[7] Richton B, Vannucci G and Wilkus S, Assisted GPS for Wireless Phone Location — Technology and Standards. 2002, pp. 129-155.

[8] Bahl P and Padmanabhan VN, RADAR: an in-building RF-based user location and tracking system, *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2000, pp. 775-784 vol.2.

[9] Bahl P, Padmanabhan VN and Balachandran A, Enhancements to the RADAR User Location and Tracking System, *Microsoft Research*, **2000**.

[10] Bahl P, Padmanabhan VN and Balachandran A, A Software System for Locating Mobile Users: Design, Evaluation, and Lessons, *Microsoft Research Technical Report MSR-TR-2000-12*, **2000**.

[11] Yan Qiu C, Damper R.I and Nixon M.S, On neural-network implementations of k-nearest neighbor pattern classifiers, *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on* , vol.44, no.7, 1997, pp.622-629.

[12] Tran Q, Tantra J, Foh C, Tan A, Yow K and Qiu D, Wireless Indoor Positioning System with Enhanced Nearest Neighbors in Signal Space Algorithm, *IEEE Vehicular Technology Conference, 2006. VTC-2006 Fall*, 2006, pp. 1-5.

[13] Kelly D, Behan R, Villing R and McLoone S., Computationally tractable location estimation on WiFi enabled mobile phones, *Signals and Systems Conference (ISSC 2009), IET Irish* , 10-11 June 2009, pp.1-6.

[14] Jayaraman P, Zaslavsky A and Delsing J, Intelligent Processing of K-Nearest Neighbors Queries Using Mobile Data Collectors in a Location Aware 3D Wireless Sensor Network. *Trends in Applied Intelligent Systems, Lecture Notes in Computer Science*, 2010, pp. 260-270.

A Novel Weighted Clustering Algorithm Based On Mobility For Mobile Ad Hoc Networks

Amir Massoud Bidgoli[1], Mohammad Shayesteh[2]

[1] MIEEE, PHD Manchester University, Islamic Azad University-North Tehran Branch, am_bidgoli@iau-tnb.ac.ir

[2]Department Of Computer, Bandar abbas Branch, Islamic Azad University, Hormozgan, Iran, shayesth.au@gmail.com

ABSTRACT

In recent years, many studies have been conducted in Mobile Ad Hoc Networks field in order to make a virtual infrastructure consisting of nodes. The common goal of all was to select a node called clusterhead which guarantees relationships between nodes. In this paper, we have presented a new clustering algorithm in Mobile Ad Hoc Network based on nodes weight and their relative mobility. In this algorithm selection of a clusterhead is done during two stages. In the first stage, each node calculates its primary weight by using a new presented weighted function. In the second stage, each node calculates its relative mobility in present time and predict relative mobility in future to other neighborhoods. Then based on the primary weight of the specified neighborhood, it assigns a final weight to it. Next, based on the final weight clusterhead is selected. The goal of this algorithm is to decrease the number of cluster forming, maintain stable clustering structure and maximize lifespan of mobile nodes in the system. In simulation, the proposed algorithm has been compared with WCA, MOBIC and the Lowest_ID algorithm. The results of simulation reveal that the proposed algorithm achieves the goals.

Keywords: Mobile Ad Hoc Network, Clustering Algorithm, Relative Mobility, Weight.

1- INTRODUCTION

A MANET is a multi-hop wireless network in which mobile nodes can freely move around in the network, leave the network and join the network. These mobile hosts communicate with each other without the support of any preexisting communication infrastructure. Typically, if two nodes are not within mutual transmission range, they communicate through intermediate nodes relaying their messages. In other words, the communication infrastructure is provided by the nodes themselves. Through the nature of MANET, we have many challenges. The most important challenges are stability, routing and scalability. Clustering is the most way to improve the stability, routing and scalability. Have knowledge about the changes of node's status, can present useful information about the stability of it between its neighbors. This information is effective in clustering approach and in cluster head selection.

In wireless ad hoc network applications, such as outdoor teaching and the communications in the disaster area (the scenes of a fire, the flood, the earthquake and so on), a number of mobile hosts (MHs) are organized into several disjointed communication groups, which may move together and overlap with each other. Members within the same group have similar mobility patterns and can directly communicate with each other. Members of a group communicate with other nodes outside its group through the group clusterhead, which serves as a gateway to other groups. In the group mobility, the clusterhead equips with two network interfaces, one is used for local networks and the other is used for external networks. The local networks mean wireless ad hoc networks that are used in a group or between overlapped groups. The external networks denote the Internet, 2G, GPRS, and 3G…etc.

Clustering algorithms can be performed dynamically to adapt to node mobility [2]. MANET is dynamically organized into groups called clusters to maintain a relatively stable effective topology [1]. By organizing nodes into clusters, topology information can be aggregated. This is because the number of nodes of a cluster is smaller than the number of nodes of the entire network. Each node only stores fraction of the total network routing information. Therefore, the number of routing entries and the exchanges of routing information between nodes are reduced [3]. Apart from making large networks seem smaller, clustering in MANETs also makes dynamic topology appear less dynamic by considering cluster stability when they form [2]. Based on this criterion, all cluster members that move in a similar pattern remain in the same cluster throughout the entire communication session. By doing this, the topology within a cluster is less dynamic. Hence, the corresponding network state information is less variable [3]. This minimizes link breakage and packet loss.

Clustering is usually performed in two phases: clustering set-up and clustering maintenance. In the clustering set-up phase, clusterheads are chosen among the nodes in the network. The roles of clusterheads are coordinators of the clustering process and relaying routers in data packet delivery. After electing clusterheads, other nodes affiliate with its neighbor clusterhead to form clusters. Nodes which are not clusterheads are called ordinary nodes. After the initial cluster set up, reaffiliations among clusterheads and ordinary nodes are triggered by node movements, resulting reconfiguration of clusters. This leads to the second phase, the clustering maintenance.

As election of optimal clusterheads is an NP-hard problem [4], many heuristic clustering algorithms have been proposed [1-10]. To avoid excessive computation in the cluster maintenance, current cluster structure should be preserved as much as possible. however, any clusterhead should be able to change its role to an ordinary node to avoid excessive power drainage. In this way, the overall lifespan of the system can be extended.

The goal of this algorithm is to decrease the number of cluster forming, maintain stable clustering structure and maximize lifespan of mobile nodes in the system. To achieve these goals, we propose a new algorithm. In this algorithm, selection of a clusterhead is done during two stages. In the first stage, each node calculates its primary weight by using a new presented weighted function and then sends the obtained value to its neighborhoods. In the second stage, each node calculates its relative mobility in present time and predict relative mobility in future to other neighborhoods Then based on the primary weight of that specified neighborhood, it assigns a final weight to it. Next, based on the final weight clusterhead is selected. The result of simulation shows that the proposed algorithm provides better performance than WCA, MOBIC and Lowest_ID in terms of Clusterhead changes, Clusterhead lifetime and the average number of orphan clusters.

The rest of this paper is organized as follows. In Section 2, we review several clustering algorithms proposed previously. Section 3 presents the proposed algorithm for mobile ad hoc networks. The simulation of the proposed algorithm is given in Section 4. Finally, Section 5 concludes this paper.

2- RELATED WORK

A large number of clustering algorithm have been proposed according to certain environment and characteristic of mobile node in mobile ad hoc network to choose clusterhead . we will give each of them a brief description as follows:

1) Highest degree clustering algorithm [5] uses the degree of a node as a metric for the selection of clusterheads. The node with highest degree among its neighbors will be elected as clusterhead , and its neighbors will be cluster members. In this scheme , as the number of ordinary nodes in a cluster is increased , the throughput drops and system performance degrades.

2) The Lowest-Identifier algorithm(LID) [6] chooses the node with the minimum identifier (ID) as a clusterhead. The system performance is better than Highest-Degree heuristic in terms of throughput [4]. However, since this heuristic is biased to choose nodes with smaller IDs as clusterheads, those nodes with smaller IDs suffer from the battery drainage, resulting short lifetime span of the system.

3) Least Movement Clustering Algorithm[7]. In this algorithm, each node is assigned a weight according to its mobility. The fastest the node moves, the lowest the weight is. And the node with highest weight will be elect as clusterhead.

4) The Distributed Clustering Algorithm (DCA) [8] and Distributed Mobility Adaptive clustering algorithm (DMAC) [9] are enhanced versions of LID; each node has a unique weight instead of just the node's ID, these weights are used for the selection of clusterheads. A node is chosen to be a clusterhead if its weight is higher than any of its neighbor's weight; otherwise, it joins a neighboring clusterhead. The DCA makes an assumption that the network topology does not change during the execution of the algorithm. Thus, it is proven to be useful for static networks when the nodes either do not move or move very slowly. The DMAC algorithm, on the other hand, adapts itself to the network topology changes and therefore can be used for any mobile networks. However, the assignment of weights has not been discussed in the both algorithms and there are no optimizations on the system parameters such as throughput and power control.

5) MOBIC [7] uses a new mobility metric Instead of static weights; Aggregate Local Mobility (ALM) to elect clusterhead. ALM is computed as the ratio of received power levels of successive transmissions (periodic Hello messages) between a pair of nodes, which means the relative mobility between neighboring nodes.

6) The Weighted Clustering Algorithm (WCA) [4] is based on the use of a combined weight metric that takes into account several parameters like the node-degree, distances with all its neighbors, node speed and the time spent as a clusterhead. Although WCA has proved better performance than all the previous algorithms, it lacks a drawback in knowing the weights of all the nodes before starting the clustering process and in draining the clusterheads rapidly. As a result, the overhead induced by WCA is very high.

Most of previous algorithms were using only one metric for clustering purposes. Therefore, the resulted clustering topology fits just in terms of that specific metric [10]. As

Mobile Ad Hoc networks are generally complex and dynamic networks, existing of only one specific metric can not reference the whole situation of the network. Those types of clustering topologies which are optimal in terms of just one metric are suitable for particular senarios and have poor performance in other senario. For these reasons, we use different metrics in our algorithm to select the clusterhead. On the other hand, in clustering approaches based on weighted functions such as WCA, efforts are concentrated to select the best node among neighborhood nods by using of available metrics. In these methods, only those nodes are selected as clusterhead which have better properties than other neighborhood nodes such as more rest battery power, having more neighborhoods and less average distance from neighborhoods. but in most of them do not consider to the movement model of the clusterhead toward the other nodes of the cluster. This factor causes the formed clusters be unstable and increase the overload of cluster reelection process. While in approaches within move sensitive clustering category such as MOBIC, clustering is done based on the nodes movements or approaching/escaping to each other. in this approaches, the main parameter for clustering is the mobility of nodes and therefore other parameters such as the energy of the battery, the number of neighborhoods and so on, are not considered.

In the proposed approach we present an optimal method without appearing previous methods problems by combining the useful characteristics of those methods in order to reach to the goal of our algorithm.

3- The proposed Algorithm

In the algorithm presented in this paper, selection of a clusterhead is done during two stages. In the first stage, each node calculates its primary weight by using a new presented weighted function and then sends the obtained value to its neighborhoods. In the second stage, each node calculates its relative mobility in present time and predict relative mobility in future to other neighborhoods and base on the obtained result it can find out whether it gets closer to that neighborhood or escapes from it. Then based on the primary weight of that specified neighborhood, it assigns a final weight to it. Next, Next, based on the final weight clusterhead is selected.

3-1 Setup Procedure

First, we allocate IDs for the nodes. In the proposed algorithm, each node N_i (member or clusterhead) is identified by a state such as: N_i (id_{node} , id_{CH} , flag , $Weight_p$), it also has to maintain a 'node_table' wherein the information of the local members is stored. However, the clusterheads maintain another clusterhead information table 'CH_table' wherein the information about the other clusterheads and member node is stored.

In complex networks, the nodes must coordinate between each other to update their tables. The Hello messages are used to complete this role. A Hello contains the state of the node; it is periodically exchanged either between clusterheads or between each clusterhead and its members in order to update the 'CH_tables' and the 'node_tables' respectively.

We define a flag for every node which determine their role. The value of flag is 3 if the node is the clusterhead, is 1 if the node is an ordinary node, is 2 if the node is a gateway and is zero if the node has an undetermined status.

3-1-1 Weighted Function

To enhance stability of clusters we must find out problems that cause stability to be decreased and as a result cause a cluster to disappear. If we know and solve these problems, we can enhance stability of the clusters as much as possible.

The first parameter which causes clusters to disappear is Excessive battery consumption at a clusterhead. In MANETs, the nodes not only bear the responsibility of sending and receiving information, but also carry out routing for packages. As a result they consume a high rate of power.

As a result a clusterhead must have the following conditions:

- It must have a high existence of battery power.

- It must require a lower battery power for interaction with neighbors.

To meet the first condition, the amount of battery power is taken into account as one of the factors for calculation of weight. To meet the second condition, we can choose a node as a clusterhead, which has less distance with its neighbors during neighborhood duration (with using D_{AV}). $D_{AV}(i)$, is average distance between clusterhead i and its neighbors. The smaller $D_{AV}(i)$ is, the less transmission power the node i requires for interaction and communication with its neighbors and as a result it consumes less battery power.

The second parameter which causes the clusters be unstable is the mobility of nodes. In the proposed algorithm for creating stable clusters, in the first stage, the previous mobility of nodes intended, which is accessible by calculating parameter S. In the second stage, stable clusters are created through calculating relative mobility of nodes in present time and then predicting their relative mobility in future times.

The used parameters in weighted function for giving a primary weight to nodes ($weight_p$) include:

- cluster density (ρ): is defined as the ratio of the number of clusterhead neighborhoods to cluster diameter (d) (if the given node become clusterhead). Cluster diameter equals twice as much as the farthest node distance from the clusterhead. Cluster density provides us beneficial information about nodes aggregation around the clusterhead. The more cluster density, the more aggregation of clusters around the clusterhead and vice versa. We have to select such a node as the clusterhead which results in suitable density for the cluster. This can lead to optimal consumption of the battery and band width. The Equation (1) shows calculation the density of clusters by each node.

$$\rho = \frac{N}{d} \qquad (1)$$

the result is optimal is that the density in a normal range ($\alpha < \lambda < \beta$). To use a density parameter in the weighted function we use the Equation (2).

$$\rho_w = \begin{cases} \frac{\rho-\alpha}{\beta-\alpha} & \rho < \beta \\ \frac{\beta-\rho}{\beta-\alpha} & \rho > \beta \end{cases} \qquad (2)$$

- Battery remaining (B_r): every node which wants to be the clusterhead should have threshold power B_d. A clusterhead consumes more energy in a cluster comparing with an ordinary node. In addition, we prefer to choose a more powerful node to play its role as a clusterhead because such a node looses its energy later results in the late starting of new clusterhead selection process and therefore increases the stability of clusters. Equation (3) shows that the each node how to calculate battery remaining of itself.

$$B_r = \begin{cases} B_d + w_1 B_d & \rho < \propto \\ B_d + w_2 B_d & \propto < \rho < \beta \\ B_d + w_3 B_d & \rho > \beta \end{cases} \qquad (3)$$

- Number of nodes moving towards a node (N_{dm})

- Stability(S): The total time in which the neighbors of a specific node have spent their time beside the node. A Higher stability simply means that the neighbors of a certain node has spent a longer time in its transmission range, we conclude that the mentioned node has a more stable situation. The stability is used to address movements and adjacencies of nodes and is calculated by equation (4).

$$S = \sum_{i=1}^{n} T_{RF} - T_{RL} \qquad (4)$$

{n is the number of node's neighbors}

Where T_{RF} is the time of the first packet reception and T_{RL} is the time of the last packet reception.

- D_{AV} (i): is average distance between clusterhead i and its neighbors. The smaller D_{AV}(i) is, the less transmission power the node i requires for interaction and communication with its neighbors and as a result it consumes less battery power. by Equation (5) each node calculate its distance average with its neighbors (K is constant in the following formula).

$$D_{AV} = \frac{\sum_{i=1}^{N} \frac{k}{\sqrt{P_r}}}{N} \qquad (5)$$

Each node uses the above mentioned five parameters to calculate its primary weight (weight$_p$). The equation (6) shows how the nodes calculate weight$_p$.

$$Weight_p = c_1 B_r + c_2 N_{dm} + c_3 \rho_w + c_4 S + c_5 D_{AV} \qquad (6)$$

In the equation (6), c_i(s) are the weight factors of normalization.

3-1-2 The calculation of final Weight

After Each node calculate its primary weight, to calculate final weight should check its relative mobility comparing to its neighborhoods in the present time (R^N) and future time (R^F). Relative mobility indicates the variation of the nodes distance. Using of relative mobility enables us to recognize whether these two nodes are get closer to or escaping from each other or remains constant.

To calculate the relative mobility of two nodes A and B, the distance of these two nodes is calculated in two different times. This is done by calculating the power of successful signals received. D_{AB}^{t} stands for the distance between nodes A and B at time t which could be obtained by the equation (7). (K is constant in the following formula).

$$D_{AB}^{t} = \frac{K}{\sqrt{P_r}} \qquad (7)$$

And now we can calculate the present time relative mobility of nodes A and B by the equation (8) (between times t-1 and t):

$$R_{AB}^{N} = D_{AB}^{t} - D_{AB}^{t-1} \qquad (8)$$

In order to calculate the future relative mobility of these two nodes, (between times t, t+1) we should know the future location of them (in t+1). For that, we should predict the signal power received by nodes. For predicting the next signal power received by nodes, we use linear extrapolation method. In this method in order to obtain the signal power at the time t+1, we use the powers of signal at the times t , t-1 and t-2. Equation (9) shows how to calculate the power of a signal received at the time t+1 :

$$P_{t+1} = P_{t-1} + \alpha * \left(\frac{T_{t+1}-T_{t-2}}{T_{t-1}-T_{t-2}} * (P_{t-1} - P_{t-2})\right) + (1 - \alpha) * \left(\frac{T_{t+1}-T_{t-2}}{T_{t}-T_{t-1}} * (P_{t} - P_{t-1})\right) \quad (9)$$

After predicting the power of the received signal at the time t+1, we can calculate the future relative mobility of these two nodes by the equation (10):

$$R^{F}_{AB} = D^{t+1}_{AB} - D^{t}_{AB} \quad (10)$$

After predicting relative mobility in present and future times, the resulting values will represent following facts:

- If R^{N}_{AB} or R^{F}_{AB} were positive ($R^{N}_{AB}>0$ or $R^{F}_{AB}>0$) this indicates that these two nodes get close to each other now or in future.

- If R^{N}_{AB} or R^{F}_{AB} were negative ($R^{N}_{AB}<0$ or $R^{F}_{AB}<0$) this indicates that these two nodes escape from each other now or in future.

- If R^{N}_{AB} or R^{F}_{AB} were zero ($R^{N}_{AB}=0$ or $R^{F}_{AB}=0$) this indicates that the distance between these nodes remains constant now or in the future.

We can consider four possible statuses for each node and its neighborhoods based on the calculated values for R^{N}_{AB} and R^{F}_{AB}. Every status is shown as an ordered pair (R^{N}_{AB}, R^{F}_{AB}). These pairs represent the relative mobility of nodes A, B to each other in the present time and in the future times respectively. In all above mentioned statuses the value 1 indicates that the distance between these nodes gets closer or remains constant; the value zero indicates that these two nodes are escaping from each other. The possible four groups for the ordered pair function are as follow:

- Group 1 (1, 1): the two nodes are getting closer to each other in both present and future.

- Group 2 (0, 1): the two nodes are escaping from each other now but they will get closer to each other in future.

- Group 3 (1, 0): the two nodes are getting closer to each other now but they will escape from each other in future.

- Group 4 (0, 0) the nodes escape from each other both now and in future.

According the above mentioned classifying, each pair of nodes in a network lies in one of these groups. The best status for selecting a clusterhead is that both the member node and the clusterhead be within group 1; this indicates that the nodes get closer to each other both now and in future resulting to the stability of the formed cluster. The next priorities are groups 2,3 and 4 respectively. The worst status is group 4 because in this case the nodes escape from each other both now and in future so it is not good selection for stable clusters formation.

A node stores the values of neighborhood nodes according to the received massages from them. Then it calculates the relative mobility of now and future based on the power of the received signals and then defines its group with respect to all neighborhoods and finally calculates that nodes final weight comparing with itself, regarding to the primary value and group number.

For calculating the final weight according to group number, G factor is defined as follow:

- If two nodes with respect to each other are within group 1 , G=g1.

- If two nodes with respect to each other are within group 2 or 3, G=g2.

- If two nodes with respect to each other are within group 4 , G=g3

The above factors should select in such a way that always we have g1> g2 >g3 .

$$\text{Weight}_{F} = G*\text{weight}_{p} \quad (11)$$

Finally, if the primary weight of each node, is more than the final weight of all its neighbors and certain percentage of its neighbors to be nearing at it, the node declaration itself as clusterhead otherwise offer the node with the highest final weight is achieved between neighbors as clusterhead if the primary weight of this node is higher than primary weight of himself. Node do This action by sending a Best message to the desired node. Each node that receive best message from a certain percentage of its neighbors declaration himself as clusterhead. Q factor are determined This specific percentage. If at a certain time any node does not receive specific amount of the Best messages, then the node has more primary weight is selected as clusterhead.

In this way, we can select suitable clusterhead for all nodes not only by considering the suitable condition of clusterhead in the network but also with regarding to the node status comparing with the clusterhead. This leads to better local selection of the clusterhead resulting in more stable clusters creation preventing future cluster reform rippling.

Table 1 shows the messages with its description used in proposed algorithm.

Table 1. Messages used in the algorithm

Message	Description
Hello (id_{node} , id_{CH} , flag, $weight_p$)	To update the tables of the nodes
Best(id_{node} , id_{CH})	Offer the node to be a clusterhead
Join_request(id_{node} , id_{CH})	To affiliate a cluster
Join_accept(id_{node} , id_{CH})	The node accepts the welcome_ACK
CH_Wel(id_{node} , id_{CH})	The CH accepts a Join_Request
CH_NWel(id_{node} , id_{CH})	The CH rejects a Join_Request
CH_ACK(id_{node} , id_{CH})	The CH adds the node as a member
CH_info(id_{node} , id_{CH})	The CH accepts the presence of a new CH in the network
CH_change(id_{CH})	The CH notifies a CH change
Leave(id_{node} , id_{CH})	The node leaves the cluster

3.2 New Arrival Nodes Mechanism

Once a wireless node is activated, its id_{CH} field is equal to NULL since it does not belong to any cluster. The node continuously monitors the channel until it figures out that there is some activity in its neighborhood. This is due to the ability to receive the signals from other present nodes in the network. The node still has no stable state, thus its state is not full identified. In this case, it broadcasts a Join_Request in order to join the most powerful clusterhead. Thus, it waits either for a CH_Wel or for a CH_NWel.

When the entry node does receive neither CH_Wel nor CH_NWel . If this persists for certain number of attempts, the node declares itself as an isolated node, and restarts by broadcasting a new Join_Request after a period of time. We note that just the clusterheads may response by a CH_Wel or CH_NWel ; the ordinary members have to ignore any Join_Request received even if they are in the transmission range of the new entry node. This allows simplifying the management of the clusters.

In the case where the node receives a response (CH_Wel or CH_NWel), it does not take immediately any decision, this allows the node to be certain that it has received all the responses from all the neighboring clusterheads. The CH_Wel and CH_NWel messages do not indicate that the clusterhead has added the node to its table; they just signify that the clusterhead is waiting for a Join_Accept in order to add the node to its table. When the node receives multiple CH_Wels, Based on the primary weight of clusterheads calculate the final weight of them and select the node with highest final weight as the clusterhead. After that, it sends a Join_Accept to the chosen clusterhead and waits for CH_ACK from this CH. The CH_ACK has to contain a confirmation that the id_{node} has been added to the CH_table. Thus the node can fully-define its state. The reason that we use four ways to confirm the joining procedure is to prevent other clusterheads that they can serve the entry node to add this node to their tables and cause conflicts.

In the case where the node was just receiving CH_NWels, it considers these responses as rejection messages from the clusterheads. This may occur when the clusterheads are saturated and decide to reject the adhesion of new nodes. When the number of attempts reaches a certain value, the node prefers not to stay isolated, thus it declares itself as clusterhead.

3.3 Clusterhead Nodes Mechanism

A clusterhead has an id_{node} field is equal to id_{CH} field. As a clusterhead, the node calculates periodically its weight, thus it sends periodically Hello messages to its members and to the neighboring clusterheads in order to update the node_tables and CH_tables respectively. The clusterhead must monitor the channel for Leave, Hello and Join_Request messages. in the proposed algorithm this operation is limited to clusterhead to allow easier management on clusters.

When the clusterhead receives a Join_Request (id_{CH}=NULL) from a new arrival node or a Join_Request (full state) from a node which belongs to another cluster, the clusterhead can accept or reject the request basing on its capacities. This procedure gives more flexibility to the members by allowing them to leave a weak clusterhead and join another one which seems stronger than the current clusterhead. It may not be possible for all the clusters to reach the cluster size λ. We have tried to reduce the clusters formed by merging the clusters that have not attained their cluster size limit. However, in order not to rapidly drain the clusterhead's power by accepting a lot of new nodes, we define thresholds which allow the clusterhead to control the number of nodes inside its cluster.

The re-election is not periodically invoked; it is performed just in case of a higher received weight, it allows minimizing the generated overhead encountered in previous works As we explained above. the re-election may not result a new clusterhead, it depends on the stability of the new node for playing the clusterhead's role.

In the proposed algorithm clusterhead will check regularly incoming messages from neighboring nodes. if clusterhead received a message that contains higher primary weight from his weight, then it check the relative mobility with the desired node, if its relative mobility to this node were in the first group and all of the cluster members exist in neighboring of this node, assign clusterhead role to the desired node. The node do it with save the ID of this node in its CH_ID field. Then send a CH_info message to new clusterhead to declare that this node as a new clusterhead selected. Then copy their tables in to new clusterhead and send a CH_change message to neighboring nodes to defines a new clusterhead. in This new approach selecting the new clusterhead is based on stability of it in the cluster. In this case where a new clusterhead is elected, the procedure is soft and flexible in order not perturb the clusters while to copying the databases from the old clusterhead to the new clusterhead.

3.4 Member Nodes Mechanism

after joining a cluster, the node declares itself as a member of this cluster. Hence, it calculates periodically its weight and sends periodically Hello messages to its clusterhead. As a member, this node should just handle the Hello, the CH_change and the CH_info messages. This allows optimizing the resources (bandwidth, battery, etc) and minimizing the job of the nodes.

When the node receives a Hello from its clusterhead, the node has to update its node_table. When the node receives Hellos from the neighboring clusterheads, the node has the possibility to migrate to another clusterhead if there is a Hello which has a higher weight than the current clusterheads weight, Member node get this decision by calculating the final weight of the new clusterhead. it sends a Join_Request to the clusterhead which is Hello's source and continues as a member of the current clusterhead until the reception of CH_ACK. In this case, the node can send a Leave_Request to the last clusterhead. This method allows us to minimize the number of the formed clusters in the network.

When the node member receives a CH_info message as a result of the re-election procedure, thus it realizes that it is going to become the new clusterhead in the cluster. When a node member does not receive any message from its clusterhead, it considers that the clusterhead has gone brusquely down; in this case, the nodes have no choice and must restart the clustering setup procedure.

4- simulation and results

In this paper we use GloMoSim tool for simulation. The simulation environment is a Mobile Ad Hoc Network consists of 20 to 100 nodes in an 800*800 area. We assume that each node will be activated by a 2.4GHz radio frequency. The simulated area is considered as a two dimensional square and nodes move freely throughout the area. The movement of nodes has been simulated according to Random Waypoint model.

In order to evaluate the performance and efficiency of the proposed algorithm, a set of simulations were operated and duration of them was 1200 seconds. We select a set of parameters to show the efficiency of our algorithm. Our proposed algorithm was compared with WCA, MOBIC and Lowest_ID method which is the most famous clustering algorithms. These parameters include:

4-1 Clusterhead changes

Figure 1 , 2 and 3 shows that the clusterhead changes in the proposed algorithm less than the WCA, MOBIC and Lowest_ID algorithms that leads to long life of clusters so will have more stable clusters. The reason is that the proposed algorithm selected such as the node as the clusterhead that have more Presence and battery power therefore more time is left as clusterhead. Figure 1 shows the Average number of clusterhead changes against the speed of node, figure 2 shows the average number of clusterhead changes against the transmission range and figure 3 shows the count of clusterhead changes against the number of nodes.

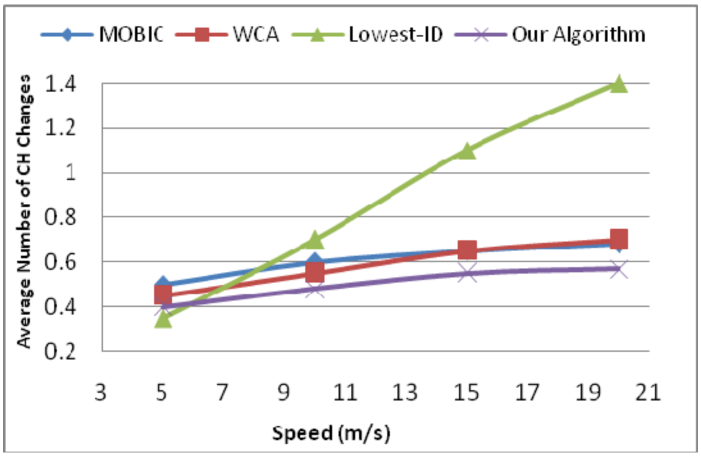


Figure 1. the Average number of clusterhead changes vs the speed of node

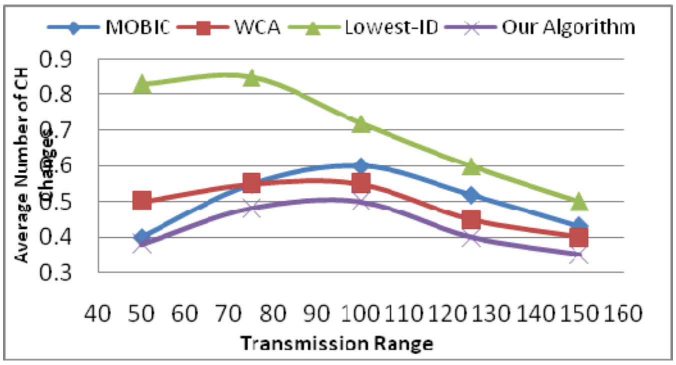


Figure 2. the Average number of clusterhead changes vs the transmission range

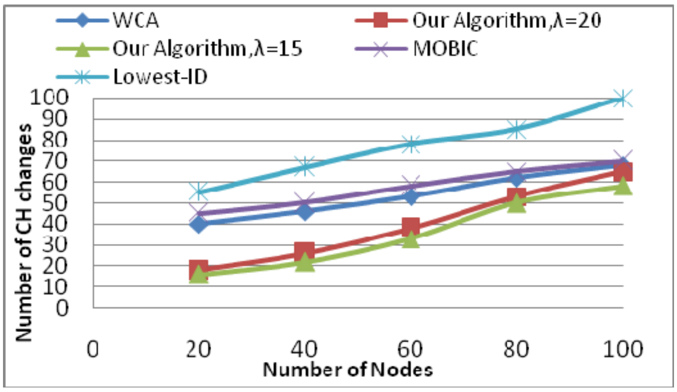


Figure 3. count of clusterhead changes vs the number of nodes

4-2 Clusterhead lifetime:

Figure 4 and 5 shows that the clusterhead lifetime in the proposed algorithm higher than the WCA, MOBIC and Lowest_ID algorithms. Figure 4 shows the average lifetime of clusterhead against the speed of node and figure 5 shows the average lifetime of clusterhead against the transmission range.

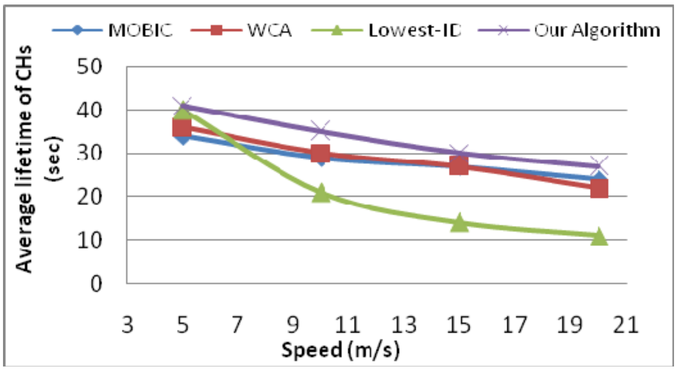


Figure 4. Average lifetime of clusterhead vs the speed of node

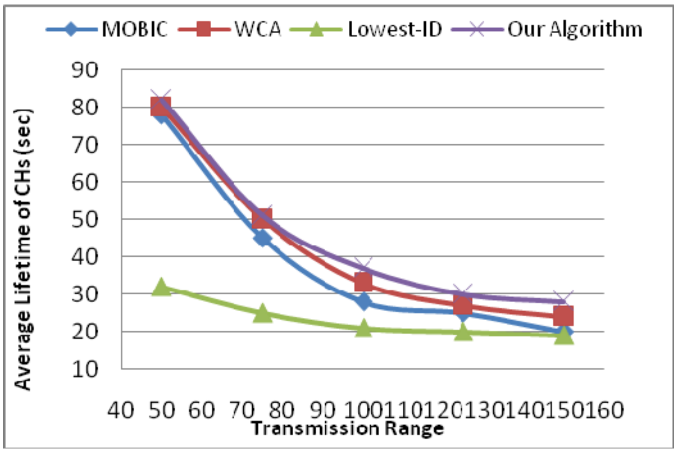


Figure 5. Average lifetime of clusterhead vs the transmission range

4-3 The average number of clusters:

As you can see in figure 6, the number of formed clusters is increased by increasing the number of nodes. Figure 6 shows the average number of clusters against the number of node.

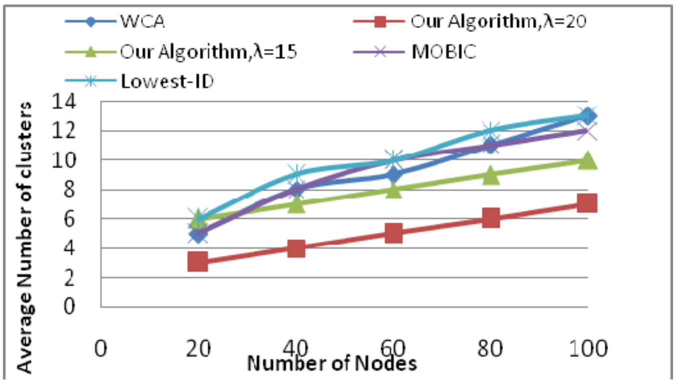


Figure 6. Average number of clusters vs the number of node

4-4 the average number of orphan clusters

As you can see in figure 7, the number of single node clusters or orphan clusters in our algorithm is less than WCA, MOBIC and the Lowest_ID algorithms. The reason is that in our algorithm, the clusterheads are selected from central safe area.

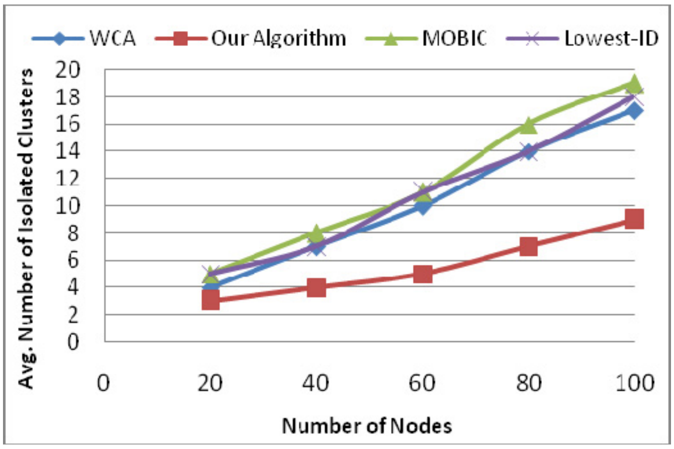


Figure 7. number of single node clusters or orphan clusters

5- conclusion

In this paper, we have presented a new clustering algorithm in Mobile Ad Hoc Network. In this algorithm selection of a clusterhead is done during two stages. In the first stage, each node calculates its primary weight by using a new presented weighted function. In the second stage, each node calculates its relative mobility in present time and predict relative mobility in future to other neighborhoods. Then based on the primary weight of that specified neighborhood, it assigns a final weight to it. Next, based on the final weight clusterhead is selected. A number of parameters of nodes were taken into consideration for assigning weight to a node. The proposed algorithm chooses the cluster-heads based on the information of neighbor nodes, and maintains clusters locally. Also it has a feature to control battery power consumption by switching the role of a node from a cluster-head to an ordinary node. We assumed a predefined threshold for the number of nodes to be created by a clusterhead, so that it does not degrade the MAC function and to improve the load balancing. We conducted simulation that shows the performance of the proposed enhancement clustering in terms of the average number of clusters formation, stability of clusters, and lifetime of a clusterhead. We also compared our results with the WCA, MOBIC and Lowest_ID . The simulation results show that our enhancement clustering algorithms have a better performance.

REFERENCES

[1] C. R. Lin and M. Gerla. Adaptive clustering for mobile wireless networks. IEEE Journal on Selected Areas in Communications, 15(7):1265-1275, Sept. 1997.

[2] A. B. McDonald and T. F. Znati. A mobility-based framework for adaptive clustering in wireless ad hoc networks. IEEE Journal on Selected Areas in Communications, 17(8):1466-1486, Aug. 1999.

[3] C. E. Perkins, editor. Ad Hoc Networking. Addison-Wesley, 2001.

[4] Chatterjee, M., Das, S., and Turgut, D., “WCA: a weighted clustering algorithm for mobile ad hoc networks,” Journal of Cluster Computing (Special Issue on Mobile Ad hoc Networks), 5, 2002, pp. 193-204.

[5] Gerla M., Tsai J. T. C., "Multicluster, Mobile, Multimedia Radio Network," ACM/Baltzer Wireless Networks Journal 95, vol. 1, Oct. 1995, pp. 255-265.

[6] Baker D.J., Ephremides A., "A distributed algorithm for organizing mobile radio telecommunication networks," Proceedings of the 2nd International Conference on Distributed Computer Systems, Apr. 1981, pp. 476-483.

[7] Basu P., Khan N., Little T. D. C.. A mobility based metric for clustering in mobile Ad hoc networks[A]. proceedings of IEEE ICDCS 2001 Workshop on Wireless Networks and Mobile computing[C], phoenix, A Z, April 2001:413-418.

[8] Basagni S., "Distributed clustering for ad hoc networks," Proceedings of International Symposium on Parallel Architectures, Algorithms and Networks, Jun. 1999, pp. 310- 315.

[9] Basagni S., "Distributed and mobility-adaptive clustering for multimedia support in multi-hop wireless networks," Proceedings of Vehicular Technology Conference, VTC, vol. 2, fall 1999, pp. 889-893.

[10] Hui Cheng, Jiannong Cao, Xingwei Wang, Sajal K. Das, Stability-based Multi-objective Clustering in Mobile Ad Hoc Networks, The Third International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, August 7–9, 2006, Waterloo,Ontario, Canada © 2006 ACM.

Host Mobility Management Across Heterogeneous Networks Using MPLS

Liren Zhang*, Hesham El-Sayed, Hadeel El-Kassabi
Faculty of Information Technology
United Arab Emirates University, Al Ain, UAE
lzhang@uaeu.ac.ae, helsayed@uaeu.ac.ae, htallat@uaeu.ac.ae

ABSTRACT

This paper presents a novel framework based on MPLS (multiple protocol label switching) protocol in support of high speed mobility management across multiple domains of heterogeneous mobile networks. The design of the proposed framework focuses on hierarchical architecture, multi-dimensional label distribution mechanism, scalable LSP (label switching protocol) configuration scheme and QoS (quality of service) management. Major performance measures in terms of handoff call blocking probability and new call blocking probability are evaluated by simulations under different traffic load scenarios. The numerical results obtained from simulations demonstrate that the proposed scheme is able to significantly improve the mobility and scalability in a networking environment with mixture of different types of communication platforms running on different protocols.

1. INTRODUCTION

Future wireless networks are expected to provide IP-based coverage and efficient mobility support with end-to-end QoS guarantees. Two enabling factors are considered as crucial: 1) maintaining the network connectivity during node mobility and 2) provisioning the network resources required by the Mobile Node (MN) in all the visited subnetworks. In recent years, many researchers have been focusing on protocols able to provide continuous network connectivity by accessing to multiple network domains. Mobile IP [1] is one of such protocols, which is able to provide a simple mobility management solution in IP-based wireless networks. According to Mobile IP, a mobile node (MN) can change its point of attachment without changing its IP address. To do so, a MN is assigned with a permanent home address in its home network, and will borrow a temporary care-of address (CoA) in each foreign network. The CoA is the foreign agent (FA) IP address of the currently visited foreign network. In this case, the home agent (HA), residing in the MN's home network, will maintain the mapping between the home address and CoA. Specifically, packets sent to MN are first intercepted by its HA, and then tunneled to the current serving FA using the MN's COA. The FA then decapsulates the packets and forward them to the MN. In turn, the packets transmitted by a MN are directly routed toward their destinations, without the need to pass through the sending MN's HA. Clearly, this routing approach has some deficiencies and induces long handoff latency and large signaling load when handoffs occur frequently [31]. For example, in Mobile IP, a mobile host needs to send an update message to its home agent for every subnet change. Thus, this may introduces significantly latency when handover occur frequently and registration messages may occupy significant valuable wireless communication bandwidth, especially when these registration messages travel long distances before packet redirection occurs. In this regard, many enhancements to Mobile IP for MNs with frequent handoffs have been proposed in the literature [10–11] to ensure service continuity. From the peer-to-peer communication point of view, information data need to be forwarded cross multiple wireless networking domains. The integration of such multiple protocols for networking certainly has the following disadvantages: (1) inefficiency due to large overhead for information data reformat in order to across different types of communication domains, (2) complicated routing process, which requires more memories, more computing power and large processing delay, and (3) no guaranteed end-to-end QoS because of the transmission status including available bandwidth, bit error rate, queuing buffer capacity, and the window size for packet forwarding are different in different communication domains. In fact, these disadvantages make the mobility management even more difficult. Furthermore, the cross layer design of such network including network capacity planning, traffic flow control, packet level scheduling, buffer dimensioning, bandwidth allocation and congestion control become challenge.

A number of solutions were proposed to provide seamless integration in different environments [12]. Vidales et al. propose PROTON to assist mobile users with multimode devices in the decision-making process related to roaming among heterogeneous technologies in 4G networks [13]. Siddiqui et al. propose a novel scheme to perform handoff decisions in a multi-access environment consisting of heterogeneous networking technologies [14]. They present a design architecture that efficiently implements automatic network selection based on several factors such as the conditions of accessible networks, requirements of active applications, and preferences of the mobile user. Luoto et al. propose a framework for improving the handover performance of Mobile IP based on a variety of cross-layer and cross-domain triggers [15]. Lo proposed architecture for mobility and QoS support in all-IP wireless networks [16]. Langar et al. propose a new a new protocol called Micro Mobile MPLS which alleviates the limitations of Mobile IP and in the same time benefits from MPLS resource provisioning capability [17].

To meet the requirements of next generation mobile networks, this paper presents an efficient networking protocol using MPLS [2], which aims to simplify the routing process with capabilities of mobility across multiple mobile heterogeneous networking domains. This paper is structured as follows. Section 2 describes the proposed framework and protocol for mobility management. Section 3 presents a case study of the proposed framework. Performance evaluation and analysis is presented in Section 4. Conclusion and future work are presented in section 5.

2. THE MOBILITY MANAGEMENT FRAMEWORK

In this paper, we consider a heterogeneous networks environment consisting of a number of mobile networks, which are operating with different configurations independently. As shown in Figure 1, the proposed middleware platform is based on a hierarchical architecture including access layer and administration layer. The access layer consists of multiple mobile access MPLS domains, where each of these mobile access MPLS domains is corresponding to an individual mobile network in the heterogeneous environment. The administration layer consists of a number of administration MPLS domains, which are hierarchically inter-connected by VGWs (visiting gateway) to match the requirement of scalability. Each mobile access MPLS domain has an AGW (access gateway), which is connected to an administration MPLS domain through a VGW (visiting gateway) [3] [12]. When a mobile host visits a new mobile access foreign MPLS administration domain, it sends a registration message with the IP address of the visited AGW to its home AGW. Upon receiving the registration message, the home AGW updates the forwarding table so that all the packets destined to the mobile host are rerouted to the corresponding AGW. VGW performs the mapping for downlink traffic and each BS (base station) performs the mapping for uplink traffic. Quality of service (QoS) based on differential service model are supported by LSPs between AGW and VGW, in which different QoS traffic classes are carried by separate LSPs. Likewise, within a mobile access MPLS domain, multiple LSP are created between base stations and AGWs in support of QoS. Software agents are distributed on access gateways including both AGWs and VGWs. These agents communicate among themselves and they collectively represent the middleware that controls the overall operations of heterogeneous network environments using LSPs. Within a mobile access MPLS domain, multiple LSPs are created between a base station and AGW to support multiple FECs that can be mapped to signaling traffic and user data traffic with different types of QoS. Distributing label-binding information among LSRs is accomplished using downstream label allocation mechanism [8]. LSPs are created by a middleware operator based on provisioning policies. The total number of LSPs in a mobile access MPLS domain is related to the number of traffic classes, since each LSP is shared by a number of multiple mobile hosts with the same QoS requirement. In order to make the LSP to be scalable, LSP is designed in a multiple-dimension format as shown in Figure 2. However, middleware operates on top of all gateways. It is responsible for coordinating gateways, assigning labels, updating gateways and supporting QoS operations. In initial stage, mobile host detects a nearest base station according to signal strength of beacon signals from near-by base stations (BSs) and then sends registration message to AGW via the closest base station and pre-established LSP between BS and AGW. When AGW receives the registration message, it updates its forwarding table by associating MN’s IP address with registration received interface. Then, AGW forwards the registration message to VGW (visiting GW) using pre-established LSP between AGW and VGW. VGW also updates its forwarding table by associating the IP address of the mobile into the mobile access MPLS domain with registration received interface. If the MPLS administration domain is not the home MPLS administration domain of the mobile host, VGW assigns a Virtual Care-off Address (VCoA) for the mobile host and registers it to the home AGW of the mobile host [3][9]. An LSP using CR-LDP is established between home AGW and visiting AGW to satisfy the QoS requirements of the traffic. The home AGW updates its forwarding table so that packets destined to the mobile host can be forwarded from the home AGW to the current visited AGW using established CR-LSP.

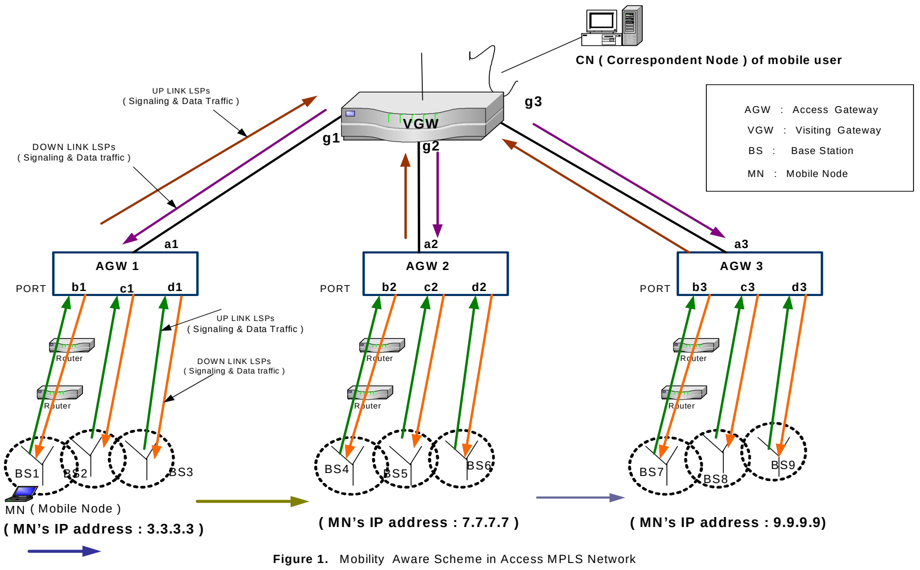
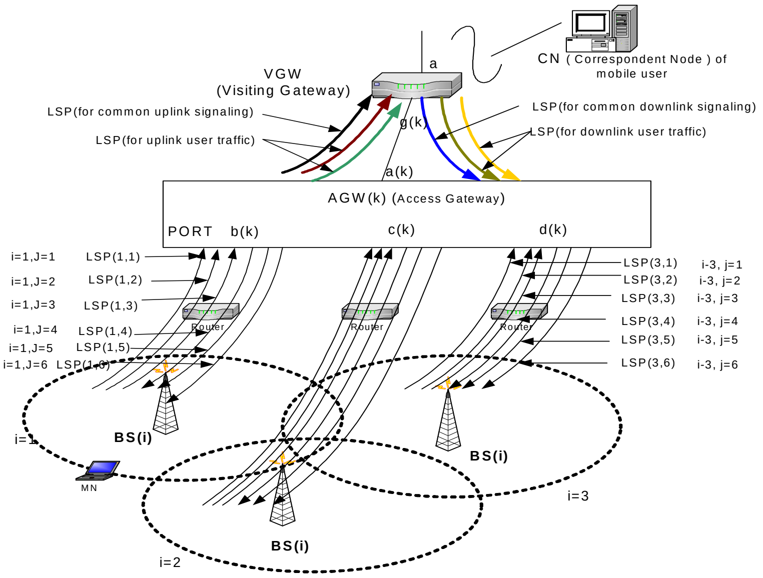

CN (Correspondent Node) of mobile user
UP LINK LSPs (Signaling & Data Traffic)
DOWN LINK LSPs (Signaling & Data traffic)
VGW
g1
g2
g3
AGW : Access Gateway
VGW : Visiting Gateway
BS : Base Station
MN : Mobile Node
a1
a2
a3
AGW 1
AGW 2
AGW 3
PORT
b1
c1
d1
b2
c2
d2
b3
c3
d3
UP LINK LSPs (Signaling & Data Traffic)
DOWN LINK LSPs (Signaling & Data traffic)
Router
BS1
BS2
BS3
BS4
BS5
BS6
BS7
BS8
BS9
MN (Mobile Node)
(MN's IP address : 3.3.3.3)
(MN's IP address : 7.7.7.7)
(MN's IP address : 9.9.9.9)

Figure 1. Mobility Aware Scheme in Access MPLS Network

Note that Virtual Care-off Address (VCoA) assigned to mobile host is valid as long as the mobile host is roaming within the administration domain. In terms of transport efficiency, IP-in-IP tunnels are not used between home AGW and visiting AGW. Instead, MPLS labels are employed for the same purpose. As illustrated in Figure 1, when a mobile host enters into the coverage area of new base station located in different MPLS access domain, the mobile host sends a registration request to the new AGW via the new base station. Then, this new AGW forwards the request to its VGW in the administration domain. If both home MPLS access domain and new access MPLS domain are under the same MPLS administration domain, in this case, mobile host is not required to be reregistered with its home agent. Therefore, the benefits of such procedure include less delay, less packet loss during handoff and eliminating registration between mobile hosts and home agent. Eliminating registration is able to reduce the signaling load experienced by core network in support of mobility and provides scalability for the mobile access network.

3. CASE STUDY

As shown in Figure 2, for illustrative purpose, six LSPs are pre-established between BS1 and AGW1, where LSP (1,1), LSP(1, 2) and LSP(1, 3) are assigned as uplinks and LSP(1, 4), LSP(1, 5) and LSP(1, 6) are assigned as downlinks. In BS1, LSP (1, 1) is used for common uplink signaling and LSP (1, 4) is used for common downlink signaling. In this case, LSP (1, 2) and LSP (1, 3) are used for uplink user traffic where as LSP(1, 5) and LSP(1, 6) are used for downlink user traffic. To support QoS on DiffServ basis[4], LSP (1, 2) and LSP (1,5) are mapped as to Assured Forwarding (AF) for uplink and downlink, respectively and LSP(1, 3) and LSP(1,6) are mapped as Expedited Forwarding (EF) for uplink and downlink, respectively. When a mobile host named MN enters the access MPLS domain of base station 1 (BS1), it sends a registration message to BS1. BS1 updates its forwarding table correspondingly with MN's IP address and also forwards the registration message to AGW1 through the common uplink signaling LSP(1, 1). Upon receiving of the registration message, AGW1 updates its routing table with MN's IP address and sends a registration message to VGW. Finally, VGW updates its routing table and sends the registration message to MN's home AGW if current access MPLS domain is not the MN's home access MPLS domain. Therefore, data packets destined to MN can be forwarded to the corresponding BS1 correctly through established LSPs via VGW and AGW. Note that MN's IP address in the forwarding tables at VGW, AGW and BS1 is maintained as soft-state and it is refreshed by a periodic route update message from MN. When VGW receives data packets with QoS class of AF from a remote host, named as CN to mobile host MN located in BS1, VGW assigns an appropriate label to the packets and sends them onto the corresponding LSP forwarding to AGW1. At the same time, VGW also updates its forwarding table accordingly. Hence, MN's IP address is associated with AGW1 and corresponding LSP 60. When these labeled packets arrive AGW1, label switching is performed at AGW1 using LSP(1, 5) as shown in Table 1. Such label switching from LSP 60 to LSP(1, 5) is performed by examining the label header and PHB (AF) associated with the label to maintain the QoS assigned to packets. Furthermore, when these packets arrive at BS1 through LSP(1, 5), label header is removed from packet and then is forwarded to mobile host MN. Likewise, when mobile host MN sends a packet to remote host CN, BS1 performs mapping of the QoS class for the packet onto appropriate LSP.



LSP(i,j) : where i represents cell & BS number, i={1,2,3,… … .,n} and

j represents uplink & downlink LSP between BS(i) and AGW(k) , j={1,2,3,… ..,n} and

k={1,2,3,… ,n}

Figure 2. Multiple Dimension LSP Configuration with Mobility Aware MPLS

Table 1. Forwarding table of AGW1 after packet arrival from VGW

Input		FEC	Output		Per-hop behavior (PHB)	Associated base station BS1
I/F	Label		I/F	Label		
b1	1	-	a1	-----	-	LSP(1, 1)
b1	2	-	a1	-----	AF	LSP(1, 2)
b1	3	-	a1	-----	EF	LSP(1, 3)
a1	-		b1	4	-	LSP(1, 4)
a1	**60**	**3.3.3.3**	**b1**	**5**	**AF**	**LSP(1, 5)**
a1	-	-	b1	6	EF	LSP(1, 6)

Handoff Procedure

When mobile host MN moves into the coverage area of base station BS2, it initiates handoff from BS1 to BS2 by sending a route update message to BS2. Then, BS2 forwards the route update message to AGW1 through LSP(2, 1) (common uplink signaling LSP between BS2 and the AGW1). AGW1 updates its forwarding table so that MN's IP address can be associated with LSP(2, 5) as shown in Table 2. After handoff, packets from remote host CN to MN are forwarded to BS2 through LSP(2, 5). Since PHB contained in packets are maintained the same format in the handoff process, thus, the same QoS is maintained after handoff. Likewise, the same handoff procedure is performed at AGW1 when MN moves from BS2 to BS3. Therefore, the proposed handoff procedure is able to provide lower handoff delay to the traffic heading to the MN during the handoff period. This technique leads to much better scalability than an approach based on Mobile IP. Maintaining the IP address of the MN unchanged across movements within the same Mobile Access domain results in straight forward support for QoS.

Table 2. Forwarding table of AGW1 after handoff from BS1 to BS2

Input		FEC	Output		Per-hop behavior (PHB)	Associated base station BS1&BS2
I/F	Label		I/F	Label		
b1	1	-	a1	-----	-	BS1 [LSP(1, 1)]
b1	2	-	a1	-----	AF	BS1 [LSP(1, 2)]
b1	3	-	a1	-----	EF	BS1 [LSP(1, 3)]
a1	-		b1	4	-	BS1 [LSP(1, 4)]
a1	**60**	**---**	**b1**	**5**	**AF**	**BS1 [LSP(1, 5)]** ***(before handoff to BS2)***
a1	-	-	b1	6	EF	BS1 [LSP(1, 6)]
…	…	…	…	…	…	…
c1	7	-	a1		-	BS2 [LSP(2, 1)]
c1	8	-	a1		AF	BS2 [LSP(2, 2)]
c1	9	-	a1		EF	BS2 [LSP(2, 3)]
a1	…	…	c1	10	-	BS2 [LSP(2, 4)]
a1	**60**	**3.3.3.3**	**c1**	**11**	**AF**	**BS2 [LSP(2, 5)]** **(after handoff to BS2)**
a1			c1	12	EF	BS2 [LSP(2, 6)]
…..	……	…..	…..	…..	…..	…..

4. PERFORMANCE EVALUATION

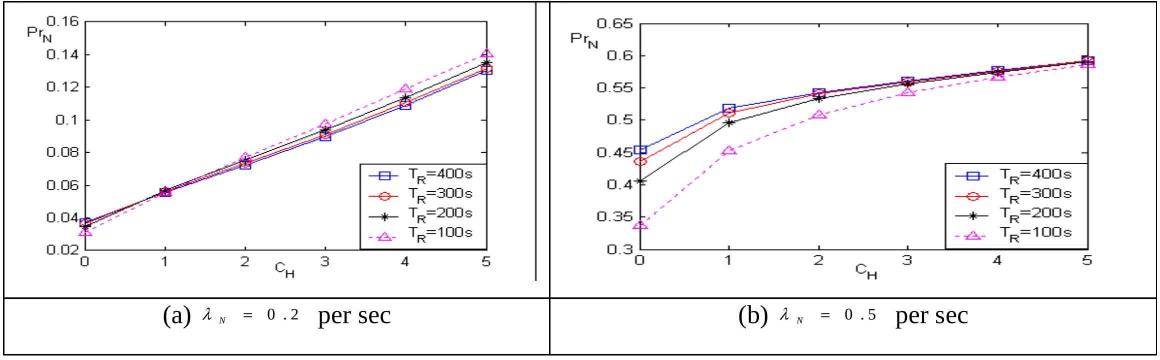
In this section, the major performance metrics are investigated in terms of new call blocking probability and handoff blocking probability, which are denoted by $\Pr_N$ and $\Pr_H$, respectively. The new call blocking probability $\Pr_N$ is defined as the probability that a new arrival call is blocked due to unavailability of mobile channels. The handoff blocking probability $\Pr_H$ is defined as the probability that an existing call is eventually not completed due to lack of mobile channel when handoff attempt is requested. However, in the proposed LSP configuration scheme, the handoff connection request is given higher priority than the new call connection request to improve handoff related system performance that a certain number of mobile channels are reserved for handover exclusive use. The performance of the LSP configuration scheme based on MPLS protocol is evaluated by simulation experiments. The simulation system is shown in Figure 2, where each BS is able to provide sufficient mobile channels for the existing mobile connections and the size of packets generated by a mobile host is identical. Either uplink or downlink between BS and AGW can simultaneously accumulates the bandwidth in support of up to C_N mobile calls, among which the bandwidth for C_H mobile calls is exclusively assigned for handoff and the remaining bandwidth for $C_N - C_H$ mobile calls are shared by both new arrival calls and handoff calls. When a new call arrives, if the number of existing mobile connections is greater than $C_N - C_H$, then this new arrival call is blocked. On the other hand, a handoff attempt is unsuccessful if there is no bandwidth available between the targeted BS and AGW.

Table 3. Simulation Parameters

Items	Value
C_N	30
C_H	1~5
T_C	120s
T_R	100~400s
λ_N	0.2~0.5 per sec

A discrete simulator is developed using the handoff simulation model [5], where new call arrival is implemented as a Poisson process with an average rate λ_N. The call holding time is exponentially distributed with mean value of T_C, and residence time of each mobile host is also implemented as an independent random process with exponential distribution at a mean value of T_R, which can be used to indicate the mobility degree of mobile hosts. The simulation parameters are shown in Table 3. All statistics of new call blocking probability and handoff blocking probability are estimated using the data obtained from M independent simulation runs. Since each simulation run depends on a particular stream of pseudo-random numbers to drive the simulation process, where the pseudo-random stream is generated by computer using a given random seed number, the obtained results may typically vary from one simulation run to another. To ensure the accuracy of the simulation results, the confidence intervals are calculated using independent replication method [6] as follows:

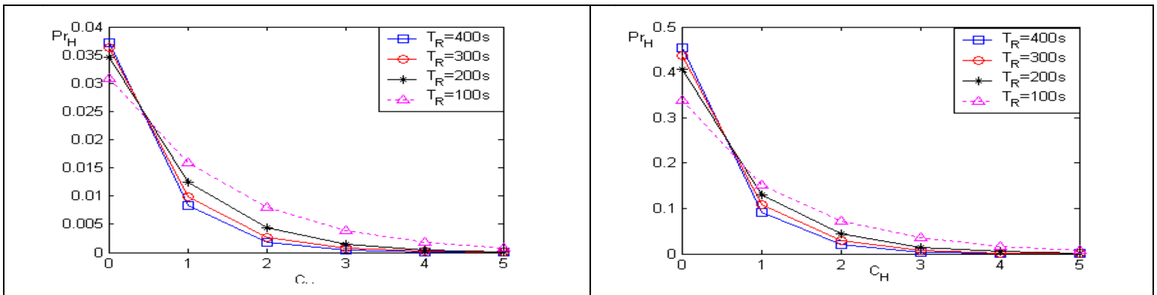
1. The simulation is independently repeated M = 30 times, and M groups of data are thus obtained. Each simulation run has a length of 10^8 slots excluding a warm up period of 1000 slots, which is setup to ensure that the results are obtained on the basis of a stable simulation process, because of each simulation run starts with an empty network environment. Actually, before the 1000 slots are decided as the warm up period, it has compared the results, which were obtained from the simulations runs using different warm up periods of 100, 500, 1000, 2000 and 5000 slots, respectively. The comparison has shown that a warm up period of 1000 slots or more is reasonable adequate.
2. The confidence interval δ is calculated as that at least 90 percent of estimated values for ρ obtained from these M simulation runs fall in the interval (ρ – δ, ρ + δ), where (δ << ρ).
3. We note that this performance evaluation is mainly devoted to the statistics of network mobility under high traffic density condition during the busiest hours of the day. This is feasible for relatively high new call blocking probability and high handoff call blocking probability. In this case, the confidence intervals of the simulations are not difficult to be implemented. However, when these blocking probabilities are low, the simulations require an excessive amount of computing resources and the confidence interval becomes more difficult. The simulation experiments conducted for this research has not addressed that region.



(a) $\lambda_N = 0.2$ per sec (b) $\lambda_N = 0.5$ per sec

Figure 3. New call blocking probability versus number of channels reserved for handoff call

Figure 3 and Figure 4 show the performance statistics in terms of new call blocking probability and handoff blocking probability under different traffic load conditions. Figure 3 shows new call blocking probability $\Pr_N$ versus C_H, which is the number of bandwidth exclusively assigned for handoff calls in a cell. It can be seen that $\Pr_N$ increases with C_H when the average cell residence time T_R has a fixed value. This is because that the available bandwidth for new call decreases with the increase of C_H. Moreover, the offered traffic load changes when the new call arrival rate λ_N changes. For example, as shown in Figure 3(a), when the traffic load is $\lambda_N = 0.2$ per second, $\Pr_N$ increases with C_H following an approximately linear rate. However, when $C_H \geq 2$, $\Pr_N$ increases with the decrease of residence time T_R. On the other hand, for the traffic load case of $\lambda_N = 0.5$ per second as shown in Figure 3(b), when C_H increases, the increase rate of $\Pr_N$ is much less than that for the traffic load case $\lambda_N = 0.2$ per second. However, $\Pr_N$ increases with the increase of T_R. The difference between the values of $\Pr_N$ for variable T_R decreases with the increase of C_H. For example, when $C_H = 0$, $\Pr_N$ for $T_R = 400s$ is 10% higher than that for $T_R = 100s$; when $C_H = 5$, the values of $\Pr_N$ for $T_R = 400s$ and $T_R = 100s$ are almost the same. Comparing Figure 3(b) to Figure 3(a), it can also be found that new call blocking probability significantly increases with λ_N when both T_R and C_H are kept as fixed values. Figure 4 shows handoff blocking probability $\Pr_H$ versus C_H for different values of residence time T_R. The curves of $\Pr_H$ shown in Figures 4(a) and (b) have the similar trend with the change of C_H. This means that, for both the traffic case of $\lambda_N = 0.2$ per second and the traffic load case of $\lambda_N = 0.5$ per second, $\Pr_H$ always decreases with the increase of C_H. This is because that the bandwidth reserved for handoff calls increases with C_H. Comparing Figure 4 with Figure 3, it can be found that for fixed values of T_R andλ_N, $\Pr_H$ has the same value of $\Pr_N$ when $C_H = 0$, since both handoff calls and new calls are treated same in this case. However, when $C_H \ge 1$, $\Pr_H$ is much smaller that $\Pr_N$. For example, when$T_R = 100s$, $C_H = 4$and $\lambda_N = 0.5$ per sec, the value of $\Pr_N$ is more than 25 times of $\Pr_H$. Moreover, it is obvious that there is tradeoff between handoff blocking probability and new call blocking probability. In order to reduce$\Pr_H$, we can increase the value of C_H, but $\Pr_N$ increases accordingly. From Figure 4, the study shows that $\Pr_H$ increases with the decrease of T_R when $C_H \ge 1$, i.e., the increase of mobile host moving rate will increase handoff blocking probability and the number of forced call terminations.



(a) $\lambda_N = 0.2$ per sec

(b) $\lambda_N = 0.5$ per sec

Figure 4. Handoff call blocking probability versus number of channels reserved for handoff call

Table 4. Simulation results of handoff call blocking probability ($\Pr_H$) for load traffic $\lambda_N = 0.2$ per sec

number of BW assigned for handoff in the cell	cell residence time of mobile node (T_R)			
	T_R = 100s	T_R = 200s	T_R = 300s	T_R = 400s
0	3.08E-2±2.10E-3	3.46E-2±1.58E-3	3.62E-2±1.88E-3	3.71E-2±1.63E-3
1	1.59E-2±2.34E-3	1.24E-2±1.19E-3	9.95E-3±9.25E-4	8.27E-3±7.41E-4
2	8.03E-3±7.41E-4	4.36E-3±3.24E-4	2.685E-3±1.59E-4	1.811E-3±1.27E-4
3	3.932E-3±2.52E-4	1.49E-3±2.41E-4	7.07E-4±6.50E-5	3.87E-4±2.28E-5
4	1.867E-3±1.46E-4	4.96E-4±3.85E-5	1.81E-4±2.84E-5	8E-5±7.52E-6
5	8.56E-4±6.83E-5	1.59E-4±2.64.E-5	4.5E-5±3.35E-5	1.6E-5±1.24E-6

Table 5. Simulation results of handoff call blocking probability ($\Pr_H$) for load traffic $\lambda_N = 0.5$ per sec

number of BW assigned for handoff in the cell	cell residence time of mobile node (T_R)			
	T_R = 100s	T_R = 200s	T_R = 300s	T_R = 400s
0	3.38E-1±1.91E-2	4.06E-1±3.18E-2	4.36E-1±3.13E-2	4.53E-1±3.92E-2
1	1.52E-1±1.32E-2	1.29E-1±1.51E-2	1.08E-1±1.22E-2	9.22E-2±9.5E-3
2	7.26E-2±6.82E-3	4.35E-1±3.74E-2	2.82E-2±1.52E-3	1.96E-2±2.11E-3
3	3.49E-2±2.55E-3	1.46E-2±2.21E-3	7.22E-3±6.62E-4	4.06E-3±3.35E-4
4	1.65E-2±1.45E-3	4.74E-3±3.53E-4	1.79E-3±2.79E-4	8.11E-4±7.46E-5
5	7.57E-3±6.33E-4	1.485E-3±2.75E-4	4.25E-4±3.64E-5	1.55E-4±1.41E-5

The simulation results clearly indicate the advantage that the LSP scheme offers lower handoff blocking probability (i.e., significantly improves probability of successful handover) under different traffic load scenarios. The performance evaluation concerning new call blocking probability show that the new call blocking probability largely increases with load traffic λ_N when both T_R and C_H are kept as fixed values.

The dropout performance of new call can be enhanced by reserving dynamically adjustable number of channels exclusively for handoff requests. Such approach requires the determination of an optimum number of guard channels, acknowledgement of the traffic pattern in the area, the existing connections in the neighboring radio covering zone, and estimation of the channel occupancy time distributions. This will enable the BS to approximately reserve the actual amount of resources for handoff requests and thereby accept more new calls as compared to a fixed scheme. Another alternative solution to improve the dropout performance significantly is the combination of reserving channels for handoff and queuing of handoff requests [7] but it is more suitable to base station with a large radio covering range because of queuing new calls result in increased handoff blocking probability and the effectiveness of queuing decreases for small radio covering range.

6. CONCLUSION

This paper presented a new framework that deploys hierarchical MPLS architecture and scalable LSP configuration approach in support of mobility and QoS management across multiple heterogeneous network domains. The total number of LSPs in a mobile access MPLS domain is related to the number of QoS classes supported, since each LSP is shared by multiple mobile hosts with the same QoS class. In order to make the scheme be scalable, the MPLS label switched path (LSP) is designed in multiple dimension format. The major attention focuses on the performance evaluation by simulation experiments focusing on network mobility in terms of new call blocking probability and handover blocking probability, which are the most important technical terms being commonly used for the performance evaluation in mobile networks. Handoff priority implemented as guard channels is used to improve system performance. The simulation results clearly indicate the advantage that the proposed framework using LSP concept is flexible, scalable and efficient in support of mobile host handoff across heterogeneous network domains. After handoff, continuous QoS can be maintained by using the pre-established LSP between BS and AGW. The scheme does not require re-registration between mobile host and AGW when mobile host moves from bases station to base station as long as the mobile host remains inside of this middleware platform coverage area. On the other hand, this eliminating registration is able to reduce the overhead for handoff and QoS management between mobile host, BS and AGW, which is an important advantage able to increase network scalability and efficiency in support of large volumes of mobile hosts.

REFERENCES

[1] C. Perkins, ed., “IP Mobility Support”, IETF RFC 2002, Oct 1996.

[2] E. Rosen, A. Viswanathan, R. Callon, “Multiprotocol Label Switching Architecture “, IETF RFC 3031, Jan 2001.

[3] Z. Ren, C.Tham and C. Ko, “Integration of Mobile IP and Multi – Protocol Label Switching“ ICC2001, June 2001.

[4] D. Black et al. “An Architecture for Differentiated Services” , 1998. Internet RFC 2475.

[5] Y. Lin, S. Mohan, and A. Noerpel, "Queuing priority channel assignment strategies for PCS hand-off and initial access," IEEE Trans. on Vehicular Technology, vol. 43, no. 3, pp. 704-712, Aug. 1994

[6] T.G. Robertazzi, “Computer Networks and Systems: Queueing theory and Performance Evaluation”, Springer-Verlag, New York Inc., 1990

[7] D. Giancristofaro, M. Ruggieri and F. Santucci, “Queuing of Handover Requests in Microcellular Network Architectures“, Proc. 44st IEEE VTC, May 1994, pp. 1846-49.

[8] Asgari, A.; Egan, R.; Trimintzios, P. and Pavlou, G., “Scalable monitoring support for resource management and service assurance,” IEEE Networks, Vol. 18, No.6, , pp.6 – 18, 2004

[9] Howarth, M.P.; Flegkas, P. and Pavlou, G., “Provisioning for interdomain quality of service: the MESCAL approach”, IEEE Communications Magazine, Vol.43, No.6, pp. 129-137, 2005.

[10] K. El Malki, “Low latency handoffs in Mobile IPv4,” IETF, RFC4881, June 2007.

[11] E. Fogelstroem, A. Jonsson, and C. Perkins, “Mobile IPv4 regional registration,” IETF, RFC4857, June 2007.

[12] A. Macriga, P. Kumar, “Mobility Management for Seamless Information flow in Heterogeneous Networks Using Hybrid Handover”, IJCSNS Int. Journal of Computer Science and Network Security, VOL.10 No.2, Feb. 2010

[13] Vidales at el., “Autonomic System for Mobility Support in 4G Networks,” IEEE Journal on Selected Areas in Communications, Vol. 23, No. 12, pp. 2288- 2304, December 2005.

[14] F. Siddiqui, S. Zeadally, H. El-Sayed and N. Chilamkurti, “A Dynamic Network Discovery and Selection Method for Heterogeneous Wireless Networks”, The Int. Journal of Internet Protocol. Technology. Vol. 4, No.2, 2009

[15] M. Luoto, T. Sutinen, “Cross-Layer Enhanced Mobility Management in Heterogeneous Networks”, IEEE International Conference on Communications, pp. 2277-2281, May 2008.

[16] Lo, “Architecture for Mobility and QoS Support in All-IP Wireless Networks,” IEEE Journal on Selected Areas in Communications, Vol. 22, No. 4, pp. 691-705, May 2004.

[17] R. Langar, N. Bouabdallah, and R. Boutaba, “A Comprehensive Analysis of Mobility Management in MPLS-Based Wireless Access Networks”, IEEE/ACM Transactions on Networking, Vol. 16, No. 4, Aug. 2008

Based on Image Twist Technology to Create Funny Mobile Animation Characters

Peng Ge-gang, Li Xin- yu, Song Ying, Xiang Li- sheng, Shen Qing, Li Ren-fa
Hunan University and Talkweb Information System CO. LTD, Hunan, China，410205

Abstract: Under the powerful support from government, together with the progress of 3G business, 3G terminals, and the popularity of the animation product itself, Chinese mobile animation consumer market makes a rapidly progress. Mobile multimedia animations are some what different from the conventional "decent" cartoon films or televisions. These mobile products often work with light comedy to "open the window and lightly breathe". By re-creation a funny shape of who personal character to increase the entertainment atmosphere and meet the "debris-style refreshing after work". This paper proposes a way to fit the target above. First applied sine function as parameter to design a nonlinear scaling template, and then project a small un-rectangular grid of the source image to a linear rectangular grid to build a twist image. By using this non-linear stretch and project, a normal picture may be re-created as a funny cartoon character and meet different users' enjoy hobbies.

Key words: Mobile Cartoon；Image Stretch；Digital Image Processing

1、Introduction

Under the powerful support from government, together with the progress of 3G business, 3G terminals, and the popularity of the animation product itself, Chinese mobile animation consumer market makes a rapidly progress.

China has more than 800 million mobile phone users, even if only 5% of them use mobile animation, this market size is very optimistic indeed. Data show that mobile phone users in the country, 35 young people under the age of majority. This group includes students, fashion youth and white-collar workers; many of them are loyal fans of animation.

These mobile products often work with light comedy to "open the window and lightly breathe". By re-creation a funny shape of who personal character to increase the entertainment atmosphere and meet the "debris-style refreshing after work" greatly. [1, 2, 3]

This paper proposes a way to fit the target above. First applied sine function as parameter to design a nonlinear scaling template named as a twist template. And then project a small un-rectangular grid of the source image to a linear rectangular grid to build a twist image. The first grid above is in the twist template, and it is located by the X, Y Coordinates of the left-top and right-bottom corners of this grid. By using this non-linear stretch and project, a normal picture may be re-created as a funny cartoon character and meet different users' enjoy hobbies.

2、Technical Procedure

2.1 Design a nonlinear template

Step1. Face to an image with a width of W and height of H, we design an L lines and C columns rectangular (linear) grid template named as SquareTemplet. The length of square grid is S, so we have:

$$C = ((W / S) + 1) / S * S$$

$$L = ((H / S) + 1) / S * S$$

¶ This research was supported in part by National Technical Supporting Project Foundation of China (2007BAH14B04)

The template records the X and Y coordinates of each grid left-top endpoint as:

$$SquareTemplet[i*C+j].x = j*S$$

$$SquareTemplet[i*C+j].y = i*S$$

Here i and j is the series number of the line and column of this grid in square template.

Step2. Design a twist template for the same image. This template is a nonlinear grid template named as TwistTemplet below. It records the X and Y coordinates of each grid left-top endpoint as:

$$TwistTemplet[i*C+j].x = j*S+dx$$

$$TwistTemplet[i*C+j].y = i*S+dy$$

Here,

$$dx = Xamplify*\sin((i+Xshift)*Xalpha) \quad (1)$$

$$dy = Yamplify*\sin((j+Yshift)*Yalpha) \quad (2)$$

$$Xalpha = 2*\pi/C \quad (3)$$

$$Yalpha = 2*\pi/L \quad (4)$$

Xamplify and Yamplify in formula (1) and (2) are the amplify coefficient in direction X and Y of sine function, and also these coefficients describe a twist degree in direction X and Y. Xshift and Yshift are the shift of sine in direction X and Y. Xalpha and Yalpha are the radian of every grid in direction X and Y.

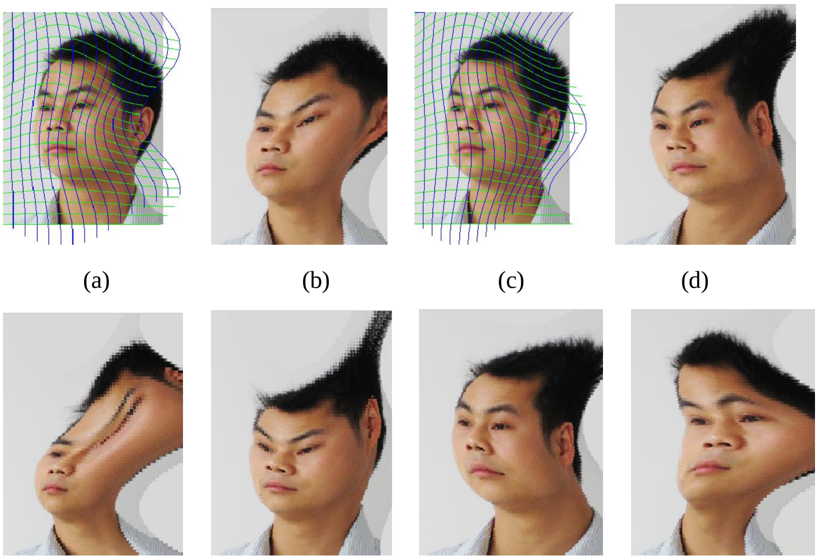
2.2 Set all parameters just on time

Analysis formula (1) and (2), you may find that if the parameters, such as Xamplify, Yamplify and dx, dy, are all set to zero, the twist template degenerates into a linear square template.

The suitable range of Xamplify and Yamplify are 0～1.5. Among them, the values below 0.5 are for the mild twist, moderate values of about 0.9, and if these value are larger than 1.3, it may introduce to a severely twist (for example, when Xamplify and Yamplify taken 1.3, a seriously twist image happened, see Figure 1(e)).

Xshift and Yshift parameters in sine function decide the twist shift in the X, Y direction, and its value may choice from 0^{o} to 90^{o} .

In our program, a user can set parameter values immediately as his needs. Not only these four set parameters can set values in continuous mode, but also the parameter values may different in direction X and Y. As a wonderful result, by select different values and make them in different combinations, an image may re-create to a lot of different funny pictures, even some of them in much unexpected effects. Fig 1 shows some different twist templates (in convex curves) and their effects.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig 1 different twist templates (in convex curves) and their effects

Due to space limitations, we can not show more images with different effects in this article. Table1 lists different parameter values in setting TwistTemplet template and its serial numbers of distorted image:

Table 1 different parameter values and corresponding different image numbers

Image number	Xamplify	Yamplify	Xshift	Yshift
1(a), 1(b)	0.5	0.5	0	0
1(c), 1(d)	0.5	0.5	90	0
1(e)	1.3	1.3	0	0
1(f)	0.3	1.3	15	75
1(g)	0.5	0.3	35	70
1(h)	1.2	0	70	0

Fig. 1(a) shows: when set Xshift to 0, corresponding nonlinear template grids in the upper ear area becomes narrower in direction X, and cause fig.1 (b) in corresponding area extended. On the other hand, fig. 1(c) taken Xshift equals 90^{o}, nonlinear template in the same grids becomes wider, so cause fig.1 (d) became narrow in the corresponding area. It is clear, that different parameters values may cause very different effects even in opposite manner.

2.3 Do twist operation on a picture to form distorted face

Section 2.1 above, we have established two templates: a SquareTemplet and a TwistTemplet. In this section, we describe a self-designed function named as bkTwist (meaning Block Twist) to finish the desired twist operation. The bkTwist function traverses on all grids in the whole template, to realize a panoramic non-linear stretch result and get some panoramic distortion effects.

To deal with this function, we first define two titles below:

l Be-rendered block: it is a target block in the target image and corresponds to a grid in the square template.

l Projection block: it is a source block in the source image and corresponds to a grid in the twist template. The local-image in this block is project (in an interpolation mode) to the area of be-rendered block.

Operation: by call "bkTwist" function implements twist operation on each be-rendered block. That is do one-by-one interpolated distorted projection operations from every projection block to every be-rendered block.

The detail of "bkTwist" function may describe as follow:

Step1. The left-top X, Y coordinates of this be-rendered block are determined under the line number i and column number j according the formula below:

$$x_{ij} = j \times scale$$

$$y_{ij} = i \times scale \tag{5}$$

Step2. The left-top and the right-bottom X, Y coordinates of this project block are determined under the line number i and column number j according the formula below (also see fig. 2):

$$Lx_{ij} = TwistTemplet[i \times C + j].x$$

$$Ly_{ij} = TwistTemplet[i \times C + j].y \qquad (6)$$

$$Rx_{ij} = TwistTemplet[(i+1) \times C + (j+1)].x$$

$$Ry_{ij} = TwistTemplet[(i+1) \times C + (j+1)].y \qquad (7)$$

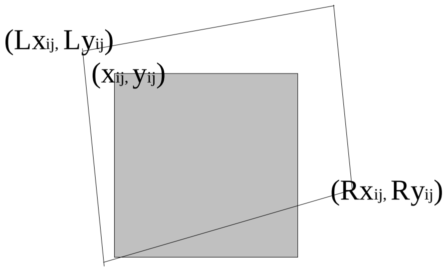


Fig 2 the relationship between a be-rendered block and a projection block

Step3. Find out the derivative of each projection block in both X, Y direction:

$$dx_{ij} = (Rx_{ij} - Lx_{ij}) \big/ scale,$$

$$dy_{ij} = (Ry_{ij} - Ly_{ij}) \big/ scale \qquad (8)$$

Step4. Find out the Y coordinate p_y on every line (b_y) in the projection block.

$$p_y = Ly_{ij} + (b_y \times dy_{ij}) \qquad (9)$$

Step5. Find out the X coordinate p_x on every column (b_x) in the projection block.

$$p_x = Lx_{ij} + (b_x \times dx_{ij}) \qquad (10)$$

Step6. Project the source RGB image pixels (located with p_y and p_x) onto target (located with b_y and b_x). In other word, take assignment operation on every pixel, one by one from source to target. The source input indicator and target output indicator are calculated as fellow:

$$\mathrm{Im}gIpt + p_y \times LineByte + p_x \times pixelByte + shift \qquad (11)$$

$$\mathrm{Im}gOpt + b_y \times lineByte + b_x \times pixelByte + shift \qquad (12)$$

Here, ImgOpt is a target image output pointer and ImgIpt is an image input pointer. lineByte is the bytes of an image line. And pixelByte is the bytes of an image row (lineByte and pixelByte are in the same values for both output and input image). And shift is the offset gradually to RGB image projection operation, respectively, 0, 1, 2.

3. Conclusion

This project is designed to re-create funny mobile cartoon characters. The core of this project is doing nonlinear stretch operation according the assigned parameter values on TwistTemplat.

Through our practice we found that the operating parameters: Xamplify, Yamplify and Xshift, Yshift have a great influence on the twist results. Some of them may get strongly unexpected effect. Readers are welcome to try and to prove this conclusion. If you have any question or need the source code you are welcome to contact us with E-mail.

The inadequacies of this paper want to be criticized and corrected.

We would like to thank the State Science and Technology Support Program and the Ministry of Culture to support the project.

References

1. PENG Ge-gang, Li Xinyu, Song Ying, Xiang Lisheng, Shen Qing, A Facial Organs Positioning Method Based on Grayness Mutation，ICWN'10.

2. Li Xinyu, Song Ying, Xiang Lisheng, Shen Qing, A Mobile Cartoon Creating

Scheme Based on the Materials Reuse, ICWN’09.

3. Peng Ge-gang, Li Xin- yu, Song Ying, Xiang Li-sheng, Shen Qing, Li Ren-fa, Research and Implementation of Mobile Phone Comic and Animation Assisted Production and Creative Platform ， ICWN’10.

Applications And Studies On Mobile Client For Mobile E-Business

Peng Ge-gang, LI Xin- yu, SONG Ying, XIANG Li- sheng, SHEN Qing, Li Ren-fa
Hunan University and Talkweb Information System CO. LTD, China，410205

Abstract - *With the rapid increase of mobile users and the further development of mobile communication, mobile e-business has become a hot trend nowadays. Among all these applications, mobile payment is the cornerstone and motive power of mobile e-business. Compared with other channels, such as website, SMS, wap etc., mobile phone client is the most important channel for mobile payment, and it has many advantages on portability, user experience and convenience indeed. This paper puts forward a way on mobile client solutions and emphatically focuses on the architecture and key problems. Actual operation situation shows that such a system has a good safety and practicality, and may become good payment tool for the mobile e-business applications.*

Keywords: Mobile e-business, Mobile phone client, Security, Terminal adapter, UI design

1 Introduction

With the progress of science and technology, especially the wide application of Internet technology, mobile e-commerce has become a new branch in e-commerce field. From the application point of view, its development is complementary to wired e-commerce and it is a new form of e-commerce development.

The mobile e-commerce refers to electronic transactions via mobile phones, personal digital assistant (PDA) and other mobile communications devices. It fully supports Internet services. Users can use the smart phone or PDA anytime, anywhere to find, select and purchase goods and services, and make use of electronic means to achieve payment. Mobile e-commerce has become a new model of world economic development and has broad development prospects [1, 2].

Mobile payment is the core of mobile e-commerce system. Mobile client has great significance for changes of payment means, expansion of payment scope, improvement of payment security and improvement of customer satisfaction[3].This paper presents a mobile client solution for mobile e-commerce and emphatically introduces the architecture of mobile client and the key issues to be resolved.

2 Architecture

As Fig1 shown, the mobile client system is divided into front-end and back-end.

- **Mobile client front-end**: The front part is mainly responsible for the user interface to show as well as the legitimacy check of a small amount (for example: the legality of the transaction amount and date). The front part is not responsible for business logic processing. In order to meet client requirements for business expansion, the entire front part is built on the basic framework for a variety of plug-ins (or applications).

  The basic framework is responsible for support functions of front-end. Corresponding plug-ins perform the implementation of a variety of applications and business. This design not only ensures that the client application can get a good scalability but also decreases the size of client install package. The basic framework consists of plug-in layer and protocol layer. Through the protocol layer, Plug-in layer provides a variety of plug-ins or applications for presentation layer. The protocol layer provides the underlying support for the basic framework.

- **Mobile client back-end:** Back-end is mainly responsible for providing strong data support for front-end. Logically, it is composed of back-end application layer, service layer and data layer.

  Back-end application layer provides a variety of background management functions, including plug-in application management, content query management, content distribution management and statistical analysis. Service layer provides a variety of services for support functions, including file services, data synchronization services, adaptation services and so on. Data layer provides data which Back office systems needs.

3 Research on Key Technologies

There are some key technology issues to be solved during the construction process of mobile electronic payment client system. We put forward our solutions combined with the characteristics of mobile payment after we fully study many traditional solutions. These research and solutions for key technology have played a crucial role in the mobile client system's building. The following discussion will focus on

† This research was supported in part by National Technical Supporting Project Foundation of China (2007BAH14B04)

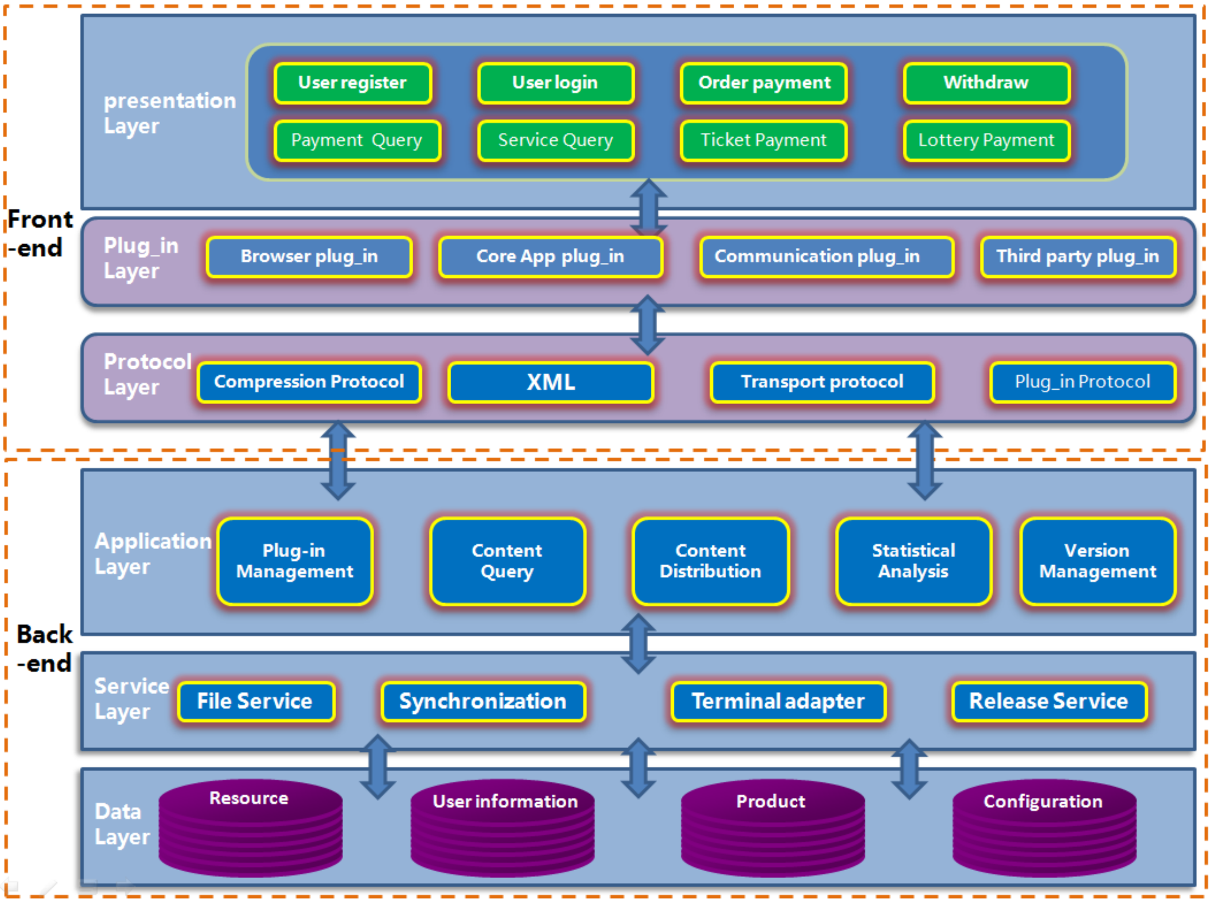


Fig1 the architecture of the mobile client system

safety issue, the terminal adapter and UI design.

3.1 Security Solution

Due to the special nature of mobile e-commerce, mobile e-payment security is particularly important. Security issue has become the most critical factors for mobile e-commerce success [4]. During the building of the system, we put forward practical solutions to solve the access control, access security, data storage security and availability.

In access control, we mainly use user name and password authentication because payment business quota has been controlled from business requirement. Client software randomly generates transfer key which are stored in memory of mobile phone. The client uses the server public key encrypted with transmission key. The back-end authentication platform converts login PIN or payment PIN into cipher text to be tested which is encrypted with PVK. Access control involves password initialization, login password's authentication and payment password's authentication. Among these processes, payment password's authentication process is the most important and it is shown as Fig2.

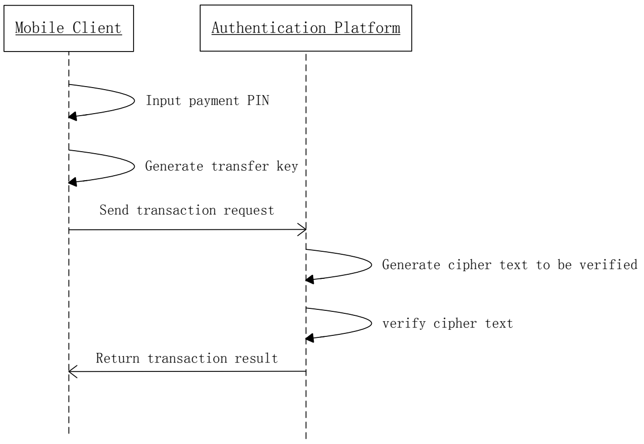


Fig2 the process of payment password's authentication

Specific steps are described below.

1) User inputs payment PIN (at this point, the user has logged in)

2) Mobile client randomly generates transfer key.

3) The client uses the server public key encrypted with transmission key (public key has been downloaded in the

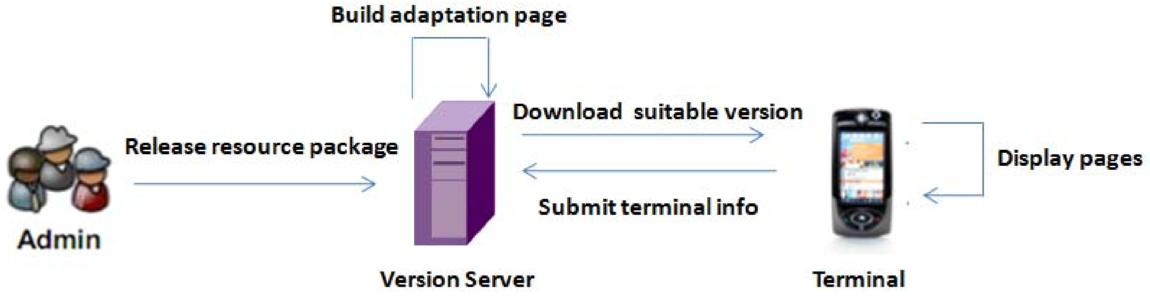


Fig3 the process of dynamic adaptation

login process and stored in memory). Payment pin is encrypted with transfer key. These two parts together constitute cipher text of payment Pin.

4) The mobile client transfers payment PIN cipher text along with transaction request to back-end authentication platform. The transaction request includes transaction type of payment PIN.

5) Back-end authentication platform invokes encryption machine interface to convert payment PIN into cipher text to be verified with PVK encryption.

6) Mobile payment platform verifies cipher text.

7) Authentication platform returns transaction result to the mobile client.

We also take other aspects in security solution into account. In communications security, we use the HTTPS security protocol to communicate with the backend server. In data storage security, we present strict requirements for sensitive data stored in the mobile terminal. Password in the client Appears only in the memory and should be deleted immediately after is verified. In the client availability, some measures also are taken to provide security protection. These measures include exception handling for call, power interruption and network interruption during transaction process.

3.2 Terminal adapter

As the core technology of mobile payment client, terminal adapter is a self-adaptation technology for mobile terminal. Its main function is identifying different types of mobile phone. According to operating system, screen size and pixels, it is responsible for converting UI design to display UI for meeting different terminal feature. In addition, the terminal adapter provides good scalability for terminal library and dynamic analysis based on basic terminal information. In the construction of the mobile payment client, the dynamic adaptation is used and the main process is shown as Fig3.

Specific steps are described below.

1) Administrator uploads resources packages of mobile client UI to terminal version server.

2) Client with terminal basic information requests version server for suitable version.

3) According to basic terminal information, version server obtains the most suitable UI resource package.

4) Version server returns the most suitable UI resource package to the mobile client.

5) The mobile client uses local UI engine to display UI.

3.3 UI Design

Visual effects for UI and usability for software operation are the key factor for product success. Excellent UI design can give the user bring a very good product experience and increase user stickiness for the product [5].

Compared to the traditional desktop and web design, the most significant constraint in product design of software product for mobile phone is the design space sharply reduced. It is the design goal of mobile client UI to better display complex functions on the small mobile phone interface. It demands UI elements need to maintain clear, reasonable, simple, and nice. So the basic layout of the design becomes particularly critical.

Button layout and label layout are the most important layout of client UI design. Button layout is suitable for application with independent functions and clear structure. The main advantage of button layout is nice and simple. Button layout also can effectively organize multiple functions of applications. The drawback is that there are more steps for switch between different functions. Switch between different functions becomes very complicated when each function interface is more complex and hierarchy is deeper. For this case, each time user has to return to main interface for accessing other functions. These inconvenient operations bring about bad user experience.

Label layout can solve the problem of function discrete in button layout. Label layout is very suitable when functions are very close each other and user need to frequently switch between the various functions. However, the current cursor may remain in the body when label page is more complex. For this case, a lot of additional key operations will bring about since user has to move the cursor to the label. These additional key operations will inevitably cause bad user experience [6].

In order to combine the advantages of button layout and labeling layout, we adopt the hybrid layout in the design of mobile payment clients. The main navigation uses label layout to switch between different functions and each label uses button layout. The hybrid layout can better organize different functional key. In this layout, user can very conveniently switch label or move between buttons. Throught the layout, all functions of mobile payment client could be clearly displayed on the phone and it is very convenient for user operation. The layout is shown as Fig4.

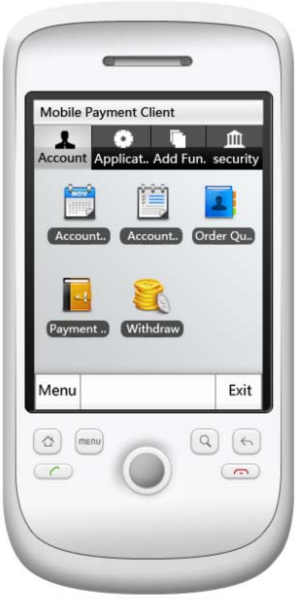


Fig4 the hybrid layout

4 Conclusion

This paper presents the solutions of mobile client based on mobile e-commerce. The key issues to be resolved in the process of building the system are discussed based on the introduction of the architecture of the system. These key issues mainly include client’s safety issue, the terminal adapter and UI design. Actual operating condition shows that the mobile e-commerce client has a good safety and practicality, and could be an excellent tool for mobile payment applications.

5 ACKNOWLEDGMENT

We would like to thank deeply to the National Technical Supporting Project Foundation of China and the Ministry of Culture of China for their support on this project.

Reference

[1] A. Tsalgatidou and J. Veijalainen. “Requirements for Mobile E-Commerce”. in EBussiness: Key Issues, Applications, Technologies, B. Stanford-Smith and P. Kidd (eds.),IOS Press, pp. 1037-1043. Proceedings of the E-business and E-work Conference, Madrid,Spain, 18-20 October 2000.

[2] Chan Yeob Yeun, and Tim Farnham. "Secure M-Commerce with WPKI". proceedings of 1st International Workshop for Asian PKI, October 2001, Korea.

[3] David McKitterick. "A Web Services Framework for Mobile Payment Services".A dissertation submitted to the University of Dublin, in partial fulfilment of the requirements for the degree of Master of Science in Computer Science, September 2003.

[4] Peter Tarasewich,Robert C. Nickerson and Merrill Warkentin. "ISSUES IN MOBILE E-COMMERCE". Communications of the Association for Information Systems. Volume 8, 2002

[5] Lee Byung-Rae． "Ticket based Authentication and PaymentProtocoI for MobiIe TeIecommunications Systems "． IEEE Comm． Magazine，2001

[6] Longman T R． Roberts T． "Creating a Successful Payment Architecture" ． IEEE 3G Mobile Communi cation． Technologies, 2002

SESSION

SENSOR NETWORKS

Chair(s)

TBA

ERoS: Role Sharing for Improved Energy Efficiency in Cluster-Based Wireless Sensor Networks

J.L. Liu[1] **and C. V. Ravishankar**[2]
[1]Department of Information Management, I-Shou University, Kaohsiung, Taiwan
[2]Department of Computer Science & Engineering, University of California, Riverside, CA, USA

Abstract - *We propose ERoS, an energy-efficient protocol that greatly improves the lifetime of cluster-based wireless sensor networks. The central idea in ERoS is to restrict the cluster head to cluster management, and offload aggregation and transmission functions from the cluster head to other appropriately chosen nodes. Cluster heads are self-selected randomly in ERoS, using a single probability parameter, so that clusters are formed autonomously and in a distributed manner. Cluster distribution is uniform in ERoS since we use a crowding distance check to ensure that clusters are sufficiently spaced apart. Energy consumption is uniform in ERoS, since it distributes energy-intensive roles across several nodes, and rotates clusters (and roles) across nodes in the network. We present analytical and simulation results comparing ERoS with other protocols, and show that it outperforms competing methods.*

Keywords: Energy-efficient Protocol, Lifetime, Cluster Head, Aggregation and Transmission Functions.

1 Introduction

Wireless sensor networks (WSNs) usually comprise a large number of tiny, inexpensive battery-powered sensor nodes. Such networks typically deployed over a wide region in order to measure physical properties, such as temperature, sound, humidity, pressure, luminosity and concentration of chemical materials, and so on, with acceptable accuracy and reliability. The nodes detect the values of the parameters of interest from the environment and transmit the collected data to a base station (BS) for further analysis and processing. Sensors have been effectively applied in combat scenarios, habitat monitoring, home security, smart hospital care, real-time communications, and so on [1-5]. WSNs comprise a large number of battery-powered devices, so energy-efficient network protocols are critical in WSN applications.

Various routing protocols have appeared in the literature for WSNs, and can be classified as direct, multi-hop relay, or as based on clustering. The direct transmission (DT) [6] and minimum transmission energy (MTE) [7-8] methods are easy to implement, but do not result in well-balanced distributions for node energy consumption. Since some sensor nodes die quickly, data for a part of the sensor field may not be detected, resulting in insufficient sensed information for analysis across the field. In the DT protocol, sensor nodes transmit their sensed data directly to BS in a single hop. This is inefficient, since single-hop transmission requires a lot of energy. Nodes located far from the BS will be particularly affected, and will dissipate their energy very quickly. In the MTE protocol, data packets are sent to the BS by way of multi-hop relay. As a result, nodes located near the BS die first, since they frequently relay data for remote nodes. Simulation results for both these protocols appear in [9]. It appears clear that a clustering communication protocol could be a better approach to extending node lifetimes [10].

The first low-energy adaptive clustering hierarchy was LEACH, proposed by Heinzelman *et al.* [9, 11]. LEACH dynamically creates clusters with cluster heads (CHs) in the set-up phase, using a threshold function T(s). Nodes sense data and transmit them to CHs using a time division multiple access (TDMA) MAC protocol. The CHs also aggregate data received from their cluster members, and then forward them to the BS. From the simulation results of Heinzelman *et al.*, each node tends to dissipate the same level of energy over time, since the CHs role is periodically rotated among all nodes. However, LEACH did not consider the states of neighbors nor energy status during CH selection. Therefore, Heinzelman *et al.* also proposed a centralized LEACH protocol, called LEACH-C, to control CH selection using a centralized clustering algorithm. Each node sends its location and energy information to the BS during the set-up phase, so the BS can determine the optimal clusters of nodes by minimizing the amount of energy required for nodes to send data to their respective CHs, using a simulated annealing (SA) method. However, LEACH-C results in limited gains in network performance [11]. Younis and Fahmy proposed a hybrid, energy-efficient, and distributed clustering approach (HEED) to improve the performance of long-lived ad-hoc sensor networks [12]. HEED periodically selects CHs based on node's residual energy and a secondary parameter. As a result, the protocol is more effective in prolonging network lifetimes, as compared to DT and LEACH. Misra *et al.* [13] proposed EEEAC, an enhanced energy efficient adaptive clustering protocol based on the residual energy of each node in the network. They pre-divide the network domain into several sub-areas, and assign a single CH in each sub-area. Their results demonstrate that EEEAC results in more balanced energy dissipation.

Lindsey and Raghavendra have proposed PEGASIS, a power-efficient gathering in sensor information system [14], which is a chain-based power efficient protocol based on LEACH. Each node receives data from its neighbors, combines this data with its own, and then transmits the aggregated data to its neighbor in the chain. Although PEGASIS is more robust and shows better performance than LEACH, the protocol collects information from the entire network, so that it incurs high maintenance cost. Also, the end nodes in the chain become network bottlenecks. A threshold sensitive energy efficient sensor network protocol (TEEN) [15] uses a hard threshold and a soft threshold to determine a node's decision of data transmission to CH. The protocol is quite good for time-critical applications.

2 Energy-efficient communication protocols

In this section, we briefly describe how CHs are selected in LEACH, and then present an improved adaptive clustering protocol with dynamic load-sharing assignment. This work assumes that the BS is static and located far from the sensors with a high enough energy. Moreover, all nodes in the network are homogeneous and distributed uniformly over the sensor field with limited energy, the links between nodes are symmetric, and the communication of all nodes is able to reach the BS.

2.1 Clustering hierarchy in LEACH

LEACH is a stochastic cluster head selection algorithm, which selects CHs dynamically and periodically according to a threshold in every round. The operation of LEACH is divided into several rounds, each round consisting of set-up and steady-state phases. To reduce data transmission costs, each node transmits its sensed data to the closest CH. The CH for each cluster receives and aggregates the data from the cluster members and then transmits the aggregated data to the BS through a single-hop relay. The clustering hierarchy in LEACH is shown in Fig. 1.

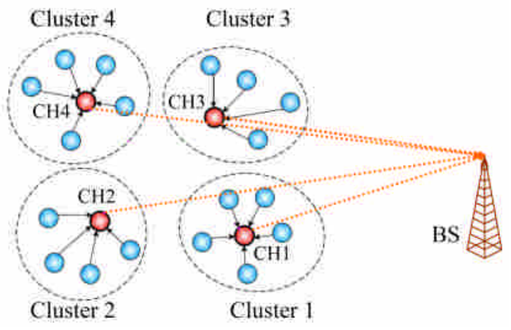


Figure 1. Clustering communication protocol of LEACH in the wireless sensor network.

As shown in Fig. 2, LEACH uses a set-up phase for CH selection, and a steady-state phase for time slot scheduling and data transmission. Each sensor *s* decides independently of other sensors whether or not to be a CH, by first choosing a random number *r* between 0 and 1 and then comparing *r* with a threshold T(s) based on a pre-specified probability *p*. The threshold is derived as follows [9]:

$$\mathrm{T(s)} = \left\{ \begin{array}{ll} \dfrac{p}{1-p\left(r \bmod (\dfrac{1}{p})\right)} & \text{if } \mathrm{s} \in G \\ 0 & \text{otherwise} \end{array} \right\} \qquad (1)$$

where *G* is the set of nodes that have not been CHs in the last *1/p* rounds. When a node decides to be a CH, it broadcasts an advertisement message to the entire sensor field, with its ID and a header, using a non-persistent carrier-sense multiple access (CSMA) MAC protocol, to ensure the elimination of collisions. This message is short, and can be broadcasted to reach all of the nodes in the network. Non-CH nodes, or member nodes, choose to join the cluster headed by the CH with the strongest received signal strength. In the next period of cluster set-up phase, the member nodes inform the closest CH that they become a member to that cluster with a join-request message containing their IDs using CSMA.

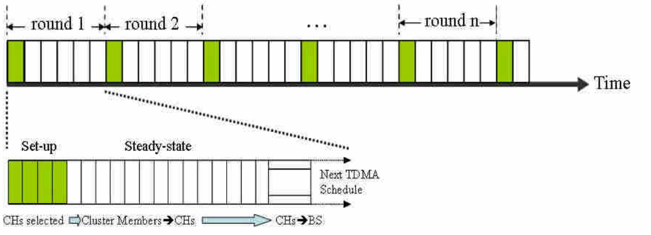


Figure 2. Time sequence for the set-up and steay-state phases.

After the cluster-setup sub-phase, the CHs recognize the number of nodes in their cluster, and their IDs. Based on all join-request messages received within the cluster, the CH creates a TDMA schedule in addition to a unique spreading code, and transmits them to cluster members at the beginning of the steady-state phase. Thereafter, each node in the cluster transmits its data packets to the CH only in its pre-specified TDMA slot to avoid collisions among transmitters using the same spreading code, and to decrease energy dissipation by allowing each node to remain in sleep mode except during its assigned time slot. When all the data packets have been received, the CHs aggregate and send them to the BS. These actions are repeated each round.

Although LEACH performs better than the direct and multi-hop protocols, it has several shortcomings. First, CHs are randomly selected using (1), without regard to node energy. Also, the CHs are not well-distributed over the sensor field; two or more CHs may crowd in a small zone. Finally, some CHs may be located far from the cluster centers resulting in more energy dissipation for member nodes in transmitting their data.

2.2 ERoS: A proposed randomized clustering protocol

In current protocols, the CH is responsible for cluster management, as well as data aggregation and transmission. This places an excessive energy burden on the cluster head. Most protocols try to distribute this energy burden across the network by rotating the CH role between nodes chosen either randomly, or according to some residual-energy metric. Our work shows, however, that the use of residual energy in CH selection still yields sub-optimal energy balance and network lifetimes.

Based on this insight, we propose a new protocol called ERoS (Energy-efficiency via Role Sharing), which chooses CHs randomly in each round, yet achieves excellent energy balance by off-loading the data aggregation and transmission functions to other nodes. The cluster head is needed in ERoS mainly for cluster formation. Data aggregation within each cluster is performed by an aggregation node AN, and data transmission by a transmission node TN.

Each round in ERoS is divided into set-up and steady-state phases, just as in LEACH. The selection CHs is driven purely by a probability parameter *p*. Each node picks a random number *r* uniformly in the interval [0, 1]. If $r < p$, the node advertises itself as a CH.

Since CHs in ERoS are randomly selected, based only on the parameter *p*, several CHs could be located near each other, causing a local imbalance in energy consumption. To distribute the CHs more uniformly, we perform a crowding distance check. If a node selects itself as a CH, but hears an announcement from another CH claimant within a distance threshold d_t (see Fig. 3), it gives up its claim to be a CH and joins a cluster as already described. The distance threshold d_t used in our work is defined as

$$d_t = \sqrt{\frac{4}{\pi}} \times \sqrt{\frac{M \times M}{p \times n}} \quad , \tag{2}$$

where *M* is the edge length of a square sensor field, and *n* is the total number of sensor nodes. As Figs. 4 (a) and (b) show, CHs tend to more uniformly distributed in the sensor field when the crowding distance check is used.

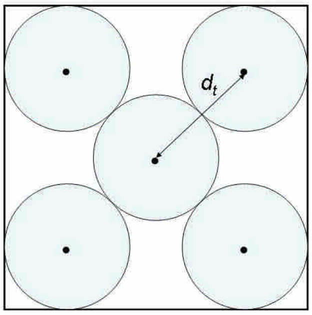


Figure 3. Specified distance threshold (d_t) between two cluster heads.

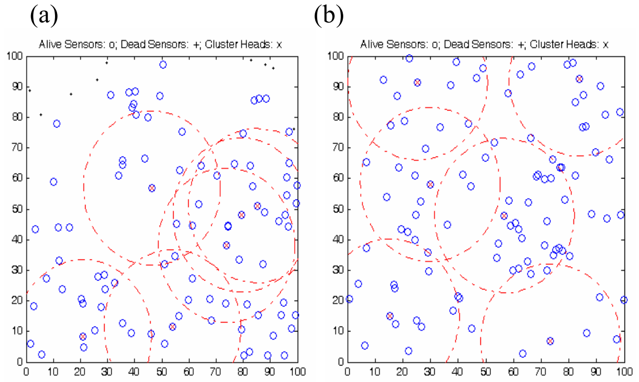


Figure 4. Distributions of cluster heads (a) without and (b) with crowding distance check.

After this crowding distance check, the surviving CHs advertise themselves to the other nodes in the network via broadcast messages. Cluster formation proceeds with each of the non-CH nodes selecting the closest CH, that is, the CH whose broadcast signal appears the strongest. Picking the closest CH minimizes the energy required for member nodes to communicate with the CH. Each node sends a join-request message to its chosen CH, with its ID, geographical position, and a header. Cluster formation is complete when all nodes in the network have joined a cluster.

2.3 Choosing the aggregation and transmission nodes

The energy required to transmit a wireless message over a distance *d* is proportional to d^{α} , where the value α depends on the channel model. The values $\alpha = 2$ and $\alpha = 4$ represent the free-space and multi-path fading models, respectively. Since data is aggregated in ERoS by the aggregation node AN, the total data transmission energy used by nodes within the cluster is minimized when the AN is at the center of the cluster. Accordingly, the AN is chosen by the CH to be the node closest to the cluster center with residual energy higher than the average value within the cluster (see Fig. 5(a)). The AN accepts data packets and aggregates them to eliminate redundancy for reducing the size of data.

Each CH also selects the node with the highest residual energy level within the cluster to be the transmission node (TN), as shown in Fig. 5(a). The TN receives aggregated packets from the AN and forwards them to the BS. Since the TN has the highest residual energy in its cluster, it is the best candidate to transmit data packets to the BS located far from the cluster. Fig. 5(b) shows how the CH, AN, and TN are distributed within the sensor field in a typical simulation. Next, the CH sends a message with the IDs of AN and TN to its cluster members via a unique sub-area code. Finally, each CH creates a TDMA schedule and a unique spreading code, and transmits them to the members of its cluster. Clearly, the CH plays the role cluster administrator in ERoS.

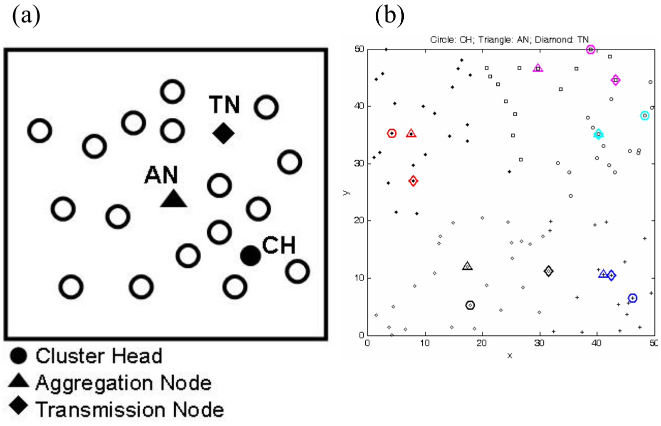


Figure 5. The distribution of CH, AN, and TN in the field.

After clusters have been formed, the steady-state phase begins, and the network starts its transmitting and receiving operations. In this phase, all cluster members transmit their sensed data to the local AN. The time assigned for each data transmission slot depends on the number of nodes in the cluster. Except when transmitting, the member nodes remain in sleep mode to save energy. As described, ANs aggregate the received data packets and send them to the TN. At the end of the schedule, the TN receives aggregated data from the cluster AN, and retransmits them to the BS.

3 Algorithm of present protocols

The pseudo-code of the proposed energy-based adaptive clustering protocol with efficient transmission routing is described as follows.

Pseudo-code of the Proposed Protocol

```
BEGIN
Specify the probability (p), number of nodes (n);
$E_{init}(s)=E_0$,  s=1,2, …, n;
do {                          //repeat for r rounds
I. SET-UP PHASE
  t←random(0,1);
  if ($E_{init}(s)>0$)  then
    if (t < p)  then
      CCH{s}=TRUE; //node s be a candidate of CH
    else
      CCH{s}=FALSE; //node s not be a candidate of CH
    end if
  end if
  if  (CCH{s}=TRUE) then
      if (distance>distance threshold $d_t$) then
          CH{s}=TRUE;      //crowding distance check
      else
          CH{s}= FALSE;   //give up to be a CH claim;
      end if
  end if
  if  (CH{s}=TRUE) then
      BC (ADV) ← broadcast an advertisement message;
      Join($ID_i$, ($x_i$,$y_i$), E(i)); //non-cluster head node i join
                                         into the closest CH
      Cluster(c);          //form a cluster c;
      GC{c}← ($x_c$, $y_c$);  //compute the geometric center
      do{
        AN{u}=TRUE; //node u be the aggregation node
           } while (E(u)> $\overline{E}$(c) & min{dist(u,GC(c))})
      do{
        TN{v}=TRUE; //node v be the transmission node
           } while (E(v)=max{E(c)})
  end if
II. STEADY-STATE PHASE
  If (AN(s)=TRUE) then
     Receive($ID_i$, DataPCK) //receive data from members;
     Aggregate($ID_i$, DataPCK) //aggregate received data;
     TansToTN($ID_{AN,}$ DataPCK); //transmit received data;
  else
     If (MyTimeSlot=TRUE) then
        TansToAN($ID_i$, DataPCK); //transmit sensed data;
     else
        SleepMode(i)=TRUE;      //node i at a sleep state
     end if
  end if
  If (TN(s)=TRUE) then
        Receive($ID_{AN}$, DataPCK); //receive data from AN
     If (MyTimeSlot=TRUE) then
        TansToBS($ID_{TN}$, DataPCK); //transmit data to BS;
     end if
  else
        SleepMode(s)=TRUE;      //node s at a sleep state
  end if
}             // one round is completed
END
```

4 Analysis of energy dissipation in ERoS

We use a first-order radio model [9] in evaluating ERoS. The parameter values used in our simulation model are listed in TABLE I. According to the radio energy dissipation model illustrated in Fig. 6, the energy required by the transmit amplifier $E_{Tx}(l,d)$ for transmitting a l-bit message over a distance d between a transmitter and a receiver is given by

$$E_{Tx}(l,d)=\begin{Bmatrix} l\times E_{elec}+l\times\varepsilon_{fs}\times d^2 & if\ d\le d_0 \\ l\times E_{elec}+l\times\varepsilon_{mp}\times d^4 & if\ d\ge d_0 \end{Bmatrix} \quad (3)$$

where $d_0=\sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$ expresses the threshold distance, E_{elec} represents the energy consumption in the electronics circuit to transmit or receive the signals, and the terms of $\varepsilon_{fs}d^2$ and $\varepsilon_{mp}d^4$ represent amplifier energy consumption for a shorter and longer distance transmissions, respectively. To receive the l-bit message, the energy $E_{Rx}(l,d)$ required by the receiver is given by

$$E_{Rx}(l,d)=l\times E_{elec} \quad (4)$$

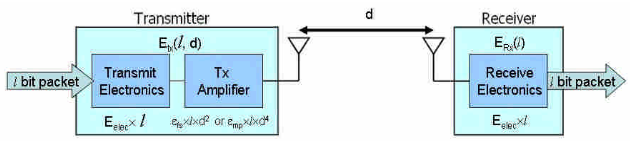


Figure 6. First-order radio model.

We analyze the energy consumption under the first-order radio model for ERoS as follows. Let a total of n sensor nodes be distributed uniformly in the sensor field of size $M \times M$ (m^2), and be grouped into k clusters. The energy costs required to transmit/receive control messages are neglected in the following analyses of energy consumption, since data packets (l) are far larger than control messages (l_{ctrl}). The energy required per round for an AN to receive data packets from member nodes, and aggregate the received data and forward them over a distance d_{toTN} to the TN is

$$E_{AN}(l,d) = l \times \left[E_{elec}\left(\frac{n}{k} - 1\right) + E_{DA}\,\frac{n}{k} + E_{elec} + \varepsilon_{fs} \times d^2_{ANtoTN} \right] \quad (5)$$

where E_{DA} represents the energy dissipation for aggregating data. In addition, the energy required per round for a TN to receive aggregated data packets from AN and forward them over a distance d_{toBS} to the BS is

$$E_{TN}(l,d) = \begin{cases} 2l \times E_{elec} + l \times \varepsilon_{fs} \times d^2_{toBS} & if\ d_{toBS} < d_0 \\ 2l \times E_{elec} + l \times \varepsilon_{mp} \times d^4_{toBS} & if\ d_{toBS} \geq d_0 \end{cases} \quad (6)$$

The energy dissipation for a member node, or a non-aggregation node, is

$$E_{non\text{-}AN}(l,d) = l \times E_{elec} + l \times \varepsilon_{fs} \times d^2_{toAN} \quad (7)$$

where d_{toAN} represents the distance between a cluster member and its AN. Since the nodes are assumed to be uniformly distributed in the sensor field, the expected value of squared distance from a member node with coordinate at (x, y), to its AN, which located approximately at the center of a cluster in our protocol, is given by

$$E\left[d^2_{toAN}\right] = \frac{1}{A}\iint (x^2 + y^2)\,dx\,dy \quad (8)$$

Assuming the shape of clusters is a circle, (6) becomes

$$E\left[d^2_{toAN}\right] = \frac{1}{2\pi}\frac{M^2}{k} \quad (9)$$

The expectation of d^2_{toAN} in (9) matches that of d^2_{toCH} in the work of Heinzelman *et al.* [11], since they assumed the CH to be at the center of cluster. This assumption needs to be corrected, since the CHs are not located at the center of clusters in most cases. In the general case, the value of d^2_{toCH} should be the twice that of d^2_{toAN} [16]. Similarly, the expected value of the squared distance from the AN to TN, assuming the TN at (x', y'), also can be approximated as

$$\begin{aligned} E\left[d^2_{ANtoTN}\right] &= \frac{1}{A}\iint (x'^2 + y'^2)\,dx'dy' \\ &= \frac{1}{2\pi}\frac{M^2}{k} \end{aligned} \quad (10)$$

Moreover, the energy dissipated in a cluster is obtained as

$$\begin{aligned} E_{total} &= k \times \left(E_{AN} + E_{TN} + (\frac{n}{k} - 1)E_{non-AN} \right) \\ &\approx k \times \left(E_{AN} + E_{TN} + \frac{n}{k}E_{non-AN} \right) \end{aligned} \quad (11)$$

Thus, the total energy dissipation for a round is given by

$$E_{Total} = \begin{cases} l \times \left[2(n+k)E_{elec} + nE_{DA} + k\varepsilon_{fs}E[d^2_{toBS}] + \varepsilon_{fs}\dfrac{(n+k)M^2}{2\pi k} \right] & if\ d_{\mathrm{toBS}} < d_0 \\ l \times \left[2(n+k)E_{elec} + nE_{DA} + k\varepsilon_{mp}E[d^4_{toBS}] + \varepsilon_{fs}\dfrac{(n+k)M^2}{2\pi k} \right] & if\ d_{toBS} \geq d_0 \end{cases} \quad (12)$$

where $E\left[d_{toBS}\right]$ is the expectation of d_{toBS}. Equation (12) shows that the total energy dissipation is most significantly affected by the distance between TN and BS, and the size of the sensor field. The corrected equations for total dissipation in LEACH are presented in [16] as follows.

$$E_{Total} = \begin{cases} l \times \left[2nE_{elec} + nE_{DA} + k\varepsilon_{fs}E[d^2_{toBS}] + \varepsilon_{fs}\dfrac{nM^2}{\pi k} \right] & if\ d_{\mathrm{toBS}} < d_0 \\ l \times \left[2nE_{elec} + nE_{DA} + k\varepsilon_{mp}E[d^4_{toBS}] + \varepsilon_{fs}\dfrac{nM^2}{\pi k} \right] & if\ d_{toBS} \geq d_0 \end{cases} \quad (13)$$

From Eqs (12) and (13), we can see that although ERoS increases required power by $2lkE_{elec}$, this increase is small, since number of cluster k is small. However, it reduces the energy consumption roughly by $\dfrac{l\varepsilon_{fs}nM^2}{2\pi k}$ when $(n-k) \approx n$. Therefore, it is beneficial to assign the AN function to a node other than the CH.

Table I. Parameters of the first-order radio model

Parameters	Values
Initial energy (E_0)	0.25 J, 0.5 J, 1 J
Transmitter Electronics (E_{elec})	50 nJ/bit
Receiver Electronics (E_{elec})	50 nJ/bit
Data Packet Size (k)	2000 bits
Transmitter Amplifier (ε_{fs}) if $d \leq d_0$	100 pJ/bit/m^2
Transmitter Amplifier (ε_{mp}) if $d \geq d_0$	0.0013 pJ/bit/m^4

5 Simulation results

We distributed the nodes randomly in a $M \times M$ sensor field with M = 50 m. The BS was located at (25, 150), (25, 250), and (25, 350) for investigating the effect of BS distance on our protocol. The probability p was set at 5%, identical to the settings in [9, 11]. All our simulations were performed using MATLAB 6.5 software on the Windows XP platform. Each simulation was repeated for 30 independent runs, and the results of the runs averaged. In addition, control packet sizes for broadcasting packet and packet header were 50 bits long in the present simulations.

5.1 ERoS simulation and evaluation

TABLE II lists the simulated results obtained using MTE, DT, LEACH, and present protocol for BS located at (25, 150). Three initial conditions of energy for all nodes are prescribed, and the ratios of dead nodes to all nodes at 1%, 20%, 50%, and 100% were recorded during simulations. For the case of E_0=0.25(J), the sensor nodes all died after 284, 119, and 658 rounds in MTE, DT, and LEACH, respectively, whereas all nodes were still alive till 645 rounds in ERoS. Moreover, the nodes died evenly between rounds 645 and 686, with Δ=41 only. In contrast, the Δ values were 282, 65, and 191 for MTE, DT, and LEACH, respectively. It is clear that energy dissipation in ERoS is far more uniform than in the other three protocols. Similarly, for the case of E_0=0.5(J), ERoS outperforms MTE, DT, and LEACH for the lifetime of network. All nodes survived up to 1313 rounds in ERoS, but no nodes were still alive after 1298 rounds for LEACH.

Table II. Data of lifetimes using different amount of initial energy and protocols

Energy (J/node)	Protocol	Node Death				Duration of Node Death
		1%	20%	50%	100%	Δ(1%→100%)
0.25	MTE	2	30	78	284	282
	DT	54	62	76	119	65
	LEACH	467	513	549	658	191
	Present	645	668	678	686	41
0.5	MTE	4	56	142	536	532
	DT	108	123	153	236	128
	LEACH	951	1027	1098	1298	347
	Present	1313	1336	1346	1355	42
1	MTE	8	106	272	1054	1046
	DT	216	246	304	469	253
	LEACH	1939	2055	2180	2588	649
	Present	2677	2699	2710	2719	42

Also in LEACH, the number of surviving nodes stays at the initial value of 100 before round 951 and then reduces gradually, with all nodes being dead by round 1298. In this situation, it is very likely that some regions of the sensor field were populated by dead nodes only, so no sensed information was available for such regions. As a result, the data transmitted to BS could be insufficient for data analyses. For the case of E_0=1.0(J), ERoS showed the best results in terms of lifetime of network as compared to MTE, DT, and LEACH. The values of Δ obtained using ERoS are small and within 41—42 rounds. Network lifetimes for the cases of E_0=0.5(J) and E_0=1.0(J) are shown in Fig. 7. ERoS clearly outperforms MTE, DT, and LEACH. Figures 8(a) and 8(b) show that the distribution of cluster numbers under ERoS is more consistent than for LEACH. The number of clusters is also smaller than for LEACH because ERoS uses a crowding distance check. As a result, the required energy cost for transmitting data packets to far BS is reduced.

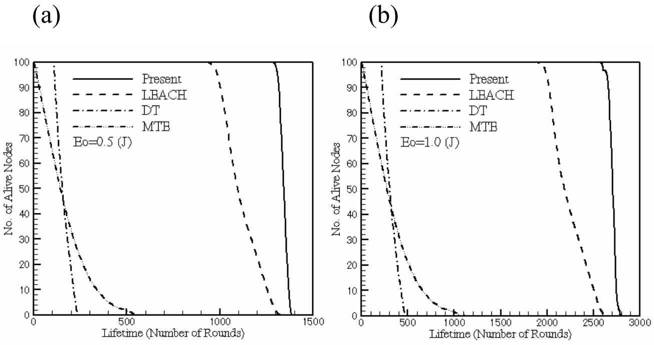


Figure 7. Comparison of lifetime for (a) E_0=0.5(J) and (b) E_0=1.0(J) cases.

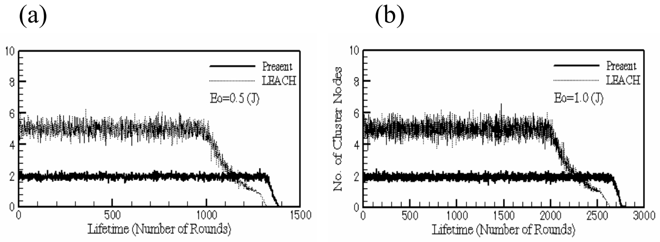


Figure 8. Distributions of cluster number for (a) E_0=0.5(J) and (b) E_0=1.0(J) cases.

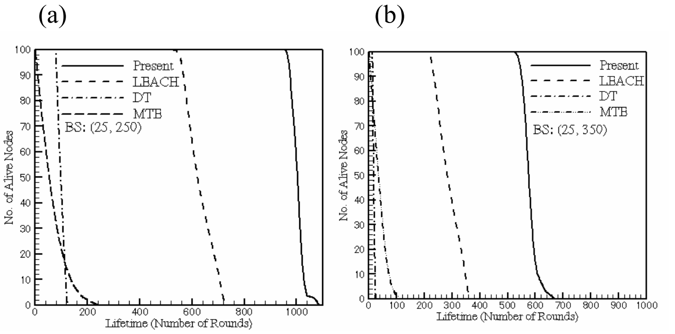


Figure 9. Comparison of lifetime for the BS located at (a) (25, 250) and (b) (25, 350).

5.2 Simulation results of BS located at far positions

We also studied the performance of ERoS when the BS was located far from the field. We considered two cases, with the BS located at (25, 250) and (25, 350). The initial energy for each node was set as 0.5(J). Figures 9(a) and (b) compare the protocols with the BS located at (25, 250) and (25, 350), respectively. ERoS showed excellent performance in terms of lifetime of the network. As the BS moves farther from the sensor network, ERoS showed a higher ability to prolong the network lifetime. The improvements with ERoS are 40.1% and 66.0% as compared to LEACH, with the BS located at (25, 250) and (25, 350), respectively. Both DT and MTE showed a poor performance in the lifetime of network for these cases.

6 Conclusions

We have proposed ERoS, a protocol for improving the energy efficiency in WSNs using dynamic role-sharing. The dynamic role-sharing is realized by reassigning some of the roles played by the cluster head in traditional protocols to nodes other than the cluster head. ERoS results in far more uniformity in energy dissipation than competing protocols. Nodes self-select themselves to be cluster heads using just a single probability parameter, so cluster-head selection is fully autonomous and distributed. In each cluster, the data packets are received from cluster members by an aggregation node (AN) located near the cluster center, reducing energy usage within a cluster. The AN aggregates them and send the aggregated data to a transmission node (TN), which has the highest residual energy in a cluster, to ensure that it has enough power to transmit data packets to far BS.

We also analyzed the total energy dissipation required in each round in ERoS. ERoS increases the energy expenditure for electronics slightly, but reduces the transmission energy between member nodes and their aggregation nodes. Our experimental results show that ERoS outperforms MTE, DT, and LEACH in terms of network lifetime and node survival rates. The nodes died evenly in ERoS. Our experiments show that ERoS is energy-efficient, and suitable for WSN applications.

7 Acknowledgment

This work was supported in part by grant N00014–07–C–0311from the Office of Naval Research.

8 References

[1] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci, “A Survey on Sensor Networks,” IEEE Communications Magazine, Vol. 40, No. 8, pp. 102–114, Aug. 2002.

[2] Kemel Akkaya and Mohamed Younis, “A Survey on Routing Protocols for Wireless Sensor Networks,” Ad Hoc Networks, Vol. 3, pp. 325–349, May 2005.

[3] Ameer Ahmed Abbasi and Mohamed Younis, “A Survey on Clustering Algorithms for Wireless Sensor Networks,” Computer Communications, Vol. 30, pp. 2826–2841, Oct. 2007.

[4] Alan Mainwaring, Joseph Polastre, Robert Szewczyk, David Culler, and John Anderson, “Wireless Sensor Network for Habitat Monitoring,” ACM WSNA’02, Atlanta, Georgia, pp. 88–97, Sept. 2002.

[5] Shanzhong Zhu, Wei Wang, and Chinya V. Ravishankar, **“**PERT: A New Power-efficient Real-time Packet Delivery Scheme for Sensor Networks,**”** International Journal of Sensor Networks**,** Vol. 3, Issue 4, pp. 237–251, Jun. 2008.

[6] Chalermek Intanagonwiwat, Ramesh Govindan, Deborah Estrin, John Heidemann, and Fabio Silva, “Directed diffusion for Wireless Sensor Networking,” IEEE/ACM Transactions on Networking, Vol. 11, Issue 1, pp. 2–16, Feb. 2003.

[7] Timothy J. Shepard, “A Channel Access Scheme for Large Dense Packet Radio Networks,” ACM SIGCOMM Computer Communication Review**,** Vol. 26, Issue 4, pp. 219–230, Oct. 1996.

[8] Bhaskar Krishnamachari, Deborah Estrin, and Stephen Wicker, “Modeling Data-Centric Routing in Wireless Sensor Networks,” Wireless Communications, Vol. 1, Issue 4, pp. 660–670, Oct. 2002.

[9] Wendi R. Heinzelman, Anantha P. Chandrakasan, and Hari Balakrishnan, “Energy-efficient Communication Protocol for Wireless Microsensor Networks,” IEEE Proceedings of the 33rd Annual Hawaii International Conference on System Science, pp. 1–10, Jan. 2000.

[10] Dharma. P. Agrawal and Qing-An Zeng, “Introduction to Wireless and Mobile Systems”, Pacific Grove, Thomson Brooks/Cole, 2003.

[11] Wendi B. Heinzelman, Anantha P. Chandrakasan, and Hari Balakrishnan, “An Application-specific Protocol Architecture for Wireless Microsensor Networks,” IEEE Transactions on Wireless Communications, Vol. 1, Issue 4, pp. 660–670, Oct. 2002.

[12] Ossama Younis and Sonia Fahmy, “HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks,” IEEE Trans. Mobile Computing, Vol. 3, No. 4, pp. 366–379, Oct. –Dec. 2004.

[13] Iti Saha Misra, Sudipto Dolui, and Ambarish Das, “Enhanced Energy-Efficient Adaptive Clustering Protocol for Distributed Sensor Networks,” IEEE International Conference on Networks, Vol. 1, pp. 16–18, Nov. 2005.

[14] Stephanie Lindsey and Cauligi S. Raghavendra, “PEGASIS: Power-Efficient Gathering in Sensor Information System,” Proceedings of 2002 IEEE Aerospace Conference, pp. 1–6, Mar. 2002.

[15] Arati Manjeshwar and Dharma P. Agrawal, “TEEN: A Routing Protocol for Enhanced Efficient in Wireless Sensor Networks,” 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, Apr. 2001.

[16] Jenn-Long Liu and Chinya V. Ravishankar, “LEACH-GA: Genetic Algorithm-Based Energy-Efficient Adaptive Clustering Protocol for Wireless Sensor Networks,” International Journal of Machine Learning and Computing, Vol. 1, No. 1, pp. 79–85, Apr. 2011.

A Partial Sorting Algorithm in Multi-Hop Wireless Sensor Networks

Aboubecrine Ould Cheikhna
Department of Computer Science
University of Picardie Jules Verne
80039 Amiens, France
Ould.cheikhna.aboubecrine @u-picardie.fr

Jean Frédéric Myoupo
Department of Computer Science
University of Picardie Jules Verne
80039 Amiens, France
Jean-frederic.myoupo@u-picardie.fr

Abstract— ***The partial sorting problem is to sort the k smallest elements of a given set of n integers such that 1 <k <n in descending order. The algorithms that exist in the literature for the partial sorting in wireless networking solutions are based on single-hop model. In this paper we consider a multi_hop sensor network. Initially, we partition the network into several levels by using the method of Gerla and Tsai. We consider that each node has a data item and two linked lists. Once the network is partitioned with multiple levels, each cluster runs the algorithm MaxCluster seeking the maximum element in that cluster. Then the algorithm Partial_Sorting finds separators (separators are identified by the algorithm MaxCluster) and helps to identify the remaining elements of the list to be sorted. We give the upper bound of our algorithm in terms of broadcast rounds. We also extend our approach to the case where each node has multiple data items instead of a single datum. Finally Experimental results highlight our approach***

Keywords- Sensor Networks; Cluster head; Clustering; Partial Sort; Wireless communication.

I. INTRODUCTION

The sorting problem has been extensively studied for decades. They proposed different techniques to solve this problem. Most of these techniques focus on how to sort a set of n elements as quickly as possible. Hoare suggested in the early 60's quicksort. Until recently, Martinz, Shiau and Yang have studied how to sort the k smallest elements of a given set of n elements as 1 <k <n. MARTINZ showed that this problem is known as a partial ordering. Today, due to development of personal networks, the wireless communication (WN) has become more interesting and attractive. Single-hop model is a complete graph and consists of n stations where each station listens to all the others. Each station in this model can communicate with each other via a common channel only. If two or more stations want to send messages simultaneously, sending a conflict occurs. In this model we assume that each station has the ability to detect the conflict. When a conflict is detected, the pattern of resolution being put in place.

A. Related work

Quicksort is one of the most influencing algorithms in sorting. It has been studied and chosen to appear among the best ten algorithms in the 20th century [7]. The average number of comparisons for the quick sorting is *O(nlogn+ O(n))*[7]. One of the variants of quick sorting called median-of-three, runs in *(12/7nlogn+ O(n))* [7, 12, 13]. The amelioration of *2n* to *12/7n* is proved by Chern and Hwang [3] who discussed the generalized quicksort. Martinez [9] has recently introduced partial sorting. The problem is not the sorting of *n* elements, but rather selecting the smallest *k* elements among *n* elements, *1 < k < n*. Martinez [9].

Now, the question is how to resolve the partial sorting problem when *1 < k <n*. Martinez [9] has proposed an algorithm called partial quicksort to solve the partial sorting. Shiau and Yang [11], also proposed another algorithm " generalization of partial sorting" for resolving the same problem. Both two algorithms were based on Quicksort and some of their results are almost the same. The main difference between the two papers is the original function and the methods of analysis they use. The method in [7] consists of selecting the j*-th* element of a file containing *n* elements. Prondinger [11] derives an algorithm to select the *k* elements j_1-th, j_2-th,..., j_k -th. It may also resolve the partial sorting problem if we let *k* selected elements be the k smaller elements.

Martinez [9] mentioned that his algorithm can resolve the partial sorting problem but with two constraints. The first constraint is that *k* must be large enough. The author does not mention what value of *k* is large enough. The second constraint is that the *n* elements must be treated offline. Furthermore, Sedgewick and Flajolet [13] show that the standard deviation the number of comparisons used by Quicksort is 0.6n. This means that the accuracy of performance prediction is not very good and need to be further discussed. The first constraint means that some algorithm better results than the algorithm of Martinez [9] when k is not big enough. The extreme case is when k = 1. Of Clearly, resolution of this case, the smallest element selected by the algorithm is the best and it will take (n-1) comparisons. The algorithm of Martinez [9] needs $2H_n$ 2n comparisons, where H_n is harmonic number. The two algorithms have less selection good results of the algorithm when k = 1.

How quickly can we sort the two smaller elements of set *n* elements? A simple method is by executing the algorithm that allows the smallest elements fall twice. The second iteration will choose the second smaller element starting from *(n-1)* of remaining elements. The time of execution of the method of selection is *n* to select the smallest *n* elements in a set of *n* elements. Generally, this method for researching the *k* smaller elements among *n* elements needs

$$\frac{1}{2}[(n-1)+(n-k)]k$$

comparisons. It proves that partial sorting loses its efficiency with respect to Partial Quicksort when *k>=3*. One can think that the algorithm of Martinez [7] and the algorithm of Shiau and Yang [11] may take leadership position for *k>=3* on the traditional model. It is not the case: Shyau-Horng Shiau has proposed an algorithm for resolving the partial sorting problem online. The algorithm outperforms both algorithms and assumes the position of leader *k =1* with $k = \frac{3}{5}\sqrt{n}$ comparisons. The algorithm is an adaptation of sorting insertion and can be treated online.

Some sorting algorithms for single hop wireless networks appeared in the literature [1,4, 8, 10, 14, 15]. In [10] a wireless network on p station is considered to sort n data items using k channels, k<p. Their algorithm runs in $4\frac{n}{k}+\frac{p}{k}+O(\sqrt{n})$ broadcast rounds provided that $k \le \sqrt{\frac{p}{2}}$.

Shiau et al [14] proposed an algorithm for partial sorting in single up network with the number of slot time bounded by $\frac{23}{6}k + 5\ln(n-(k-1)) - \frac{7}{6}$ where k is the number of channels.

B. Main contribution

This paper proposes an algorithm for solving the partial sorting problem (sorting *k*, the smallest elements in a set of *n* elements), in wireless multi-hop sensor networks. Initially, we consider that each node in the network has one and only one data. Once we succeed in solving the problem, we proposed a generalization of our approach to solve the partial sorting problem in the case where a node can have one or more data items. Our approach does not need to have complex nodes: each node is capable of turning a sorting algorithm in order to locally select the largest value in the list it holds. More precisely we show the number of broadcast rounds for a partial sorting problem cannot exceed

$$O(n)+|S|*P_r*\left\lceil\frac{n}{|c_{\min}|}\right\rceil |c_{\max}|*|c_{\max}|\log(|c_{\max}|)$$

where
n is the number of sensors in the network
S is the st of separators
Pr is the number of layers obtained after the clustering process
C_{max} is the cluster of maximum size among all clusters of the different layers.
C_{min} the cluster of minimum size among all clusters of the different layers

The rest of paper is organized as follows: the following Section presents our approach in the case where each node has a single data. Section 3 is a detailed description for the generalization for our approach for the case when one node has multiple data's. Section 4 presents a simulation for our approaches and finally a conclusion to our paper.

II. PARTIAL SORITNG ALGORITHM ON THE MODEL MULTI-HOP

After the network clustering, the method consists for each station to send the value it contains to the clusterhead, which is its leader. Once the clusterhead receives all messages, it seeks the maximum among all the received elements.

We will use the notation $< x_0, x_1, \ldots, xt >$ to denote a linked list, where x_t is the head of the list. Let us suppose $L_1 = < x_0, x_1, \ldots, xt >$ and $L_2 = < y_0, y_1, \ldots, y_t >$, the notation $< L_1, L_2 >$ represents concatenation of L_1 and L_2, which is $< x_0, \ldots, x_t, y_0, \ldots, y_t >$. In the description of our algorithm, we also use the term " data elements " that represents the data given by the station.

A. Clustering algorithm

We use the algorithm *Gerla &Tsai* [5] for the hierarchical clustering of our network. This algorithm consists in finding a set of interconnected clusters. More precisely, the topology of the system is separated into scattered partitions. Once the network is partitioned, we repeat the process until we find a single cluster that is named the Super Clusterhead. The decision is based on data held by the nodes: we consider that the node holds the smallest value and prioritize the most adapted for this task to be clusterhead.

An interesting point in this algorithm is that the cluster head nodes and the ordinary nodes all do the same task when working in the construction of cluster. Thus, they spend as much energy as each other. The complexity of this algorithm is *O(|V|)*, which *V*, all nodes of the graph represent the network.

B. Assumptions

Let M be a set of data items associated with a network which consist of a group of sensors that can communicate with each other via a communication channel. *M* is also associated with a graph where vertices are the sensors. In this directed graph, there is an arc from *u* to *v* if *u* can send a message toward *v*. We assume also that the *n* nodes are deployed; each nod has a data and two linked lists. We consider the following notations:

- *M*: set of data associated with a sensor network;
- $\cup C_{ij}$: The tree partitioning of the network *M* using the algorithm of *Gerla and Tsai*;
- P_r : The depth of tree clustering;
- Node v: Each node in progressive L'_g and L_g, with L'_g contains the values hold by their neighbors in their cluster, L_g list that contains the most great value in the cluster
- C_{ij} : The cluster j in layer i;
- $|C_{ij}|$: The size of the cluster j in layer i (number of nodes);
- x_{ij}^{k} : Evaluate the hold by the node k in cluster j layer i;
- x_{ij}^{c} : Evaluate the hold by the clusterhead of the cluster j in layer i;
- c, k : Represents respectively clusterheads and ordinary nodes;
- $L_{g_{ij}}{}^{c}$: The chained list that detains the largest values of the clusterhead of node cluster j in layer i;
- $L_g = k$: Chained list named list of separators, which contains the largest values of each, cluster that is part of the tree portioning.
- L_m: Chained list named list of separators, which contains the largest values of each, cluster that is part of the tree partitioning.

III. DESCRIPTION OF OUR ALGORITHM

A. Algorithm MaxCluster

i.- Step 1: Multi-level clustering

We use Gerla & Tsai clustering algorithm to yield a multi-layer cluster organization: $\cup C_{ij}$ with C_{ij} in the cluster *j* in layer *i*. As mentioned above, it is obvious that each cluster has a leader named clusterhead.

ii.- Step 2: Finding the greatest data item of each cluster.

The algorithm MaxCluster is a method of finding the maximum: it consists for an ordinary node in a cluster to send its value to its cluster head. Once it receives all data items, finds the greatest data item among all those it received. The cluster heads work in parallel in this step.

Step 3: Concatenation method of the data in the chained list of cluster head of the cluster using the algorithm MaxCluster that is presented in algorithm 1 below.

Algorithm 1: MaxCluster

```
Input: $\cup C_{ij}$ where $C_{ij}$ is cluster j in layer i.
Output: Chained list $L'_g$, the list of separators
$L^{k}_{ij}$ chained list of node k of cluster j in layer i, $x_{max}$=NULL, $L'_g$ = $L'_p$=NULL, $P_r$, the depth of the tree
For i=1 Until $P_r$-1
{
    For j=1 Until $q_i$ where $q_i$ is the number of cluster in layer i.
    {
        For l=1 Until $|C_{ij}|$
        {
            each sensor in $C_{ij}$ sends its value to its cluster head
            $L_{g_{ij}}{}^{c} =< L_{g_{ij}}{}^{c}, x^{l}{}_{ij} >$
            If $x^{l}{}_{ij} > x_{\max}$
                $x_{\max} := x^{l}{}_{ij}$
        }
        $L_{g_{ij}}{}^{c} := Quicksort(L_{g_{ij}}{}^{c})$
        $L'_g :=< (L'_g, x_{\max} >$
    }
}
$L'_g := Quicksort(L_g)$
Return $L'_g$
```

B. *Partial soritng algorithm*

For distinct *n* elements, the partial sort problem is to find the first *k, k ≥1*, the most k greatest elements in the non-increasing order. Our partial sorting algorithm may be represented as a function of recursive research: (L_g, c)= *Sort-partial (M, k)*, where L_g, is the chained list that stores the sorting sequence. *c* is the number of elements in L_g and *M* represents a set of data items (sensors network). In our algorithm each sensor has exactly one data item and maintains two chained lists as mention above. Finally, we will get a chained sorting list L_g containing the first greater *k* elements. For any set of data items, there exists at least one sensor network that is its representation

For our partial sorting algorithm, we use the algorithm *MaxCluster* that is a method of researching the maximum. A set of elements are considered as separators to bring the data items remaining in the layers properly. Then the lists of separators will be decomposed in the form of intervals. The recursive function is based on these intervals in order to identify the remaining elements, which will be in the sorted list. The algorithm can be represented as a recursive function of research framed through interval. Our algorithm of partial sort is depicted in the algorithm 2

Algorithm 2: Partial Sort *(L_g, c)=Partial-Sort(M, K)*

```
Input: M the set of items to sort (sensor network)
Output: chained list $L_g$ that contains the k greatest Values
Step 1: Hierarchical clustering of the network
Step 2: Algorithm MaxCluster
Step 5: $L_m$=MaxCluster, $L_m=<x_1, x_2, \dots, x_{t-1}, x_t>$ with $x_{i-1} \le x_i$, $1 \le i \le t$.
Set $L_g$=NULL
For $i=0$ Until $|L^c g_{1(p_r-1)}|$
{
    $L_g =< x_t, L_g >$
    $|L_g| \ge c$
    Return
    Else
    {
        If $(x_{t-1} < L^c g_{1(p_r-1)}[i] < x_t)$
            $L_m =< L^c g_{1(p_r-1)}[i], L_m >$
    }
}
Return $(L_g, c)$
```

Theorem: *The number of broadcast rounds required by our partial sorting algorithm in a multi hop Wireless sensors networks cannot exceed*

$$O(n) + |S| * P_r * \left\lceil \frac{n}{|c_{\min}|} \right\rceil |c_{\max}| * |c_{\max}| \log(|c_{\max}|)$$

Where n is the number of sensors in the network
S is the set of separators
Pr is the number of layers obtained after the clustering process.
C_{max} is the cluster of maximum size among all clusters of the different layers.
C_{min} is the cluster of minimum size among all clusters of the different layers.

Proof:
It is well known that the complexity Gerla and Tsai [5] clustering algorithm in terms of broadcast rounds is

$$O(V)$$

V is the set of sensors of the original network. Since we set *n* as the number of sensors in the network, we have clearly

$$O(V) = n \quad (1)$$

In our partial sorting algorithm, we have the following symbols:
Pr :The number of layer in the clustering process ;
$|c_{ij}|$: The size of cluster *i* in layer *j* ;

Also note that our partial sorting algorithm uses the quick sort technique in the algorithm MAxCluster to identify the greatest element in each cluster. The clusters in all layers run MaxCluster in parallel. Thus the complexity of this algorithm is a function of cluster size and is given by the following formula:

$$|c_{ij}|_{\max} \log(|c_{ij}|_{\max})$$

for the clusters in layer j. Note C_{max} the cluster of maximum size among all clusters of the different layers. Thus the complexity of MaxCluster is dominated by

$$|c_{\max}| \log(|c_{\max}|$$

We use the following strategy to derive the number of broadcast rounds of our partial sorting algorithm :
$|c_{ij}|_{\max}$: The size of the cluster that contains the largest number of elements from all the clusters in layer *j* ;
$|c_{ij}|_{\min}$: The size of the cluster that contains the smallest number of elements from all the clusters in layer *j*;
$\left\lceil \frac{n}{|c_{ij}|_{\min}} \right\rceil$: The largest number of clusters in layer *j* ;

One can easily deduce that the largest number of rounds in the layer *j* is:

$$\left\lceil \frac{n}{|c_{ij}|_{\min}} \right\rceil |c_{ij}|_{\max} * |c_{ij}|_{\max} \log(|c_{ij}|_{\max})$$

Note $C_{\min}$ the cluster of minimum size among all clusters of the different layers. Thus the largest number of broadcast rounds in any layer is dominated by

$$\left\lceil \frac{n}{|c_{\min}|} \right\rceil |c_{\max}| * |c_{\max}| \log(|c_{\max}|)$$

Since P_r is the number of layers after clustering, we can easily deduce that the total number of broadcast rounds after clustering cannot exceed

$$P_r \left\lceil \frac{n}{|c_{\min}|} \right\rceil |c_{\max}| * |c_{\max}| \log(|c_{\max}|) \quad (2)$$

Our partial sorting approach uses a list of separator for the construction of intervals of search: let S be the set of separators and $|S|$ the number elements in this set. According to (1) and (2), we can conclude that the number of rounds of our sorting partial algorithm cannot exceed:

$$O(v) + |S| * P_r * \left\lceil \frac{n}{|c_{\min}|} \right\rceil |c_{\max}| * |c_{\max}| \log(|c_{\max}|)$$

C. An example

The input file contains 20 elements, which are {4, 19, 3, 2, 18, 11, 1, 20, 12, 6, 10, 10, 13, 5, 16, 8, 17, 9, 14, 7, 15}. Let us suppose we want to sort the largest first 9 elements. Let us also consider the example of topology shown in Figure 1 for a good understanding of the construction method. Subsequent to partitioning to several level we obtain Figure 2, where we find a tree composed of a set of clustering that are: {1_0, 2}, {3_0, 4, 11}, {10_0, 12, 13}, {14_0, 15, 16, 17}, {5_0, 6, 7, 8, 9}, {18_0, 19, 20}, {1_1, 3, 18}, {5_1, 10, 14}, {1_2, 5}, the index represents the layer in the tree partition.

From here, it is easy to set up that any node, at the end of the algorithm, finally got to take but single cluster. Indeed, the identifier of the cluster to which the node is reattached is either holds the value of the node itself, or the greatest value holds by its neighbors. It is also important to note that with this algorithm, even in a cluster, two nodes are at most at a distance of *2* hops from each other. For this, it is enough to consider several nodes of the same cluster. Each node must be able to directly reach the Clusterhead of his cluster. Thus, two nodes of the same cluster must be at distance of at most *2* from one of the other.

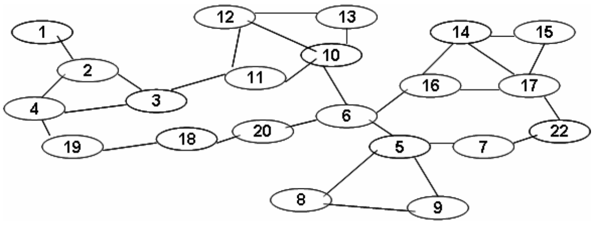


Fig. 1. Network that represent the data set.

As already mentioned above, our algorithm proceeds through the hierarchical partitioning of the network, which corresponds to the whole data elements. The following figure shows the results of the hierarchical clustering applied to the network, which represents the whole data's.

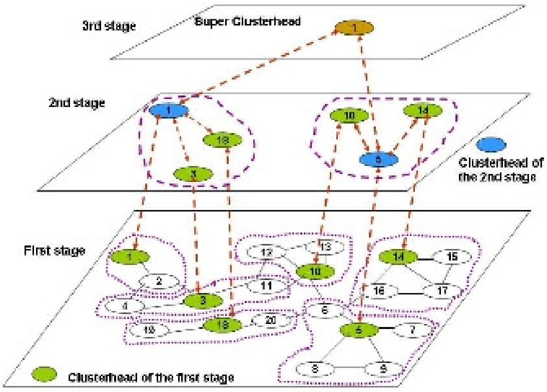
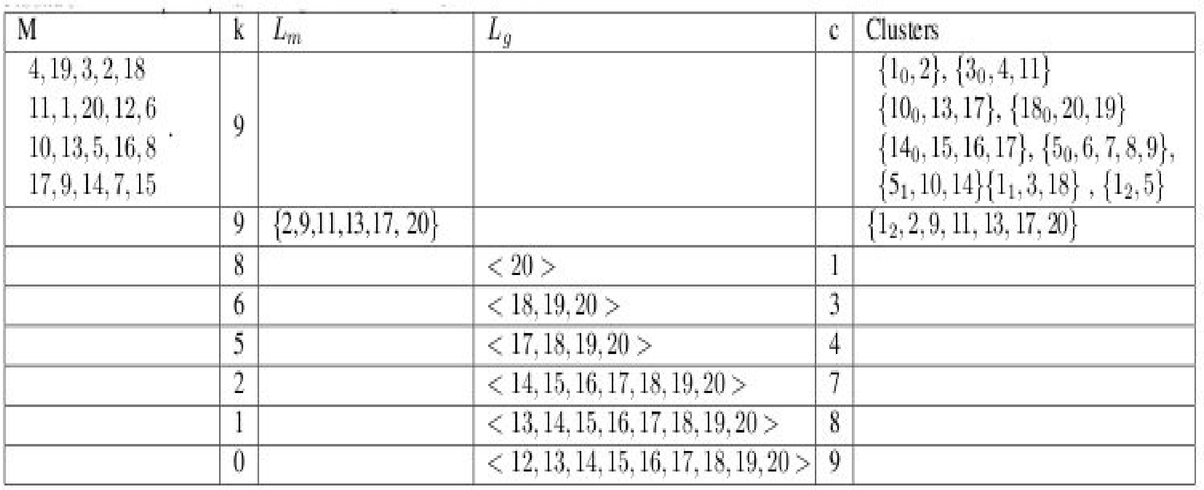


Fig. 2. The hierarchical tree corresponds to the data set.

For example, the input file contains 20 elements, which are {4, 19, 3, 2, 18, 11, 19, 20, 1, 6, 10, 13, 5, 16, 8, 17, 9, 14, 7, 15}. Let us suppose we want to sort the first 9 greatest elements. The whole process is shown in Fig. 3. After the execution of the algorithm *MaxCluster*, we find the greater elements, which are 20, and the list separator composed is as follows: ((2, 9, 11, 13, 17, 20). Once have the separator list, we can easily construct intervals for the rest of elements to be sorted. These intervals are: [17 20], [13 17], [11 13], [9 11] and [2 9]. The algorithm of partial sorting is called in a recursive manner with each of these precedent intervals. For the first interval, we find the two elements 18, 19 that will be injected into the set containing the sorted sequence. The partial sorting algorithm stops once the size of the whole stuff $L_g = k$.



M	k	L_m	L_g	c	Clusters
4, 19, 3, 2, 18 11, 1, 20, 12, 6 10, 13, 5, 16, 8 17, 9, 14, 7, 15	9				$\{1_0, 2\}, \{3_0, 4, 11\}$ $\{10_0, 13, 17\}, \{18_0, 20, 19\}$ $\{14_0, 15, 16, 17\}, \{5_0, 6, 7, 8, 9\},$ $\{5_1, 10, 14\}\{1_1, 3, 18\}$, $\{1_2, 5\}$
	9	{2,9,11,13,17, 20}			$\{1_2, 2, 9, 11, 13, 17, 20\}$
	8		< 20 >	1	
	6		< 18, 19, 20 >	3	
	5		< 17, 18, 19, 20 >	4	
	2		< 14, 15, 16, 17, 18, 19, 20 >	7	
	1		< 13, 14, 15, 16, 17, 18, 19, 20 >	8	
	0		< 12, 13, 14, 15, 16, 17, 18, 19, 20 >	9	

Fig. 3. An example of partial sorting algorithm with 20 elements.

IV. GENERALIZATION OF OUR PARTIAL SORTING ALGORITHM ON THE MODEL MULTI-HOP

In this section we genralize our algorithm the general soritng problem on the multi-hop. For this, we consider that the sensors the network can contain multiple data items. These data items must be sorted in a linked list. So we end up with a system which is composed of p sensors, in which the a entire sequence of n elements such that p < n is stored.

For this generalization, we model our network sensors such that each node contains three-linked list of data items. These lists are L_c, L'_g and L_g. L_c is the list of data items contained in their clusters and L_g is the list that contains the greatest data items in the clusters.

The algorithm of generalization (L_{gG}, *c*) is based on our algorithm for the classical sorting problem part. It assumes that each node in the network contains one and only one data item. The algorithm for generalized sorting problem uses the partial search method developed earlier in *MaxCluster*.

We can describe the stages of our algorithm as follows:

STEP 1: Hierarchical Clustering the networks by using algorithm Gerla & Tsai;

STEP 2: The partial sorting algorithm (L_{gG}, *c*) = (L_g, *c*)= *Partial-sorting (M, k).*

V. SIMULATION RESULTS

To evaluate the accuracy of our algorithms compared to the results of the theoretical analysis that exists in the literature for sort partial, we perform simulations of our approach. Our algorithm has been executed on the platform OMNET ++ 4.1 with a network of sensors. The sensors are distributed in a uniform manner. Fig. 4 show the simulation of our approach to sort partial a set containing 1000 data, each sensor has one and only data item.

For the generalized approach for sort partial, we have a variable p that represents the number of data items held by a sensor. Fig 5 show the results for *p=10*, *p=20* and *p=40.*

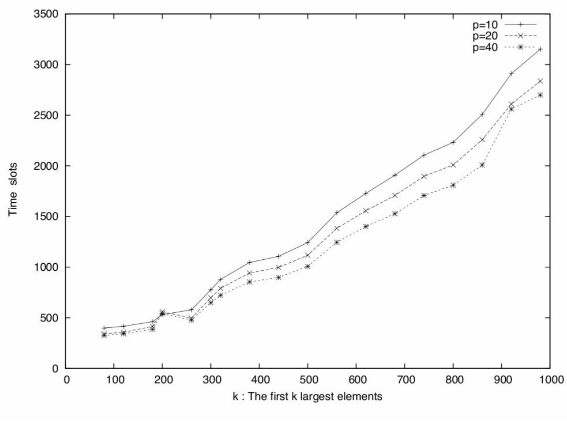


Fig. 4. Simulation performance of our sort partial approach with 1000 elements.

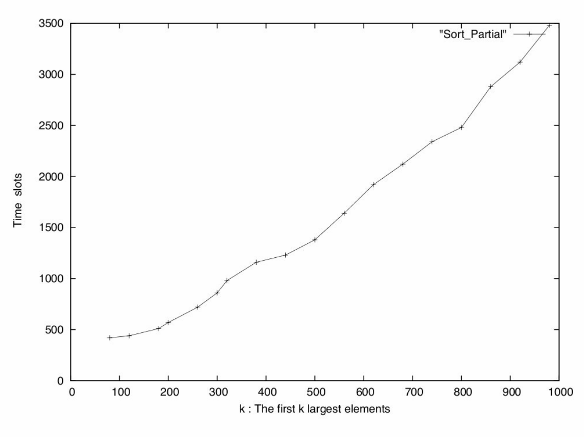


Fig. 5. Simulation performance of our generalized sort partial approach with 1000 elements and *p=10*, *p=20* and *p=40*.

The simulation results show that our algorithms match the theoretical bound. The simulations also show that when we need to sort the first k largest elements, not all elements of the set, our approaches can be good choices.

VI. CONCLUSION

The main motivation of this work is to give a solution of the partial sorting problem in multi-hop sensor networks. Our algorithms are the based on multi-layer clustering of the sensor network based on the algorithm of Gerla & Tsai. The first approach resolves the problem of partial sorting with a number of broadcast rounds not exceeding

$$O(n)+|S|*P_r*\left\lceil \frac{n}{|c_{\min}|} \right\rceil |c_{\max}|*|c_{\max}|\log(|c_{\max}|)$$

where

n is the number of sensors in the network

S is the st of separators

Pr is the number of layers obtained after the clustering process

C_{max} is the cluster of maximum size among all clusters of the different layers.

C_{min} the cluster of minimum size among all clusters of the different layers

The second algorithm provides a generalization of first approach to the classical sorting problem. This generalization for the partial sorting problem consists in assuming that each node that is part of the network may contain several data items.

REFERENCES

[1] J. L. Bordim, Koji Nakano, and Hong Shen Sorting on Single-Channel Wireless Sensor Networks. Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN.02). P. 133-138, 2002

[2] H. H. Chern and H. K. Hwang Phase changes in random m-ary search trees and generalized quicksortIn Phase changes in random m-ary search trees and generalized quicksort, Random Structures and Algorithms, Vol. 19, pp. 316358,2001.

[3] H. H. Chern and H. K. Hwang Transitional behaviors of the average cost of quicksort with median-of-(2t+1) In Algorithmica, Vol. 29, pp. 4469,, 2001

[4] R. Dechter and L. Kleinrock. Broacast communication and distributed algorithms, IEEE Transactions on Computers, C-35, 210-219, 1986

[5] M. Gerla and J T.C. Tsai. A "Multicluster, Mobile, Multimedia Radio Network ", Wireless Networks, vol. 1, no.3, p. 255-265, 1995

[6] C. A. . Hoare Find, (algorithm 65). In Communications of the ACM, Vol. 4, pp. 321332, 1961.

[7] J. JaJa A perspective on quicksort,In Computing in Science and Engineering,2000.

[8] J. M. Marberg and E. Gafni. Sorting and selection in multi-channel broadcast neworks, ICPP, pp. 846-850, 1985

[9] C. Martinez Partial quicksort,In Proc. of the 6th ACM-SIAM Workshop on Algorithm Engineering and Experiments and the 1st ACM-SIAM Workshop on Analytic Algorithmics and Combinatorics, pp. 224228, 2004.

[10] K. Nakano, S. Olariu and J. L. Schwing Broadcast-Efficient protocols for Mobile Radio Networks. *IEEE T.P.DS.*, vol.10, pp.12, 1276-1289, 1999

[11] H. Prondinger Multiple quickselect - hoares nd algorithm for several elements In Information Processing Letters, Vol. 56, No. 3, pp. 123129, 1995.

[12] R. Sedgewick Algorithms In C. USA : Addison-Wesley Publishing Company, 3 ed., 1998.

[13] R. Sedgewick and P. Flajolet An Introduction to the Analysis of Algorithms. In USA : Addison-Wesley Publishing Company,1996.

[14] Shyue-Horng Shiau *and* Chang-Biau Yang. Generalization of Sorting in Single HopWireless Networks. IEICE Trans. Inf. annd Syst., Vol. E89–D, no.4 , p. 1432-1439, 2006

[15] C. B. Yang, R. C. T. Lee and W..-T. Chen. Conflict-free sorting algorithms under single and multi-channel broadcast communication models. ICCI, LNCS 497, P. 350-359, 1991.

Energy Efficient Clustering Algorithms for Wireless Sensor Networks

S. Cui and K. Ferens
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada
umcuis@cc.umanitoba.ca, ferens@ee.umanitoba.ca

Abstract— *Four novel methods for improving the energy efficiency of clustering algorithms for wireless sensor networks are presented. The first method uses a single parameter, residual energy, to determine cluster head suitability in a passive cluster head election process; this is intended for energy constrained applications. Second, gateway nodes are allowed to distribute their sensor readings across different cluster heads, to better balance the service load. Third, this paper shows that up to a maximum of six new re-elections may only be required due to a single cluster head resignation, and, therefore, the algorithm is scalable. Forth, a sensing window is applied in TDMA slots to improve the energy efficiency of intra-cluster communications. MATLAB simulations show that the proposed algorithm has longer lifetime than basic passive clustering, and has improved intra-cluster communications scheduling (i.e., better energy efficiency).*

Keywords- Wireless Sensor Network; passive clustering; energy efficient clustering, re-election bound, TDMA.

1 Introduction

Recent discoveries of alternative and relatively less expensive sensor technologies, developments of energy efficient digital controllers, and advances in low-power radio frequency (RF) devices have given rise to the application of the Wireless Sensor Network (WSN) to environmental monitoring and control. In remote applications, such as lake water quality monitoring, a WSN consists of a large number of spatially distributed autonomous sensors nodes, each of which consists of a set of sensors, a controller, and wireless radio equipment. These nodes monitor their immediate surroundings (e.g., dissolved oxygen and pollutants) and cooperatively pass their data through the network to a central destination (e.g., a scientist's computing device). Due to the remote nature of the deployment, grid power is not available to the sensors, and, therefore, energy consumption is the most critical factor that must be carefully managed by the sensors and the communications protocols of the network.

Clustering has become a prominent approach to reduce energy consumption in these applications. In clustering, the network is arranged in clusters of nodes, where each cluster consists of member sensors, gateways, and a cluster head. Sensor nodes of a cluster send data to their cluster head, which performs data aggregation, such as data compression, suppression, min, max, or averaging. Cluster heads transmit aggregated data through the network of gateways and other cluster heads to a central destination. In this way consumption of energy is reduced, since, if every sensor node were to transmit data directly to a central destination, more intermediate relay nodes would be required to relay data (and consume energy to do so). Furthermore, clustering and data aggregation eliminates data duplication, and so, reduces the amount of energy consumed otherwise. The data averaging function of the cluster heads also provides fault tolerance, minimizing the effect of failed sensor nodes or bad reads.

Clustering also facilitates load balancing and extends network lifetime. For example, if a cluster head's energy becomes depleted due to its tasks of intra-cluster communications, performing the aggregation function, and inter-cluster communications, the cluster head may choose to resign its position; new clusters may be formed; and, other nodes may become cluster head to relieve the current cluster head of its duties. In this way, nodes in the network share the duties of being cluster head based on some parameter. Accordingly, clustering strives to maximize the lifetime of the network by balancing the duties of being cluster head.

Finally, clustering is proposed because of its network scalability; many nodes can be added or removed from the network without significantly affecting the performance, because of the clustering architecture.

However, clustering algorithms have disadvantages and pose certain research challenges, such as protocol overheads for cluster maintenance (e.g., cluster formation, cluster member assignment, and cluster head selection). Other problems are to control cluster size, granularity, and density; frequency of cluster changes by the nodes and cluster re-election; minimizing interference and collision for intra- and inter-cluster communications; and the domino effect of cluster reformation.

This paper presents several methods to improve the scalability, load balancing, and energy efficiency of clustering algorithms. The paper proposes to use a node's residual energy, exclusively, in the passive cluster head election process; this minimizes protocol overhead, network traffic, and energy consumed. In addition, the proposed algorithm allows a gateway node to probabilistically distribute transmission of its sensor readings to multiple cluster heads, and, thus, the work load of cluster heads is better balanced. In addition, this paper gives an upper bound for the number of new cluster formations caused by a re-election; the bound is six, and, therefore, the algorithm is scalable. Finally, this paper also proposes a method to improve the energy efficiency of TDMA scheduling for intra-cluster communications.

The remaining parts of the paper are organized as follows: section II discusses related work and identifies the extensions and contributions of this work. Section III gives the details of the four methods to improve the energy efficiency of clustering algorithms. Section IV discusses the simulation experiments performed to compare the proposed algorithm with other leading algorithms. Finally, conclusions and future work are given.

2 Related Work

In [1] passive clustering was introduced. In a cluster head election, the node that becomes cluster head is the first node to broadcast a cluster head declaration message. All other nodes "hearing" this message cease to compete in the cluster head election process and automatically become cluster members. This paper proposes the same kind of passive election process as in [1], except that this paper reports on an energy efficient passive clustering by proposing to use a node's residual energy to determine when that node can make a cluster head announcement message.

In [2] [3], cluster head candidates compete to become cluster head for a given round. Candidates broadcast their residual energy to neighboring candidates, and the node with more residual energy becomes cluster head. While the proposed method also used residual energy as a parameter to determine suitability to become cluster head, this paper differs in that less protocol transmissions for the election process were used, since passive clustering requires only one packet to be sent to establish the cluster head.

In [4], a node uses both residual energy and Euclidean distance from sink to determine its suitability for becoming a cluster head. The work [4] requires that each source node to know the Euclidean distance from sink nodes, which may not be possible. For example, in a scenario where a helicopter sprinkles a battlefield with sensors, the sensor and sink locations are not known a priori. Furthermore, GPS may not provide sufficient resolution for accurate location estimates. In addition, nodes that are closer to the sink (in the Euclidean sense) may be chosen cluster heads more often, especially in the case of a prolonged and repeated sink. Consequently, a relatively higher density of cluster heads would form closer to the sink, causing quicker energy depletion in these nodes; this may result in a partition being created, which would effectively disconnect the sink from other sensors in the network, and, thus reduce network lifetime. This paper differs from the work in [4] since it does not use distance, but does use residual energy, exclusively, to determine cluster head suitability.

The frequency of cluster head re-elections and the "domino" effect of re-elections are critical potential problems in clustering. For example, as discussed in [1], a potential problem in weight based clusterhead selection algorithms is that a single re-election can spread throughout the entire network, causing a total reconfiguration of all clusters. In each of [2][3][4][5], the number of clusters that need to be reformed due to a single cluster head resignation was not discussed. This paper determines an upper bound on the number of clusters that may need to be reformed due to the resignation of a single cluster head.

Basic TDMA schedules have been introduced for intra-cluster communications for sensor networks [1]. TDMA scheduling provides, ideally, collision free communications between cluster members and the cluster head. A TDMA schedule can avoid idle listening times, since nodes know when to transmit and when they may go to sleep. Other works have introduced variants of the basic TDMA schedule, to make the schedule more energy efficient. For example, [6] dynamically adjusts the length of a TDMA frame according to the amount of active nodes within a cluster, and, thus, reduces idle listening time for nodes which are not active. However, when schedules are changed, they must be re-communicated to the members (as schedule updates), and this may be potentially expensive in terms of energy and bandwidth, especially for frequent node activity changes. This paper proposes to create a more energy efficient TDMA schedule by allowing a cluster head to switch its radio off if it has not heard from a sensor node at the beginning fraction of its time slot. Compared with [6], the advantage of the proposed method is that schedule updates are not required when node activity changes, because the schedule does not depend on node activity.

3 Algorithm Description

Passive clustering is a cluster formation technique which requires a minimum amount of control packets [1] [5]. Conventional clustering algorithms require the exchange of control packets to negotiate which node becomes cluster head, while in passive clustering the cluster head is determined by the "first declaration wins" rule. In the proposed algorithm, each candidate node determines a waiting time before declaring itself cluster head. The waiting time is inversely proportional to the node's residual energy; the higher is the energy, the lower is the waiting time. Each node *i* which receives a cluster head declaration packet from node *j* will cease to compete in the election, and will assign node *j* its cluster head. The node *j* becomes the cluster head, since no other node in its radio range will make a similar declaration.

3.1 Cluster Head Load Balancing

A node *i* may receive a cluster head declaration packet from multiple cluster head nodes CH_k, since it is possible that the nodes *k* are out of radio range between themselves, but they could all be in radio range of node *i*. In this situation, node *i* becomes a gateway node, and can be used to relay messages from one cluster head to another, i.e., it may be used for inter-cluster communications as a relay node, relaying packets from CH_i to CH_j.

Unlike traditional clustering algorithms, the proposed algorithm allows a gateway node to be served by multiple cluster heads. This gateway node, which is also a sensor node, is allowed to apportion the amount of transmitted sensor readings to each of its cluster heads. The gateway node decides what portion of its sensor readings to send to a cluster head *h* based on the workload of the cluster head *h*. The greater the workload of a cluster head *h*, the lower the amount of sensor readings the gateway node will send to cluster head *h*. Thus, the data aggregation task of cluster heads is better balanced. Since the energy consumption will be better balanced amongst the cluster heads, this reduces cluster head resignation due to the energy constraint, and therefore, reduces cluster reformation.

To determine what portion of sensor readings to send to a particular cluster head, a gateway node *i* uses (1) to compute the probability of using CH_j. The gateway node *i* creates bins and orders them in the range from 0.0 to 1.0, where the width of a bin is given by the probability of using CH_j. Each bin is associated with the CH_j via an ID number *j*. When a sensor-reading packet is ready to transmit, the GW node *i* picks a random number, determines into which bin the number falls, and then uses the bin identifier to identify which cluster head to send the packet. Fig. 1 shows an example.

$$P_{ij} = \frac{\frac{E_j}{N_j}}{\sum_j^{N_i} \frac{E_j}{N_j}} \qquad (1)$$

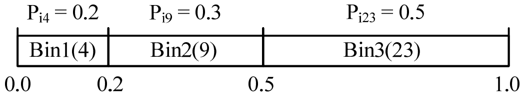


Fig. 1. Example of three cluster heads for gateway *i*. For instance, if the random number chosen is 0.45, it will fall in Bin2, and, therefore, the gateway chooses cluster head 9 to send its sensor readings for aggregation.

3.2 Re-election Upper Bound

A cluster head may need to resign, for example, because its residual energy drops below a safe operating threshold. Consequently, re-elections must occur to establish new cluster head(s) for the members of the resigned cluster. At least one new cluster must be formed, but there could be many more clusters and cluster heads needed, depending on which node takes over the role of cluster head. The concern is to minimize the effect of a resignation, to prevent a chain reaction of a resignation to spread to other parts of the network, and certainly, to avoid a network wide re-clustering. An election process can be expensive in terms of energy consumption, overhead bandwidth, and time and other resource usage. An upper bound on the number of elections caused by a single resignation is required.

The proposed algorithm inhibits the spreading of cluster reformation to other cluster heads in the network. As shown in Fig. 2(b), any neighbor cluster head (CHj) will be out of the radio range of a resigning cluster head CHi, and, thus, will not "hear" the resignation message from CHi. Neighbor cluster heads are out of range of a resigning cluster head because the algorithm does not allow more than one cluster head per cluster, and, furthermore, it specifies that all nodes have the same radio range. Accordingly, any neighbor cluster head will not take part in the re-election. Only the member nodes "hear" the resignation message, and only they may take part in re-elections. The spreading of re-elections is also prevented by allowing a node to be a member of more of than one cluster. This prevents a neighbor cluster head from initiating a re-election due to members migrating away from its cluster. Finally, the spreading of re-elections is also prevented by not allowing gateway nodes of the resigned cluster head from taking part in a re-election. If gateway nodes were allowed to take part in a re-election, and if a gateway node were to win an election, this would violate the one cluster head per cluster rule.

When a cluster head CH_i of cluster *i* resigns, at least one member node *k* of cluster *i* will need to become a new cluster head, and re-elections must occur. The problem of determining the maximum number of re-elections due to a single cluster head resignation becomes a problem of determining the maximum number of spheres required to cover all of the member nodes of the cluster in which the cluster head had resigned. The maximum number of spheres is determined by recognizing that the further (physical distance) a potential cluster head node *k* is to the resigned cluster head CH_i, the less original member nodes of CH_i will have found a new cluster, and the more re-elections will need to occur. Fig. 2(a) show an example with node n = 1 and 2 are potential cluster heads in a re-election. The worst case is when a node *k*, located at the periphery of CH_i's radio coverage, becomes a new cluster head.

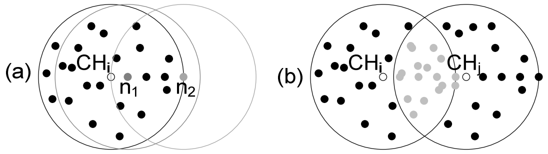


Fig. 2. (a) Shows that, as potential cluster heads, node n1 covers more sensor nodes than node n_2. (b) Shows that cluster head CH_j does not participate in a relection because it is out of range of the resigned cluster head CH_i. Gateway nodes (shaded grey) cannot take part in the re-election, since, if they had won, there would be two cluster heads in cluster j.

Therefore, the maximum number of clusters that will be formed due to a single cluster head resignation can be determined by considering that, in the worst case, member nodes near the periphery of the resigned cluster head's radio range become new cluster heads, since this presents the greatest available volume for new clusters. As given by (2), and depicted by Fig. 3, this number is 6.

$$\begin{aligned} &C = 2\pi\, r \Rightarrow 2\pi r > nr \Rightarrow 2\pi > n \Rightarrow n < 6.28 \\ &\therefore n = 6 \text{ (i.e.,) a maximum of six reelections may occur.} \end{aligned} \qquad (2)$$

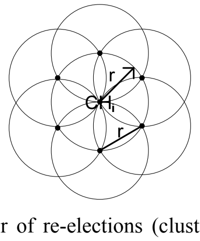


Fig. 3. Maximum number of re-elections (clusters) that may be required due to a single cluster head resignation.

3.3 Energy Efficient Intra-Cluster Communications

This paper also proposes a method to improve the energy efficiency of TDMA scheduling for intra-cluster communications. Many sensor nodes, which are members of a cluster, may not have data to transmit to the cluster head all of the time, simply because no change in environmental variables or the lack of requests from a sink. When no data is sent from a sensor node to cluster head during the node's slot in the TDMA schedule, the cluster head will be effectively wasting energy listening and waiting for sensor data.

To better conserve energy, each TDMA slot is divided into two windows: sensing and traffic windows (Fig. 4). During the sensing window, the cluster head senses whether there is transmission from the cluster member, which was allocated to this slot. If there is no transmission by the end of the sensing window, the cluster head switches off its radio during the traffic window in this slot in order to save energy. Otherwise, the cluster head receives packets from the cluster member during the traffic window. In addition, the proposed algorithm allows a cluster head to switch its radio off for the entire TDMA slot for a node that has gone to sleep. A node will have announced its intentions to enter a sleep mode due to its energy dropping below a certain threshold. During sleep mode, a node will initiate power harvesting, and wake up when its energy reaches a certain operating level. Nodes use a predefined control slot in the TDMA schedule for sleep/wake messages. The algorithm is called *TDMASense*.

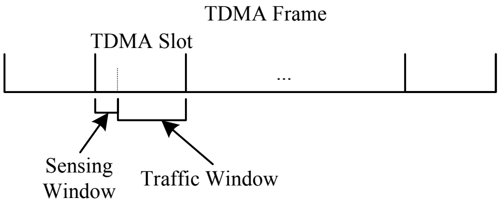


Fig. 4. Energy efficient addition to TDMA schedule.

3.4 Energy Efficient Clustering Algorithms

The proposed energy efficient cluster formation process is summarized as follows:

1. There are three possible states: Candidate Node (CN), Cluster Member (CM), Cluster Head (CH), GateWay (GW).

2. At the beginning of cluster creation, all nodes are in the CN state. Each node takes part in the neighbor discovery process to determine the number of neighbors each node has. In a static network, i.e., nodes are not mobile, neighbor discovery is performed at deployment time, and whenever new nodes are added or nodes are removed.

3. Each node *i* in the CN state will calculate its waiting time $T_{ci} = 1/random(norm, E_i, 1)$, where E_i is the residual energy of node *i*. Each node will wait T_{ci} time before transmitting a CH_ADV (cluster head advertisement message) message, declaring that it is the cluster head. The CH_ADV message sent by node *i* contains the node ID, residual energy E_i, and number of neighbors N_i. Note that the node with largest residual energy is the most likely to become CH. Neighbor nodes with the same residual energy E_i will be randomly differentiated.

4. Each Node in the CA state receiving the CH_ADV message will change its state to CM, and add the node ID of the CH_ADV message into its CH list.

5. If a node in CM state receives more than one CH_ADV packets in the same round, it will change state to GW. At this point, the network is stable until a CH needs to resign.

6. A CH may resign when its residual energy drops below an application defined threshold. The CH will broadcast a CH_RES packet to its members and gateways.

7. For CMs receiving a CH RES packet, they may go back to Step 2 and start a new process of clustering. No more than 6 new clusters can be formed for a previous cluster, as given by (2), and shown in Fig. 3.

The proposed energy efficient TDMA schedule for intra-cluster communications is summarized as follows:

1. The cluster head of *cluster i* divides a TDMA frame into N_i slots, where each slot is allocated to a CM or GW node of *cluster i*. Each cluster head broadcasts a cluster head schedule (CH_SCH) packet, which contains the CH's ID and slot allocation information, to members of *cluster i*.
2. Each CM node will use its TDMA slot to transmit data.
3. After GW node *i* receives all CH_SCH packets from its CHs, it uses (1) to compute the probability of using CH_j for sensor reading transmission and aggregation.
4. When a sensor reading packet is ready to transmit, the GW node *i* picks a random number, determines which bin the number falls, and then uses the bin identifier to identify which cluster head to send the packet.
5. If a cluster head senses no traffic in the sensing window, then the cluster head will switch off its radio in the traffic window of this slot in order to save energy.

4 Experiments and Results

A MATLAB simulation was created to test the energy efficiency of the proposed algorithm and to compare it with others. A *100m* by *100m* grid was created, and different numbers of nodes, ranging from *N = 25, 50, ..., 250* nodes were placed in this area. Each node was given equivalent resources; same type and speed of processor; same amount of memory for buffering; radio range of *R = 5m*; and initial energy of E=1x10^4 Joules. During the operational state of the network, the energy comsumption was set as follows: E_ctl=1 Joules for transmitting one control packet; E_dat=10 Joules for one data packet; and E_sng=0.1 Joules for neighbor detection.

4.1 Cluster Head Load Balancing

In the cluster head load balancing experiment, the lifetime of the proposed algorithm was compared with basic passive clustering [1]. The only difference between the proposed algorithm and basic passive clustering was that the former allowed gateway nodes to distribute their sensor readings across the cluster heads, of which they were members. The distribution of sensor readings was done according to (1).

Fig. 5 shows the comparison of the lifetime of the proposed algorithm with that of the basic passive clustering algorithm. The lifetime was defined as the total time before a single node ran out of energy and died [6]. The lifetime decreased with increasing node density for both algorithms. The decrease in lifetime with increasing node density occurred because as the node density increased, the density of gateway nodes increased at a faster rate than that of cluster head candidates. Since gateway nodes were not allowed to take part in re-elections (as explained in Section 33.2), cluster head duties were taken on by a relatively decreasing number of candidates. This resulted in an inevitable loss of energy of a single node and network expiration.

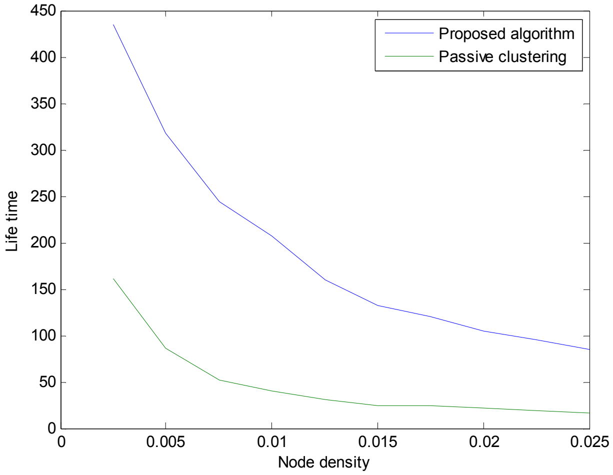


Fig. 5. Life time of the network (in unit of rounds) V.S. Node density (number of nodes/m2).

The decrease in lifetime with increasing node density is a limitation of both algorithms since both do not allow gateway nodes to take part in re-elections. A possible solution is to allow gateway nodes to take part in re-elections, but this would increase the number of required re-elections due to a single resignation, and a possible chain reaction of re-elections to the entire network could result.

As shown in Fig. 5, the proposed algorithm "lives" longer than basic passive clustering because the proposed algorithm balanced energy consumption between clusterheads, and this extended the lifetime of the network compared with basic passive clustering, which did not balance energy consumption. To show the relative balance of energy consumption, our next simulation measured the variance in energy consumption.

4.2 Residual Energy Variance

Fig. 6 shows the variance of residual energy of all nodes in the network taken at each epoch of the simulation. In both algorithms the variance was measured to be zero at the beginning of the simulation, since each node was given the same initial amount of energy. As the simulation progressed, nodes took on different duties, such as cluster head, gateway, and basic sensor duties, which consumed different amounts of energy, and, therefore, the residual energy variance increased. However, the residual energy variance in the proposed algorithm was always lower than that of the basic passive clustering algorithm at each epoch of the simulation and throughout the entire lifetime of the basic passive clustering algorithm. Finally, Fig. 6 also shows that the proposed algorithm lived on average approximately 4.5 simulation epochs longer than the basic passive clustering algorithm, due to the nodes having sufficient residual energy to carry on their networking tasks (cluster head, gateway, or basic sensor).

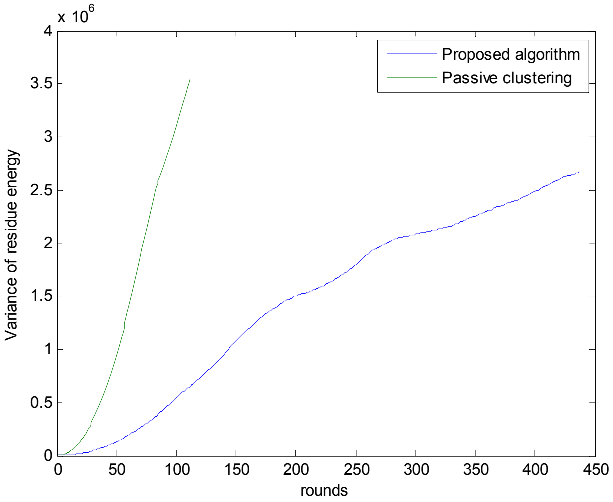


Fig. 6. Variance of residue energy at different steps of the simulation.

4.3 Energy Efficient TDMA Scheduling

The proposed algorithm TDMASense was compared with the dynamic frame size method EC-TDMA [6] for implementing intra-cluster communications. In the simulation, when *node i*'s energy dropped below a threshold, it broadcasted a sleep message to the cluster head, and then entered the sleep mode and commenced power harvesting. In response, TDMASense switched its radio off during *node i*'s slot to conserve energy. The EC-TDMA method created a new TDMA schedule, which excluded *node i*'s slot from the frame. In addition, EC-TDMA broadcasted the new schedule to the cluster members.

Similarly, when *node i*'s energy became replenished, it broadcasted a wake message to the cluster head. In response, TDMASense switched its radio on during at least *node i*'s *sensing window*, and kept the radio on for the *traffic window* depending on if *node i* had anything to transmit. The EC-TDMA method created a new TDMA schedule, which included *node i*'s slot into the frame. In addition, EC-TDMA broadcasted the new schedule to the cluster members.

An additional difference between TDMASense and EC-TDMA was that the latter will have lower intra-cluster communication delay for a smaller TDMA frame, which corresponds to a node entering sleep mode.

To compare the relative amount of energy consumed by the two methods (EC_TDMA and TDMASense), consider the amount of energy required when a node awakens from sleep mode and once again becomes an active member of a cluster. In this case, the EC_TDMA method requires creating and transmitting a schedule update, which size is proportional to the number of cluster members, while the TDMASense method requires only that the cluster head switch its radio on for at least the small fraction of the sensing window within the member's slot of the TDMA frame. Accordingly, in the simulation, the cost to transmit a schedule update packet was chosen to be ten times the cost of performing sensing in the sensing window. The 10:1 ratio is a conservative estimate, since there can be potentially a large number of cluster members; and, in many situations, the ratio may be higher.

As shown in Fig. 7, in the simulation, as the node density increased, energy consumption for both methods increased. The proposed algorithm TDMASense's energy consumption increased with increasing node density because sensing cost was proportional to the number of cluster members. The EC-TDMA method's energy consumption increased with increasing node density because an increasing number of nodes entered sleep mode or awaken mode. Consequently, the EC-TDMA method was required to send an increasing number of schedule updates.

In addition, as the node density increased, the energy consumption of the EC-TDMA method increased faster than that of TDMASense. When a node entered sleep mode, the cluster head of the TDMASense method switched its radio off for the entire slot allocated to this node, thus reducing energy consumption; while the EC-TDMA method needed to consume energy to transmit a new schedule. As the frequency of nodes entering sleep mode increased with increasing density, the energy consumption difference between these two methods became larger. According to the simulation, the TDMASense method outperformed the EC_TDMA method for high-density nodes.

5 Conclusions and Future Work

This paper proposed four methods to improve the energy efficiency of passive clustering for wireless sensor networks. First, residual energy was used to determine the cluster head in a passive cluster head election process. Second, gateway nodes distributed their sensor readings across different cluster heads to load balance the tasks of the cluster heads. Third, the maximum number of new clusters reformed due to a re-election was shown to be bounded by six. Forth, the energy efficiency of TDMA scheduling for intra-cluster communications was improved. MATLAB simulations show that the proposed algorithm has longer lifetime than basic passive clustering, and has more efficient TDMA scheduling, at the cost of larger delay for intra-cluster communications.

While the TDMASense algorithm shows a potential improvement over the dynamic frame size method, more work can be done in the area of optimization. For instance, the optimal size of the sensing window may be determined by modeling the window size using queuing theory. If the sensing window is too small, a node may miss its opportunity to transmit, and then must wait until the next TDMA frame. If the sensing window is too large, then the cluster head must keep its radio on longer, and the benefits shrink. By modeling the probability of a node transmitting in a sensing window, the optimal size may be determined, and it may also be dynamically adjusted according to the level of activity of a node.

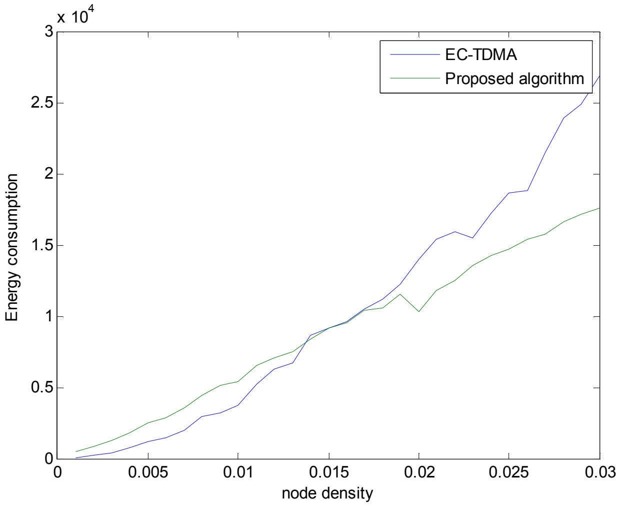


Fig. 7. Comparison of energy consumption between TDMASense and EC-TDMA for different node densities.

References

[1] M. Gerla, K. Taek Jin, and G. Pei, "On-demand routing in large ad hoc wireless networks with passive clustering," in *Wireless Communications and Networking Conference*. vol.1, no., pp.100-105, 2000.

[2] M. Ye, C. Li, G. Chen and J. Wu, "An Energy Efficient Clustering Scheme in Wireless Sensor Networks," *Ad Hoc & Sensor Wireless Networks*, vol.1, pp.1–21, 2006.

[3] O. Younis and S. Fahmy, "HEED: "A Hybrid Energy-Efficient Distributed Clustering Approach for Ad Hoc Sensor Networks," *IEEE Trans. on Mobile Computing*, vol. 3, no. 4, pp. 366- 379, 2004.

[4] Md. Mamun-or-Rashid, M. M. Alam, C. S. Hong, "Energy Conserving Passive Clustering for Efficient Routing in Wireless Sensor Network," *The 9th International Conference on Advanced Communication Technology*, vol.2, no., pp.982-986, 12-14 Feb. 2007.

[5] T.J., Kwon and M. Gerla. "Efficient flooding with passive clustering (PC) in ad hoc networks." *ACM SIGCOMM Computer Communication Review*. 32 (2002) 44–56, 2002.

[6] M. Xie, X. Wang, "An Energy-Efficient TDMA Protocol for Clustered Wireless Sensor Networks," *ISECS International Colloquium on Computing, Communication, Control, and Management*, Vol.2, No., pp. 547-551, 3-4 Aug. 2008.

[7] I. Dietrich and F. Dressler, "On the lifetime of wireless sensor networks," *ACM Trans. Sen. Netw.* 5, 1, Article 5, pg. 39, Feb. 2009.

FINITE STATE AUTOMATA BASED FAULT TOLERANT AND ENERGY EFFICIENT ROUTING (FSA-FTEER) IN WIRELESS SENSOR NETWORKS

N.Arun[1], P.Venkata Krishna[1], Aditya Ahuja[1] and V.Saritha[1]

[1]School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India

Abstract— *With the availability of low power, tiny and inexpensive devices, wireless sensor networks have been used for a wide variety of applications ranging from industrial to environmental monitoring and military uses. Wireless Sensor nodes are mainly used in mission critical applications and deployed in sensitive areas due to which they can't be replaced frequently. Certain situations may arise where the nodes in network may not function properly due to failures, and they become unavailable to the network. There is much need to develop energy efficient and fault tolerant routing for wireless sensor networks. We propose a (FSA-FTEER) protocol which creates multiple node-disjoint paths from source to destination and ensures that paths don't interfere with each other. Residual energy of sensor nodes is taken in to consideration as heterogeneous nodes are deployed in network.*

Keywords- Finite State Automata, Probability, Congestion, Interference, Faulty Nodes

1. Introduction

Wireless Sensor network's (WSN) are mostly deployed in hazardous areas to detect any change in environment and send information to nearest base station for certain action to be taken. These WSN's are used in wide variety of applications such as home automation, vehicle tracking, target detection, control of actuators, etc. Their (WSN's) most significant work has been in environmental monitoring (air, soil, temperatures, etc) and military applications where topology of the network, energy-efficiency, and fault tolerance of the network play a great role. With a rapid growth in technology and cost reducing measures used sensor nodes are mostly available for low cost. There is also great increase in demand of WSN's due to low-processing power, energy efficiency, tiny devices and the type of environments in which they operate. The communication between them is mostly distributed in nature as opposed to centralized framework .In this sense they consume very less energy in terms of processing the data and memory of node will be utilized efficiently. The main role of these nodes is to detect any changes in deployed area and use a routing protocol to transmit data to base station. The paths from node to base station (sink) are node-disjoint and multi-path in nature. Node-disjoint path will help us to send data even though there are some faulty nodes in network. Multi-path mechanism is used to reduce congestion in network and it provides load balancing. The multi-path is calculated taking intermediate link information between the adjacent nodes. These node-disjoint multi-paths should be non-interfering since interference of the paths will make node to be in active state and listen to unnecessary data which results in excess energy consumption. Congestion in the network will reduce life time of the network since lost packets need to be sent again. The bandwidth of network will be over-utilized and latency of packets will be increased. Certain congestion control mechanisms must be used such as adjusting of data rates or splitting of packets in to the available multi-paths in the network.

Based on application requirements and environmental conditions topology of network is important since nodes are highly mobile. With this mobility nodes will change their positions with respect to network and, there is a high need to guarantee certain delivery rate. Network consists of different heterogeneous nodes and these nodes vary in energy consumptions. Hence residual energy of intermediate nodes should be taken in to consideration while measuring efficiency of the path. Many fault tolerant [1-10] routing algorithms are proposed in WSN. Most of them are based on TDMA where fixed slots are allocated for data transmission and the number of faulty stations was assumed to be very small. They are mostly based on cross layer optimization approach with delay and other routing metrics. Network was assumed to be static in most cases. Energy efficient routing protocols [11-27] were based on geographic, tree, coverage ratio, flooding, cross layer strategy, fusion and slot based approach. Most of these existing protocols did not consider mobility and heterogeneous factors in them.

2. Related work

Fault-tolerant routing protocols ensure that network remains connected and communication takes place in presence of certain faulty stations. A K- degree Anycast Topology Control (k-ATC) [1] protocol was proposed where network functions effectively when there are k-1 sensor nodes. This protocol mainly consists of 3 sub-algorithms i.e. (i) greedy routing algorithm will minimize the transmission range (ii) distributed algorithm ensures that connectivity of nodes is met and (iii) k-approximation algorithm. However the constraint was protocol is unidirectional. Apart from this many other routing protocols [2], [3], [4], [5] were proposed in WSN which are fault tolerant and network is k-vertex connected. Communication happens in presence of faulty nodes which is based on mixed integer programming and it's both unidirectional and bidirectional.

A Fault tolerant protocol on permutation routing [6] was proposed where p stations in network are sender and receiver of (n/p) packets. Permutation routing problem is single hop routing and, each station is in transmission range of all other stations which are present in network. The protocol is suitable for unbalanced routing permutation where each of individual nodes has unequal no of packets. However in this protocol it was assumed that no of faulty stations were small in number.

Flow control is necessary in WSN to maintain end-end delay of packet and to avoid congestion in the nodes so that packet drop rate can be reduced. A Dynamically configurable message flow control [7] for adaptive routing was proposed which is based on pipelined networks. Flow control mechanisms are used at lower layers and 2-phase routing protocols are used to avoid faulty nodes and deadlock configurations. DFT-MSN [8] is fault-tolerant routing algorithm in WSN where it includes 2 phases i.e. synchronous and asynchronous phase. In initial phase sender will identify its neighbors to transmit packets and in second phase connection is established and sender will gain control of channel for delivery of packets. High delivery rate of packet is achieved and delay of the transmissions is reduced. Acknowledgements of packet will be delayed since second phase is asynchronous.

FLEXI-TP [9] is based on TDMA scheduling. Nodes wakeup in their scheduled slots for transmission of data and sleep for rest of the time. Nodes can build, modify and extend the schedule of slots based on local information available to them. It uses depth first schedule due to which buffering of packets will be reduced and communication slots are re-used by nodes which don't fall in each other's interference range. End-to-End data delivery rate is achieved while keeping memory constraints in consideration.

Life time of network is crucial in WSN due to its nature of use and complexity involved in replacing node or battery. Energy Efficient protocol is needed in WSN while routing and QOS requirements must be met. AFST [10] was proposed for reducing redundancy in network and achieves load balancing. Based on energy of node it decides whether a fusion can be performed at a given node or not. This protocol follows a tree approach for fusion of data and can be merged with designing of cluster based algorithms. [11] routing algorithm considers connectivity of remaining sensor network. It specifies importance of certain nodes whose disintegration will disrupt entire network. Importance of node is specified by Fiedler value of remaining network when a node expires. Each of nodes is associated with cost/metric for routing and proposes a keep-connect routing algorithms which use computable measures of network connectivity.

EECCR [12] suggests a m-coverage n-connectivity problem. This is based on heterogeneous sensor nodes where nodes are deployed in a given area. The algorithm does not need any location information of nodes. Nodes in each of scheduling set are activated periodically and energy balancing of nodes is done to prolong lifetime of network. Latency of packets is reduced and reliability of transmission is increased. This is mostly based on slot based approach. AsOR [13] is a unicast routing technique for multi-hop wireless sensor networks over different channels. The protocol has 3 different nodes namely frame nodes, assistant nodes and unselected nodes. Frame nodes are used to decode and forward a packet. Assistant nodes are used to provide security of un-successful transmissions. The protocol defines a value N i.e. minimum no of nodes which are required for transmission from source to destination. As a result Energy consumption of network is reduced.

The above protocols are mostly based on tree approach and require a slot based mechanism to transmit data. Scheduling of slots and varying them according to the given data will consume certain time and energy. Nodes use sleep and awake mechanism for transmission of data. When data is not available for that particular slot then the node has to remain idle and certain energy of the node is consumed.

Location based algorithm [14] was proposed for WSN where network is divided in to virtual grids. This helps us to reduce redundancy of data. Nodes are associated with GPS, but in real-time it's impractical to associate each and every node with GPS since a certain amount of overhead and energy is consumed with finding the location of node. Mobility, heterogeneity and faulty node issues have not been considered here. EBGR [15] selects its next node based on energy optimal mechanism as the relay node. The node uses RTS and CTS mechanism to route data. But protocol does not consider loss of packets during transmission.

Electing of clusters and cluster heads in WSN's is crucial since they involve a lot of energy and computation. An energy efficient and dynamic clustering protocol [16] was proposed. Based on signal strength received from neighboring node, each of individual nodes will compute probability of becoming cluster head. Certain factors like mobility of node, faulty nodes and heterogeneous nodes have not been specified here. A centralized routing algorithm (BCDCP) [17] was proposed for WSN's. This is a cluster based algorithm where each of cluster head is assigned equal no of nodes so that load-balancing is equal among cluster heads. However nodes were assumed to be immobile and base station is fixed.

REEP [18] is data centric, energy efficient, reliable routing protocol for WSN's. Packets are divided in two categories i.e. sense packet and information packets. Nodes process them based on their packet types. When processing packet, each node will verify residual energy present in node and if energy is below a given threshold status of the packet will be changed to negative and will be forwarded to next node or previous node for processing data. Other protocols such as [19], [20] explain about communication between sensor networks and actor networks. Communication between actors network will take place with help of sensor networks which are deployed in network. Sink is confined to points where energy of network can be utilized in an efficient way.

Some of other protocols are present in WSN [21], [22], [23], [24], [25], [26], [27] are energy based routing. Most of above protocols are either based on flooding, clustering, gradient or geographic based approach. Geographical approach will increase cost of network. In clustering approach inappropriate load balancing will consume energy and there is also an overhead for selecting new cluster head when old cluster head's threshold reaches a particular level. Flooding based mechanism will achieve reliability of transmission at cost of duplicate packets in to network. As a result cost associated with network is more or energy of network is over utilized in these circumstances. Hence we propose a protocol based on following factors 1) heterogeneous nodes 2) congestion 3) interference 4) mobility of the nodes 5) fault tolerant issues 6) residual energy of the nodes

Here we propose FSA-FTEER protocol which is based on finite state automata and is both fault tolerant and energy efficient. Routing algorithm finds a node-disjoint multipath where paths don't interfere with each other. Proposed protocol will consider faulty nodes and data are routed with help of multi paths which are found in network. The protocol will maintain information about neighboring nodes for routing and global topology of network is not needed. Mobility of nodes is considered and congestion is controlled based on load balancing of paths. Optimal Value of link is calculated by considering residual energy of node.

3. Proposed solution

3.1 Finite state automata

A total of 5 states are defined in FSA with respect to a node. SLEEP state specifies that the node is in power saving mode and no operations in the node are performed. A node is in SLEEP state when it's idle. ACTIVE state specifies that a node is processing data received by it or data generated because of event detection. TRANSCEIVER is a state where node is sending or receiving data. This is a state where maximum energy of the node is utilized. ERROR state specifies that node is a faulty node and it can't be used for any kind of transmission or processing of data. MOBILITY state specifies that node is moving. Mobility of the nodes can be achieved using some random way point model. States of the node are represented using a 3- bit sequence.

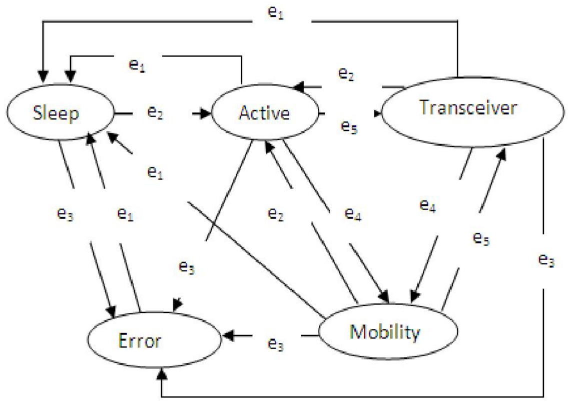


Fig.1: States of Node with transitions among them

Finite State Automata will act as a decision maker while routing is taking place. Based on parameter values which are defined in node and state in which the node is present, current node finds next best optimal node (i.e. neighbor node) for transmission of data. At a given instant of time, a node will be in a single state. Transition from one state to another is based on the type of operation performed by the node on received data. States of node are represented using a 3- bit representation. Bit representation of nodes will help to form non interfering node-disjoint multipath routes.

Table 1 Transition Table

δ	e_1	e_2	e_3	e_4	e_5
000	u	001	100	u	u
001	000	u	100	011	010
100	000	001	100	011	u
101	000	001	100	u	010
110	000	u	u	u	u

Above is the state transition table for the transition function δ. e_1 = idle, e_2 = event detected, e_3=error occurred, e_4=mobile, e_5=packet sent/received.

Deterministic Finite Automaton:-

For DFA define the set of states S = {000, 001, 010, 011, 100} and events $E= \{e_i : 1 \leq i \leq 5\}$. Then $\delta : S \times E \rightarrow S$ is a partial function given by the transition table. Entries labeled u are undefined.

Table 2 Bit Representation of Finite State Automata

NODE_STATE	BIT REPRESENTATION
Sleep	000
Active	001
Transceiver	010
Mobility	011
Error	100

Table 3 Parameters of node with respect to state

Parameter	Abbreviation	Description
Node Id	**NID**	Each of the nodes in the network is represented by a unique node id.
Type of Node	**TON**	Specifies type of node as network will consist of heterogeneous nodes deployed in the area.
Total Energy	**TE**	Initial energy present in node.
Residual Energy	**RE**	Remaining energy present in node.
Energy used for current Activity	**ECA**	The activity can be in-node data processing, sending receiving of data.
Threshold Energy	**TRE**	Residual energy to which node can process data or perform transmissions.
Current State	**CS**	Specifies current state of node.
Type of Packet	**TOP**	Packet can be a data packet or control Packet.
Length of Packet	**LOP**	Specifies length of packet.
Priority of Packet	**POP**	Specifies priority of packet
Path ID	**PID**	Specifies a path id of packet. This path id is stored in all nodes which a data packet traverses.
Path Group ID	**PGID**	Specifies path group id. Path group id is for, set of packets generated from same source for a particular event, where packets traverses in different paths.

3.2 Routing Algorithm

Initialize

Initialize the complete network where all nodes know their positions in terms of direction with respect to Base Station (Sink). Initialize parameters of node which are important in decision making while finding the routing paths

α_{TE} =10,000 mW
α_{TRE}=1000mW
α_{I}= 2mW
α_{EUS}=15mW/32bit
α_{EUR}= 30mW/32bit

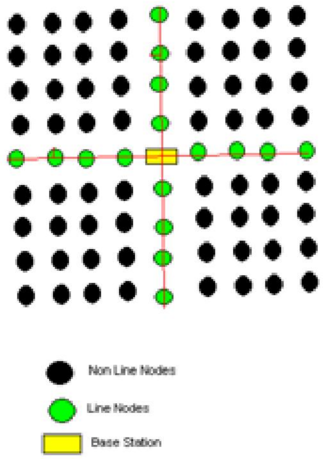


Fig.2: An Example Network

Input

Node with its parameter values at the given instant of time when the packet is ready to arrive or leave and the value of contents which are present in the packet

Output

Routing path with optimal weight to route the packets from source to destination

Steps

1: Repeat
2: Remove the Packet from Queue
Process the Packet
3: Identify neighbor nodes and send a request to the entire neighbor nodes for identifying its position with respect to Base station
4: Based on Response received from neighbor nodes calculate the probability of direction
5: Direction probability:-
Define the set of directions D as D = {N, E, W, S, NE, NW, SE, SW}
Now the probability of direction d ε D is given by

$$\Pr[d]=\frac{\sum_p 1_d(p)}{\sum_p \sum_{d' \in D} 1_{d'}(p)} \quad (1)$$

Here 'p' is the received packet and '1_x' is the indicator function in direction 'x'. The above indicator functions, for more accuracy, may be weighted with a function which is inversely proportional to euclidean distance of the node sending 'p', to the source node. This distance may be

```
approximated using timestamps (to be incorporated in the packet).
6: Select nodes which are in direction of max probability
7: Define F = {$\alpha_{RE}$, $\alpha_{ECT}$, $\alpha_{POP}$, $\alpha_{QL}$} where F denotes the family of parameters based on which the node calculates the optimal weight.
8: Find weight of paths from source node to neighbor node
9: $W(n_{adj})=\sum_{S \varepsilon F} R_S(S(n_{adj}))$ where $n_{adj}$ ε $Adj(n_0)$        (2)

   $W(n_{adj})$ represents the weight of adjacent node.
10: Select the node with optimal weight for transmission of Data
11: Send RTS and start a timer
12: if (CTS received before timer expired)
$n^{(1)}$ = Min {$W(n_{adj})$ : $n_{adj}$ ε $Adj(n_0)$}        (3)
where $n^{(1)}$ is optimal node used for transmission of data.

13: else
$n^{(2)}$ = Min {$W(n_{adj})$ : $n_{adj}$ ε ($Adj(n_0)$- {$n^{(1)}$}) }        (4)
where $n^{(2)}$ is optimal node used for transmission if $n^{(1)}$ is not ready to receive the data.
14: end if
15: until no more packets left in the Queue
```

Table 4 Family of parameters used for finding optimal node.

PARAMETER	DESCRIPTION
α_{RE}	Residual Energy of the node
α_{ECT}	Energy used for Current transmission activity
α_{POP}	Priority of Packet
α_{QL}	Queue Length Present in the Node
α_{TE}	Total energy of the node
α_{I}	Energy utilized by node when its idle
α_{EUS}	Energy used for sending data
α_{EUR}	Energy used for Receiving data

3.3 Packet format

Table 5 Structure of Message Packet

SID	DID	TOP	PID	PGID	DATA	DIRT	SEQ NO	LOP
0 6	7 13	14 15	16 18	19 21	22 53	54 56	57 62	63 68

Table 6 Description of Packet Format

CONTENTS	DESCRIPTION	NO OF BITS NEEDED FOR REPRESENTATION WITH THEIR RANGE
SID	Specifies Source ID of node which is sending packet.	7 [0,6]
DID	Specifies Destination ID of node for which packet is to be received.	7 [7,13]
TOP	Type of Packet. It can be data packet, control packet or choke packet.	2 [14,15]
PID	Specifies Path ID of packet. This path id is stored in all nodes which a data packet traverses	3 [16,18]
PGID	Specifies Path Group ID. Path group id is for, set of packets generated from same source for a particular event, where packets traverses in different	3 [19,21]
DATA	Data which is sent from source to destination.	32 [22,53]
DIRT	Specifies Direction in which packet is being sent or received.	3 [54,56]
SEQ NO	Specifies Sequential No of packet.	6 [57,62]
LOP	Length of Packet. If length of packet is large in size it's fragmented in to smaller sizes.	6 [63,68]

3.4 Numerical evaluation

Let us consider that node at position (-5, 5) has just arrived there and it needs to know its position w.r.to base station in terms of direction. It then sends a broadcast packet to all of its neighbor nodes. These neighbor nodes will inform the source node that they are in NW direction with respect to base station. The source node will now calculate the probability of direction with respect to base station

$$\Pr[d]=\frac{\sum_p 1_d(p)}{\sum_p \sum_{d' \in D} 1_{d'}(p)} \qquad (1)$$

Since maximum probability of the direction of the base station is NE, the source node will fix its position to be NE with respect to the base station.

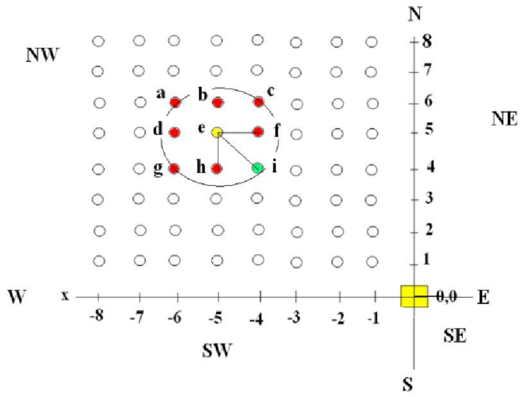


Fig.3 : Source node with other nodes and sink in co-ordinate system

n(e) is considered as source node and nodes a,b,c,d,f,g,h,i are considered as its neighbor nodes.The circle indicates the range of source node n(e).Base station is at position (0,0). Source node then needs to find next best optimal node for transmission of data. It identifies nodes n(f), n(h), n(i) as neighbor nodes for transmission of data.With ranking methodology source node needs to find optimal node for transmission.

Table 7 Nodes along with their parameter values

NODE/PARAMETER	α_{RE}	α_{ECT}	α_{POP}	α_{QL}
n(f)	80	45	1	7
n(h)	90	35	1	10
n(i)	70	40	1	5

Hence source node now finds the rank of these nodes where
n(f) = 2+3+1+2= 8
n(h)= 1+1+1+3= 6
n(i)= 3+2+1+1=7
since the node n(h) has the best ranking among all the 3 nodes its selected as the optimal node for transmission of data.

4. Conclusion

FSA-FTEER is a fault tolerant and energy efficient routing protocol. Based on finite state automata, the routing algorithm finds a node-disjoint multipath route where paths don't interfere with each other. Congestion controls is achieved with help of load balancing and choke packets. Faulty nodes and mobility of nodes in network are considered while routing the packets. Optimal Value of link is calculated by ranking method and it also takes residual energy of node in to account. Protocol is both fault tolerant and energy efficient, even when the base station is mobile. In our future work we try to implement a test bed of protocol to verify the simulation results in a real-time environment.

5. References

[1] Mihaela Cardei, Shuhui Yang, Jie Wu," algorithms for fault-tolerant topology in heterogeneous wireless sensor networks", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 19, NO. 4, APRIL 2008, pg no 545-557.

[2] Renato E. N. Moraes, Celso C. Ribeiro, and Christophe Duhamel," Optimal Solutions for Fault-Tolerant Topology Control in Wireless Ad Hoc Networks", IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 8, NO. 12, DECEMBER 2009,pg no 5970-5981.

[3] Hanan Shpungin, Michael Segal, "Low-Energy Fault-Tolerant Bounded-Hop Broadcast in Wireless Networks", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 17, NO. 2, APRIL 2009, pg no 582-590.

[4] Feng Wang, My T. Thai, Yingshu Li, Xiuzhen Cheng, Ding-Zhu Du, "Fault-Tolerant Topology Control for All-to-One and One-to-All Communication in Wireless Networks", IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 7, NO. 3, MARCH 2008,pg no 322-331.

[5] MohammadTaghi Hajiaghayi, Nicole Immorlica, Vahab S. Mirrokni , "Power Optimization in Fault-Tolerant Topology Control Algorithms for Wireless Multi-hop Networks", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 15, NO. 6, DECEMBER 2007, pg no 1345-1357.

[6] Amitava Datta," A Fault-Tolerant Protocol for Energy-Efficient Permutation Routing in Wireless Networks", IEEE TRANSACTIONS ON COMPUTERS, VOL. 54, NO. 11, NOVEMBER 2005 , pg no1409-1421.

[7] Binh Vien Dao, Jose Duato, Sudhakar Yalamanchili, " Dynamically Configurable Message Flow Control for Fault-Tolerant Routing", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 10, NO. 1, JANUARY 1999, pg no 7-22.

[8] Yu Wang, Hongyi Wu, Feng Lin, and Nian-Feng Tzeng, "Cross-Layer Protocol Design and Optimization for Delay/Fault-Tolerant Mobile Sensor Networks (DFT-MSN's)", IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 26, NO. 5, JUNE 2008, pg no 809-819.

[9] Winnie Louis Lee, Amitava Datta, Rachel Cardell-Oliver, "FlexiTP: A Flexible-Schedule-Based TDMA Protocol for Fault-Tolerant and Energy-Efficient Wireless Sensor Networks", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 19, NO. 6, JUNE 2008, pg no 851-864.

[10] Hong Luo, Jun Luo, Yonghe Liu, Sajal K. Das, "Adaptive Data Fusion for Energy Efficient Routing in Wireless Sensor Networks", IEEE TRANSACTIONS ON

COMPUTERS, VOL. 55, NO. 10, OCTOBER 2006,pg no 1286-1299.

[11] Charles Pandana , K. J. Ray Liu,” Robust Connectivity-Aware Energy-Efficient Routing for Wireless Sensor Networks”, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 7, NO. 10, OCTOBER 2008,pg no 3904-3916.

[12] Yan Jin, Ling Wang, Ju-Yeon Jo, Yoohwan Kim, Mei Yang, Yingtao Jiang,” EECCR: An Energy-Efficient m-Coverage and n-Connectivity Routing Algorithm Under Border Effects in Heterogeneous Sensor Networks”, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL. 58, NO. 3, MARCH 2009, pg no 1429-1442.

[13] Chen Wei, Chen Zhi, Pingyi Fan, Khaled Ben Letaief, ”AsOR: An Energy Efficient Multi-Hop Opportunistic Routing Protocol for Wireless Sensor Networks over Rayleigh Fading Channels”, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 8, NO. 5, MAY 2009,pg no 2452-2463.

[14] Harshavardhan Sabbineni, Krishnendu Chakrabarty,” Location-Aided Flooding: An Energy-Efficient Data Dissemination Protocol for Wireless Sensor Networks”, IEEE TRANSACTIONS ON COMPUTERS, VOL. 54, NO. 1, JANUARY 2005, pg no 36-47.

[15] Haibo Zhang, Hong Shen, ” Energy-Efficient Beaconless Geographic Routing in Wireless Sensor Networks”, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 21, NO. 6, JUNE 2010, pg no 881-896.

[16] Ming Yu, Kin K. Leung, Aniket Malvankar, “A Dynamic Clustering and Energy Efficient Routing Technique for Sensor Networks”, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 6, NO. 8, AUGUST 2007,pg no 3069-3079.

[17] Siva D. Muruganathan, Daniel C. F. MA, Rolly I. Bhasin, Abraham O. Fapojuwo, ” A Centralized Energy-Efficient Routing Protocol for Wireless Sensor Networks”, IEEE Radio Communications ,March 2005.

[18] F. Zabin, S. Misra, I. Woungang, H.F. Rashvand, N.-W. Ma, M. Ahsan Ali,” REEP: data-centric, energy-efficient and reliable routing protocol for wireless sensor networks”, IET Commun., 2008, Vol. 2, No. 8, pg no. 995–1008.

[19] Ka. Selvaradjou, N. Handigol, A.A. Franklin, C.S.R. Murthy,” Energy-efficient directional routing between partitioned actors in wireless sensor and actor networks”, I ET Commun., 2010, Vol. 4, Iss. 1, pg no. 102–115.

[20] Jun Luo, Jean-Pierre Hubaux,” Joint Sink Mobility and Routing to Maximize the Lifetime of Wireless Sensor Networks:The Case of Constrained Mobility”, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 18, NO. 3,JUNE 2010, pg no 871-884.

[21] Thrasyvoulos Spyropoulos, Konstantinos Psounis, Cauligi S. Raghavendra, “Efficient Routing in Intermittently Connected Mobile Networks: The Single-Copy Case”, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 16, NO. 1, FEBRUARY 2008, pg no 63-76.

[22] Peter Kok, Keong Loh, Hsu Wen Jing, Yi Pan,” Performance Evaluation of Efficient and Reliable Routing Protocols for Fixed-Power Sensor Networks”, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 8, NO. 5, MAY 2009, pg no 2328-2335.

[23] Michele Zorzi, Paolo Casari, Nicola Baldo, Albert F. Harri,” Energy-Efficient Routing Schemes for Underwater Acoustic Networks”, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 26, NO. 9, DECEMBER 2008, pg no 1754-1766.

[24] Tiansi Hu, Student Member, Yunsi Fei,” QELAR: A Machine-Learning-Based Adaptive Routing Protocol for Energy-Efficient and Lifetime-Extended Underwater Sensor Networks”, IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 9, NO. 6, JUNE 2010, pg no 796-809.

[25] Hojoong Kwon, Tae Hyun Kim, Sunghyun Choi, Byeong Gi Lee,” A Cross-Layer Strategy for Energy-Efficient Reliable Delivery in Wireless Sensor Networks”, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 5, NO. 12, DECEMBER 2006, pg no 3689-3700.

[26] Seung Jun Baek, Gustavo de Veciana,” Spatial Energy Balancing Through Proactive Multipath Routing in Wireless Multihop Networks”, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 15, NO. 1, FEBRUARY 2007, pg no 93-104.

[27] Yi Huang and Yingbo Hua, Fellow,” Energy Planning for Progressive Estimation in Multihop Sensor Networks”, IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 57, NO. 10, OCTOBER 2009, pg no 4052-4065.

SECURITY-ORIENTED ROBUST NETWORKING ARCHITECTURE AND KEY MANAGEMENT FOR HETEROGENEOUS WIRELESS SENSOR NETWORKS

McKenzie McNeal III
Tennessee State University
3500 John A. Merritt Blvd.
Nashville, TN 37209
mmcneal01@mytsu.tnstate.edu

Wei Chen
Tennessee State University
3500 John A. Merritt Blvd.
Nashville, TN 37209
wchen@tnstate.edu

Sachin Shetty
Tennessee State University
3500 John A. Merritt Blvd.
Nashville, TN 37209
sshetty@tnstate.edu

Stanley Aungst
Pennsylvania State University
101l Info Sci & Tech Bl
University Park, PA 16802
sga103@psu.edu

ABSTRACT

Current communication protocols used for Wireless Sensor Networks (WSNs) have been designed to be energy efficient, provide low redundancy in sensed data, and increase the lifetime of the sensor network. One major issue that must be addressed is the security of data communication. Due to the limited capabilities of sensor nodes, designing security-based communication protocols present a difficult challenge. Since current commonly used encryption schemes require large computation time and large memory, achieving security for WSNs needs unique approaches. We propose that the security should begin with the network architecture in defining the roles of each sensor node. Our work is to develop robust networking architecture and secure communication scheme. The defined roles of sensor nodes provide a measure of security through a hierarchical communication protocol and an efficient key management scheme that uses both public key and symmetric key cryptography. The simulation results show that by using the proposed networking architecture and key management scheme only a small amount of keys needs to be preloaded before deployment and stored after key setup to achieve secured communication throughout the entire network.

Key words: security, key management, sensor networks, heterogeneous

1. INTRODUCTION

A wireless sensor network (WSN) of low cost sensor nodes can be densely deployed and used for distributed data gathering, monitoring and surveillance in the applications of wildlife monitoring, military command, distributed robotics, industrial quality control, observation of critical infrastructures, smart buildings, intelligent communications, traffic monitoring, examining human heart rates, etc [1]. Some research has used WSNs for detecting the release of a poisonous gas [2]. But the information gathered by these applications is not properly protected. Security presents a major challenge in WSN design due to the limited resources and constraints of the sensor nodes. Any attacks that occur on the network could drain the limited resources. A security approach requires a certain amount of resources for implementation, including memory for storing security keys and code space, time complexity for encoding, decoding, and transmitting the encoded messages, and energy to power the sensor node. Sensor nodes do not have the capability to support traditional security methods. Due to the memory size of sensor nodes, it is necessary to limit the code size when designing cryptographic algorithms. Power is the biggest constraint in WSN. When implementing security algorithms and protocols, extra power is needed for processing security functions, transmitting encrypted data and overhead, and storing security parameters. Some applications may require a WSN to operate unattended or within a hostile environment. This situation exposes the network to physical attacks if the environment is open to adversaries, bad weather, etc. WSNs will also need to be managed remotely, which makes it virtually impossible to detect physical tampering and physical maintenance issues.

Many security methods have been proposed for homogeneous and heterogeneous WSNs. Heterogeneous WSNs (HWSNs) feature two or more node types that differ in resources and capabilities and are becoming more prevalent. A few nodes with more energy, stronger processing capability and longer communication range can be used to relax the communication bottleneck experienced in homogeneous WSNs and increase the lifetime of the network by taking on greater responsibilities than a typical resource constrained sensor node [3-5]. Some security methods developed for HWSNs show that heterogeneity helps to provide leverage in security by using high-end sensor nodes (h-nodes) to take on more security responsibilities than low-end sensor nodes (l-nodes) [6-9]. In this paper, we introduce a robust network architecture coupled with a secure communication scheme to provide security for HWSNs. Hierarchical network architecture is defined by regions and clusters. The h-nodes divide the deployment area into regions, where each l-node belongs to the region of the closest h-node. In each region, the l-nodes form clusters. Cluster heads and h-nodes form a backbone tree that can be used for data aggregation and relay. This architecture allows for self organization without the need for localization information. A combination of both symmetric and asymmetric keys is used to secure node to node communication. Elliptic curve cryptography (ECC) is feasible for sensor nodes, providing a 160-bit key which is securely equivalent to the RSA 1024-bit and thereby provide public key cryptography for WSNs [11-12]. For symmetric key cryptography, l-nodes use preloaded keying materials and neighbor knowledge to dynamically generate a pair-wise key [13-14].

The rest of this paper is organized as follows. Section II discusses related works on security for HWSNs. Section III presents the robust network architecture, the clustering algorithm and roles of h-nodes and l-nodes. In section IV, we will discuss the key management scheme and secure routing. Section V shows simulation results and comparative analysis. Finally, section VI concludes this paper and discusses future work.

2. RELATED WORK

Some research has been done to design security methods for HWSNs. In [6], a hybrid key management scheme called LIGER was proposed. LIGER uses a probabilistic unbalanced key distribution scheme where more keys are preloaded onto h-nodes than l-nodes. The scheme also has the ability to change from a standalone key-management system called LION to a key distribution center (KDC) based key management system called TIGER in case the sensor network is able to communicate with an existing backbone network.

Lu *et al.* [7] proposed two key distribution schemes, key-pool based and polynomial-pool based schemes, for HWSNs so that h-nodes and l-nodes can established at secure communication. In the key-pool based scheme, if two nodes share a key, they can establish secure communication. In the polynomial-pool based scheme, if two nodes exchange IDs, they can establish a secure link with a key only known by the two communicating nodes. The same polynomial generates a different key for different pair of nodes.

Du *et al.* [8-9] proposed a key distribution scheme that addresses the key exchange issue in homogeneous sensor networks by using both public-key and symmetric key cryptography to establish secure communication between h-nodes and l-nodes. In this scheme, they introduce the c-neighbor concept, where nodes only need to establish a key with communicating neighbors that are in route back to the sink. This helps to save resources of l-nodes so that they are not preloaded with an arbitrary number of keys before deployment. Their performance evaluation show a significant decrease in the amount of keys preloaded before deployment from that of Eschenauer and Gligor key management distribution scheme [15]. Furthermore, the Du-scheme shows better resilience against node compromise by assuming that h-nodes are tamper resistant and that any compromised l-node has minimal effect on the network because it only shares a key with communicating neighbors and not all of its surrounding neighbors.

All of the aforementioned key management system offer beneficial methods to establishing secure communication for HWSNs, but we argue that by establishing a robust network architecture that defines the different roles for h-nodes and l-nodes will offer a measure of security that helps to alleviate some of the challenges of key distribution and offer a foundation for secure communication. Those challenges include designing reliable and available network architecture, forming a high performance network infrastructure through self-organization, preloading and storing the least amount keys necessary to achieve secure communication between each node and leveraging security tasks to maximize network resources.

3. ROBUST NETWORK ARCHITECTURE

In a HWSN, we use two types of nodes, h-node and l-node, where the h-node has greater capabilities and more resources than the l-node. The network consists of a large amount of l-nodes with a small amount of h-nodes. We assume that the communication range of h-nodes is D and the h-nodes themselves is a connected wireless communication network in the sensor field. We define hierarchical cluster-based network architecture for the HWSN as follows:

- **Region Formation –** h-nodes divide the sensor field into regions using Voronoi Diagram, where each l-node belongs to the region of the closest h-node. In the region formation algorithm, the l-node selects the h-node which has the strongest signal to be its region head.
- **Clustering –** l-nodes are grouped into completed graphs in each region using a clustering algorithm that assigns one cluster head (*ch*) to each group with the remaining nodes of the cluster acting as cluster members (*cm*). If an l-node *u* has chosen a *ch v*, but u is selected to be the *ch* of another node(s), *u* will send a message to remove itself from *v's* cluster and become the *ch* of a new cluster.. To assure the head-rotation in the reconfiguration functions, the diameter of the cluster is d/2.

- **Backbone Tree Formation –** *Ch*s and h-nodes form a communication backbone that is used to send information in bi-direction between the base station to regions, between regions to clusters, and between cluster heads to cluster members through the entire network.

The network architecture is self-organized without the need of localization information of any nodes and reconfigurable in the presence of node failures, resource depletion, the addition and subtraction of nodes, and node compromise. Figure 1 shows an example of the hierarchical cluster-based HWSN architecture. The algorithms have been developed to configure and reconfigure the hierarchical network architecture and support the communication protocol as follows:

- *Cluster members (cms)* communicate only and directly with their *ch*.
- *Cluster heads (chs)* communicate with *cm*s in their cluster and with neighboring *ch*s on the backbone tree.
- Information on the backbone tree travels from child to parent (*ch* to *ch*) until it reaches the h-node of their region head.
- H-nodes send information to neighboring h-nodes in route to the base station.

We assume all collisions are avoided by using MAC protocol CSMA/CA. The network self-organization/reconfiguration is supported by the following functions:

- Node Move-out – *Ch* or *cm* leaves the network and connectivity is maintained.
- Node Move-in – New l-node joins the network and becomes a *ch* or *cm* depending on its surrounding neighbors.
- Head Rotation – New *ch* is selected when an existing *ch* is low on resources or even compromised.

Due to limitation of the number pages, details of the clustering algorithm were not included.

4. KEY MANAGEMENT SYSTEM

A. Key Cryptography

We propose a key management system supported by public key and symmetric key cryptography. Both cryptographic methods have their strengths and weaknesses, but when they are used together the weaknesses will be overcome and the security will be provided. We also propose the use of two types of keys for tasks such as data aggregation, which may occur at intermediate nodes during data transmission. Public key cryptography was initially classified as infeasible for WSNs. Recent studies have shown that ECC is feasible for existing sensor node hardware and therefore feasible for WSNs [12]. Our key management system couples ECC with the polynomial-based key distribution scheme. The polynomial-based scheme allows two nodes to generate a pair wise key using a randomly generated symmetric bi-variate *t*-degree polynomial *f(x,y)*, where $f(x,y) = \sum_{i,j=0}^{t} a_{ij} x^i y^j$ over a finite field F_q. Each sensor needs to store a *t*-degree polynomial which occupies $(t+1)\log q$ storage space. To establish a pair wise key, both sensor nodes need to evaluate the polynomial at the ID of the other sensor node. The polynomial-based scheme is secured up to a degree of *t*, where *t* is the number of nodes that needs to be compromised in order for an adversary to know the symmetric key generated between any two nodes [13, 14].

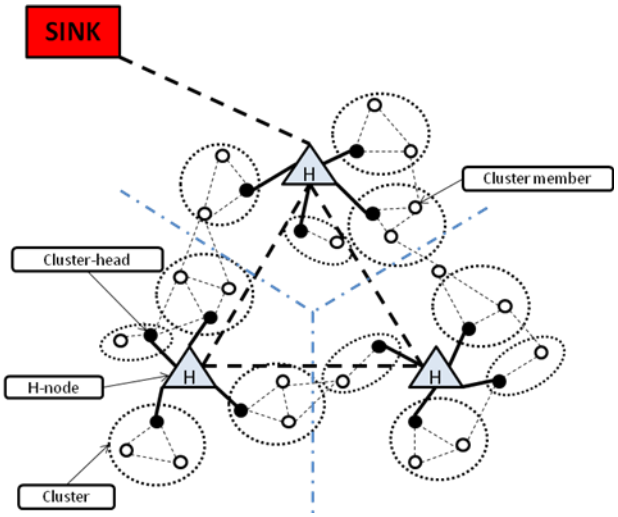


Figure 1 - Hierarchical cluster-based HWSN architecture

B. Preloaded Keys and Materials

Before the nodes are deployed, both h-nodes and l-nodes are preloaded with an initial temporary symmetric key K_G. Each h-node is preloaded with its ECC public/private key pair. Therefore, each H-node has a total of 3 keys. We propose two cases for preloading l-nodes. Preloading for h-nodes remains the same in both cases. In case 1 each l-node is preloaded with its private key from its ECC public/private key pair for a total of 2 keys. In case 2 each l-node is preloaded with its ECC public/private key pair for a total of 3 keys. Let *M* represent the number of h-nodes and *N* represent the number of l-nodes, then the number of preloaded keys for the entire network is:

Case 1: $3 \times M + 2 \times N$

Case 2: $3 \times M + 3 \times N$

The l-nodes are also pre-loaded with a randomly generated symmetric bi- polynomial that will used to enerate a symmetric key with a neighboring l-node such as the *cm* to *ch* communication.

C. Stored Keys

After the nodes are deployed, they perform neighbor discovery and clustering processes. The key K_G is used during neighbor discovery and clustering so that information is broadcasted securely. Even though we present 2 cases for preloading keys, key storage is the same for both cases. Once h-nodes have divided the sensor field into regions, l-nodes have formed clusters, and the backbone tree is formed, key exchange can start by using K_G and the preloaded symmetric bi-variate polynomial. Key exchange will be discussed in detail in the next section. The number of keys stored depends on the number of clusters throughout the entire network. The following variables help define the amount of keys stored:

- N_h – *number of l-nodes in the region of a h-node*
- K_h – *number of the neighbors of a h-node on backbone tree*
- C_{ch} – *number of the cluster members in a cluster with cluster head ch*
- K_{ch} – *number of the neighbors of a cluster head on the backbone tree*
- N_c – *number of the clusters in the network*

After key exchange each node stores a certain number of keys depending on their role in the network hierarchy:

- H-node – stores ECC public/private key pair, public keys of all l-nodes in its region and the public keys of all its neighboring h-nodes in the backbone tree. The keys stored for neighboring h-nodes are public keys of parent and children on backbone tree. Let A_h represent the number of total keys stored at one h-node, then:
$$A_h = 2 + K_h + N_h$$
- Cluster Head – stores private key of ECC pair, public key of the regional head, distinct symmetric keys with all *cms* generated by symmetric polynomial, and distinct key with parent and children on backbone tree. Let B_h represent the number keys stored by a *ch*, then:
$$B_{ch} = 2 + K_{ch} + C_{ch}$$
- Cluster Member – stores private key of ECC pair, the public key of their region head, and a distinct symmetric key with *ch* for a total of 3 keys.
$$C_{cm} = 3$$

If we sum the equations for *M* h-nodes and *N* l-nodes, $(M + N_c) - 1$ is the number of edges on the backbone tree. Let K_{all} represent total number of keys stored, therefore the estimated total number of keys stored in the entire network is:

$$K_{all} \le \sum_{all\ h} A_h + \sum_{all\ ch} B_{ch} + \sum_{all\ cm} C_{cm}$$

$$K_{all} \le 2M + 2[(M + N_c) - 1] + 2N_c + N + N$$
$$+3(N - N_c)$$

$$K_{all} \le 2M + 2M + 2N_c - 2 + 2N_c + N + N + 3N - 3N_c$$

$$K_{all} \le 4M + N_c + 5N - 2$$

$$K_{all} \le 4M + N_c + 5N$$

From this equation we find that the estimated number of keys is dependent on the number of clusters, N_c, in the network architecture.

D. Key Distribution and Set-up

As mentioned in the previous section, once the nodes are deployed, they begin to build the hierarchical networking architecture. Since nodes do not know their location, signal strength can be used to determine the proximity of a neighboring node. The following key and message notations are used to discuss this section:

- K_G – a temporary preloaded symmetric known by all nodes that is discarded and no longer needed once the network architecture established and key exchange is complete.
- x_{pb}/x_{pr} – public and private key of node *x*.
- K_{uv} – symmetric key shared between node *u* and *v*, where $K_{uv} = K_{vu}$.
- Broadcast from node x – *{x.id, Encryption_key(x.id, x.message)}*.
- Transmission from node x to node y – *{(x.id, y.id), Encryption_key(x.id, y.id, x.message)}*

The IDs are sent in plaintext to allow each node to know who is sending the message and whether it is meant for them to decrypt or forward according the communication protocol. Other materials may be sent with the message if requested by the sender or for authentication purposes. These keys are used in the different phases of building the network architecture as follows.

- **Regional Formation** – h-nodes (h) use K_G to broadcast their IDs to find neighbors and for l-nodes to select the nearest h-node as its regional head. Any h-node (*w*) and l-node that receives that message can decrypt it using K_G. The ID of the sending node would be added as a neighbor for h-nodes and as a regional head for l-nodes. Receiving h-nodes also reply back with an acknowledgement.

  - Broadcast – *{h.id, K_G(h.id, h.message)}*
  - Reply – *{(w.id,h.id),K_G(w.id, h.id, w.message)}*

- **Neighbor Discovery** – l-nodes (*u*) use K_G to broadcast their ID and find neighboring l-nodes. Any l-node (*v*) that receives this message adds the ID of the sending l-node to its neighbor list and replies back with an acknowledgement. The ID of the regional node is included with the message so that nodes only add neighbors from the same region.
  - Broadcast – *{u.id, K_G(u.id, u.message/h.id)}*
  - Reply – *{(v.id,u.id),K_G(v.id, u.id, v.message/h.id)}*
- **Clustering** – l-nodes form clusters using K_G and the neighbor list built during neighbor discovery phase. L-nodes exchange messages encrypted with K_G to determine whether they will a *ch* or *cm*. The ID of the regional head is included with message to prevent nodes from joining a cluster outside their region. After clustering all l-nodes designated *cm*s discard K_G for it is no longer needed. *Cm*s will use the preloaded polynomial to establish a pairwise key, K_{uv} = K_{vu}, with their *ch*.
  - Request *ch – {(u.id,v.id),K_G(u.id, v.id, u.message/h.id)}*
  - Reply – *{(v.id,u.id),K_G(v.id, u.id, v.message/h.id)*
- **Backbone Tree Formation** – Within each region, starting from the h-node as the root, a message is broadcast to find children (i.e., the *ch*s within *d* communication range of the h-node). All *ch*s are flagged as waiting before receiving a message. This prevents a node that is already a parent from becoming a child of another node. If an h-node receives a reply back from a *ch*(s), then that node is added to its child list. Once a *ch* has found a parent, it will send out a message to find its children. This will continue until all *ch*s have found a parent in the regional back bone tree. The parent and child node in these message are denoted by *p* and *c* respectively.
  - Child node broadcast request – *{p.id, K_G(p.id, p.message)}*
  - Reply to parent node – *{(c.id,p.id),K_G(c.id, p.id, v.message/h.id)*

Once the backbone tree is complete, each h-node needs to obtain the public key of every l-node in its region for future secure communications. We propose two ways of achieving this task. In section A, there were two cases proposed for preloading keys on l-nodes.

- Case 1 – h-nodes broadcast *cm* list request to each *ch* in its region using K_G. *Ch*s send a list of *cm*s to its regional head using $K_{G.}$ After the h-nodes receive this information from all *ch*s via the regional backbone tree, K_G is removed from all remaining nodes and is no longer needed for future communications. H-nodes send a message with a list of l-nodes in its region to the base station to request the public key of each l-node.
  - *Cm* list request - *{h.id, K_G(h.id, h.message)}*
  - *Cm* list reply - *{(u.id, h.id), K_G(u.id, h.id, u.message/u.cmlist/h.id)}*
  - Key request by h-node – *(h.id,B.id),h_{pri}(h.id, h.message/h.regionlist)}*
- Case 2 – h-nodes broadcast *cm* list and key request from each *ch* in its region using K_G. The *ch*s request the public key of every *cm* using the pairwise key it shares with each *cm*. It then sends a *cm* list with keys to its regional head via the backbone tree using K_G. After an h-node receives messages from all nodes, K_G is removed from all remaining nodes. Each l-node will also discard its public key as a security precaution.
  - *Cm* list and key request – *{h.id, K_G(h.id, h.message)}*
  - *Cm* list and key reply – *{(u.id, h.id), K_G(u.id, h.id, u.message/u.cmlist/u.keylist/h.id)}*

Even though the key storage is the same for both cases, case 1 requires h-nodes to communicate with a KDC during setup and each time a new node is added to the network. Case 2 stores more keys, but no communication with a KDC is needed allowing the network to self-organize during key setup and when a new nod joins the network.

For any new node joining the network, they are preloaded with the keys according to the cases mention in Section A. The number of keys stored by a new node depends on its role once it joins the network. If any l-node has to leave the network due to depleted resources, node failure or node compromise, all keys shared with that node is removed from its communicating neighbors. If that node is a *ch* and has remaining *cm*s, then head rotation is performed to select a new *ch* among the remaining *cm*s according to Section III. New keys are established as previously shown in each phase and according to preloaded keys for case 1 or case 2.

E. Secure Routing

Figure 2 shows a routing hierarchy and what type of key is used in communication between each type of node. Secure communication occurs from child to parent as follows:

- h-node to h-node – *h* is child, *w* is parent. Information is encrypted with an h-nodes private key and decrypted by the parent H-node with the public key.
  - *{(h.id, w.id),h_{pr}(h.id, w.id, h.messege)}*
- *ch* to h-node – *u* is child, *h* is parent. Information is encrypted with a ch's private key and decrypted by H-node with ch's public key.
  - *{(u.id, h.id),u_{pr}(u.id, h.id, u.messege/u.request)}*
- *cm* to *ch* – *u* is child, *v* is parent. Information is encrypted and decrpyted with pairwise key between cm and ch.
  - *{(u.id, h.id),K_{uv}(u.id, h.id, u.messege/u.request)}*

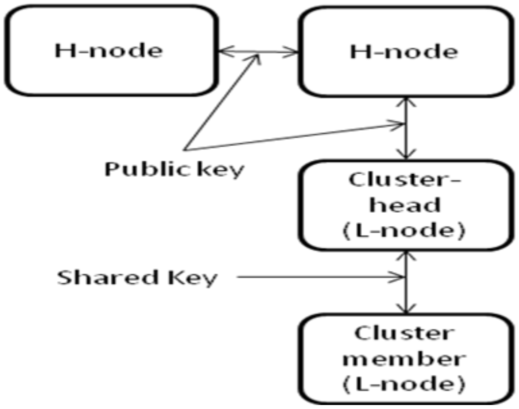


Figure 2 - Secure routing hierarchy for cluster-based HWSN.

5. PERFORMANCE EVALUATION

The simulation for this security model was developed using C-Sharp. The testing scenario features 1000 l-Nodes and 20 h-Nodes. The hierarchical network architecture was constructed with a communication range of d = 60m for l-nodes and D = 250 m for h-nodes. When forming clusters, the transmission range of l-nodes are adjusted to a range *d/4* to form the clusters, with diameter of d/2, where each cluster is a complete graph, i.e., each node in the cluster is communication range with every other node. The nodes were randomly deployed over a 500m × 500m area. These parameters where chosen to achieve connectivity over the deployment area so that each node would be in communication range with one or more nodes. Substituting the simulation parameters in the equations for the number of preloaded keys in case 1 and case 2, yields 2060 and 3060 keys respectively. Comparing this value to the centralized Du-scheme [8], we reduce the amount of preloaded keys by approximately 90% for case 1 and 86% for case 2. We also can estimate the number of stored keys. As we showed in Section IV.C, this number, $K_{all} \leq 4M + N_c + 5N - 2$, which depends on the number of cluster N_c. Both cases store the same number of keys depending on the number clusters throughout the entire network. We conducted 20 simulation runs where 1020 nodes were randomly deployed. The simulation results show that the network stored on an average; $K_{all} \leq 4M + N_c + 5N \approx 5500$ keys.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a security system for heterogeneous wireless sensor networks that couples robust network architecture with a hybrid key management scheme. The robust network architecture features a hierarchical cluster-based network architecture that defines the role of h-nodes and l-Nodes to establish a measure of security through the communication protocol. This network architecture has a direct effect on the key management scheme which uses both ECC and symmetric bi-variate polynomial-based key distribution to provide secure communication via the backbone tree. The network architecture can be self-reconfigured without localization information and it provides an efficient key management scheme for heterogeneous wireless sensor networks. With the storage being a limitation for sensor nodes, only a small amount keys need to preloaded and stored over the entire network.

Future work includes analyzing the key management scheme to determine key setup time and energy usage during key setup for both case 1 and case 2 in preloaded key scenarios. We plan to analyze the resilience of the network against node compromise and energy usage during secure routing.

ACKNOWLEDGEMENTS

This research is supported by the Defense Threat Reduction Agency (DTRA)

REFERENCES

1. Kaplantzis, S., Mani, N., Palaniswanmi, M., and Egan, G., "Security models for wireless sensor networks," *Conversion Report, Monash University*, 20th, March 2006.
2. Long, K.J., Haupt, S., Young, G.S., Rodriguez, L.M., McNeal III, M., "Source term estimation using genetic algorithm and scipuff," *7th Conference on Artificial Intelligence and its Applications to the Environmental Sciences*, January 2009.
3. Mhatre, V. and Rosenberg, C., "Homogeneous vs heterogeneous clustered sensor networks: A comparative study," *2004 International*

Conference on Communications, Vol. 6: pages 3646-3651, June 2004.

4. Al-Fares, M.S., Sun, Z., and Cruickshank, H., "A hierarchical routing protocol for survivability in wireless sensor network (WSN)," *Proceedings of the International MultiConference of Engineers and Computer Scientists IEEE 2009*, Vol I. March 18-20, 2009.
5. Khan, Z.H., Catalot, D.G., and Thiriet, J.M., "Hierarchical wireless network architecture for distributed applications," *2009 Fifth International Conference on Wireless and Mobile Communications 2009 IEEE*, 2009.
6. Traynor, P., Kumar, R., Choi, H., Cao, G., Zhu, S., and La Porta, T., "Efficient hybrid security mechanisms for heterogeneous sensor networks," *IEEE Transactions on mobile computing*, Vol. 6, June 2007.
7. Lu, K., Yi Qian, Guizani, M., and Chen, H., "A framework for a distributed key management scheme in heterogeneous wireless sensor networks," In *IEEE Transactions on Wireless Communications*, Vol. 7, February 2008.
8. Du, X., Xiao, Y., Ci, S., Guizani, M., and Chen, H., "A routing-driven key management scheme for heterogeneous sensor networks," In *The ICC 2007 Proceedings*, pp. 3407-3412. IEEE Communications Society, 2007.
9. Du, X., Guizani, M., Xiao, Y., and Chen, H., "A routing-driven elliptic curve cryptography based key management scheme for heterogeneous sensor networks," *IEEE Transactions on Wireless Communications*, Vol. 8, No. 3, March 2009.
10. He, T., Huang, C., Blum, B., Stankovic, J., and Abdelzaher, T., "Range-free localization schemes for large scale sensor networks," *MobiCom '03 Proceedings of the 9th Annual International Conference of Mobile computing and networking*, 2003.
11. Gura, N., Patel, A., Wander, A., Eberle, H., and Shantz, S.C., "Comparing elliptic curve cryptography and RSA on 8-bit CPUs," *Proceedings of the 6th International Workshop on Cryptographic Hardware and Embedded Systems*, Boston, MA, August 2004.
12. Liu, A. and Ning, P., "TinyECC: A configurable library for elliptic curve cryptography in wireless sensor networks," *7th International Conference on Information Processing in Sensor Networks (IPSN 2008),* April 2008.
13. Schmidt, S., Krahn, H., Fischer, S., and Wätjen, D., "A security architecture for mobile wireless sensor networks," *ESAS 2004,* Springer-Verlag Berlin Heidelbierg, 2005.
14. Liu, D., Ning, P., and Li, R., "Establishing pair-wise keys in distributed sensor networks," *ACM Transaction Information Systems Security,* 8(1):41-77, 2005.
15. Eschenauer, L. and Gligor, V.D., "A key management scheme for distributed sensor networks," *Proceedings of the 9th ACM CCS,* November 2002.
16. Chen, W., Miao, H., and Hong, L., "Cross-layer Design for Cooperative Wireless Sensor Networks with Multiple Optimizations," *International Journal of Networking and Computing*, (publication date).
17. Uchida, J., Muzahidul Islam, A.K.M., Katayama, Y., Chen, W., and Koichi, W., "Construction and maintenance of a novel cluster-based architecture for ad hoc sensor networks," *Ad Hoc & Sensor Wireless Networks* Vol. 00, pp. 1-31, 2008.
18. Walters, J.P., Liang, Z., Shi, W., and Chaudhary, V., "Wireless sensor network security: A survey," In Security in Distributed, Grid, and Pervasive Computing Chapter 17, Auerbach Publications, CRC Press 2006.

On Power Control Methods for Wireless Sensor Networks

Jihad Qaddour
School of Information Technology
Illinois State University
Normal, IL 61761

Keywords: Power control, Wireless Sensor Networks

Abstract

In this paper, we examine different methods of power control for wireless sensor networks. These methods fall into four broad categories: duty cycling, batching, hierarchy, and redundancy reduction.

Our review is grouped according to four phases of wireless sensor network operation: sensing, communication, computation, and storage. Under sensing, we discuss duty cycling, hierarchy, and soft deployment, which is a form of power load balancing for surveillance tasks. For communication, we address radio management in the form of polled, scheduled, and triggered operations, as well as communication batching, aggregation, and middleware improvements. Computation is discussed in terms of power-efficient hardware and TinyOS, which has built-in power management features. Finally, storage is addressed in terms of power-efficient utilization and storage vs. transmission.

II. Overview

Sensor Networks are ad hoc meshes consisting of sensors with simple network and computing capabilities [1]. The typical sensor is a small, low-energy unit consisting of a sensor, a microcontroller, a storage device, and a radio, powered by a pair of AA batteries [2]. These units, called "motes" are typically deployed in groups of tens or hundreds [1].

Current application areas for sensor networks center on monitoring and data collection for military, agricultural, medical, environmental, and research firms [3]. Sensor networks are still an emerging technology, and due to ongoing improvement, current equipment is likely to be obsolete within three years [1]. The market for sensor networks is immature and fragmented, with vendor-specific equipment and protocols being common [1]. The main customers are early adopters who view them as a tactical rather than strategic investment [1]. However, sensor networks are expected to become commonplace and have a large business impact in the future [1, 10].

III. Power Control

The typical sensor mote is powered by two AA alkaline batteries supplying 3V at 2000mAhr [4]. Solar-powered motes have been suggested but are not yet widespread [5]. Because motes are often deployed to remote locations and/or spread over a large geographic area, refreshing the power supply is often impractical [2]. Therefore, power control techniques are necessary to extend the life of the sensor network.

Power control techniques can be applied to four phases of sensor network operation: sensing, communication, computation, and storage [4, 10], and the remainder of this paper is organized by each operation. Typical power-saving techniques include duty cycling, batching, hierarchy, and redundancy reduction [4, 8, 9].

- **Duty Cycling:** Powering a subsystem down at predefined times in order to reduce its average power draw.
- **Batching:** Storing multiple operations to be executed in burst in order to minimize startup and overhead costs.
- **Hierarchy:** Using low power systems first, only triggering higher power systems when there is a need. For example, a low powered sensor can trigger a higher powered sensor if there is movement.

- **Redundancy reduction:** Making use of compression, aggregation, or message suppression techniques [4].

A. Sensing Phase

Sensing accounts for much of a mote's power use. In fact, the power demands of sensors are enough that inefficient power control during this phase can easily offset any gains from power control in other areas [4]. Sensing applications can be initiated according to three models: continuous, user-initiated, and event-driven [6]. With the continuous model, an area can be fully monitored at all times, but the power requirements make this model unsustainable for the required lifetime of a battery-operated sensor network. The user-initiated model, on the other hand, requires minimal power but is unsuitable for situations in which the phenomena to be monitored are unpredictable. Therefore, current research focuses on the event-driven sensing model in which an area is to be monitored for specific types of randomly occurring events, such as movement [6].

In the event-driven model, the network operates in two states: the surveillance state and the tracking state [6]. During the surveillance state, nodes are powered down, maintaining low-power sensing necessary to detect unusual or interesting events; at which point the node switches to the more power-intensive tracking state. Quality of surveillance is inversely proportional to power used [6].

The primary methods of power control during the sensing phase are duty cycling and hierarchy [4]. Gui and Mohapatra [6] have also proposed a method for soft deployment of sensors that reduces the cost of maintaining the surveillance state and spreads this burden across the network.

1. Duty Cycling

Duty cycling during sensing involves using a sleep-wake up-sample-compute-communicate cycle in which motes spend a majority of the time sleeping [4]. For periodic data collection, the duty cycle period (sum of the amount of time in one sleep period plus the amount of time in one wake period) should be set to the desired sampling interval. The only constraint is that the wake period must include enough time for the sensor to wake up from its sleep state and initialize in addition to the amount of time required to take a sample [4].

When the event that needs to be monitored is random, duty cycling should take detection time into account. Detection involves taking a number of quick samples to determine if the event is occurring [4]. The time that the sensor is sleeping must be short enough that the sensor can power on, detect the event, and record the event before the event is over. This is given by the following formula [4]:

$T_{off} \leq T_{event} - 2 * T_{on}$

As an example, if a randomly occurring phenomenon has a window of 10 ms in which it can be recorded, and the total time it takes for a sensor to power up, detect the event, and record it is 2 ms, the sensor's sleep period cannot be set longer than 6 ms or the sensor could end up missing the event. This presents a problem: if T_{event} is too short, the sensor cannot be powered down long enough to lower average power consumption by a significant amount through duty cycling alone [4].

2. Hierarchy

Hierarchy is another technique for increasing power efficiency during the sensing phase. The principle behind sensor hierarchy is that certain types of sensors use more power than others. For example, a magnometer draws 15 mW, while an accelerometer draws only 1.8 mW [4]. If the application is a conjunction of magnometer and accelerometer readings requiring a certain threshold for both, then the relatively low power cost of the accelerometer means it should be activated first. A software decision can then be made as to whether or not it is necessary to activate the magnometer.

3. Soft Deployment

In [6], Gui and Mohapatra propose a soft deployment scheme for target tracking applications. For tracking applications in the surveillance state, the entire field does not need to be monitored at once. It is often acceptable to introduce holes in the surveillance coverage, as long as there are enough sensors performing detection in a pattern sufficient to ensure that a moving target will be detected within an acceptable amount of time or distance travelled. Once a target is detected, nearby nodes can be alerted to wake up and begin tracking.

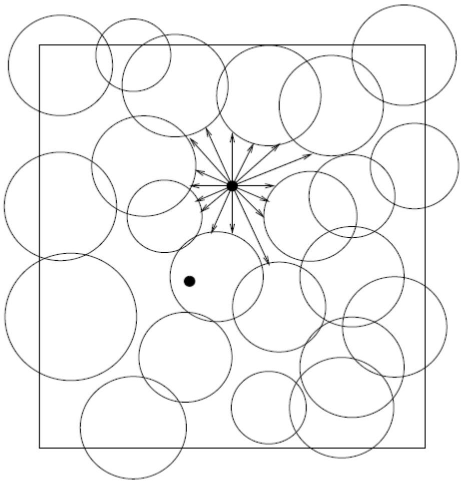


Fig. 1 Sensor network coverage with gaps in surveillance state, showing how far a target can move without being detected [6].

Soft deployment [6] is a two-stage process in which the network is first deployed physically, and then the configuration of the surveillance grid is handled in a flexible manner through software. The software aspect is advantageous because in most sensor applications, it is difficult to change the physical layout of the network once the nodes are deployed. Physical deployment is governed by the needs of the tracking application, which usually requires each point on the grid to be covered by multiple sensors. However, most of the time, there is no target to track. There is therefore an overabundance of nodes available to perform surveillance duties. Most of these nodes sleep, while software determines which combination of nodes remain awake in a surveillance state to achieve a desired level of coverage which can be changed as the need arises.

a) Soft Deployment Sleep Planning

Several methods are available for sleep planning, and these fall into two categories: uniform distribution methods and planned distribution methods. Uniform distribution methods attempt to maintain acceptable coverage without reference to node locations. The two methods for doing this are *pre-scheduled independent sleeping* and *neighbor collaboration sleeping*. Planned distribution methods, by contrast, take node location into account in order to form specific geometric patterns of coverage.

(1) Pre-Scheduled Independent Sleeping

Pre-scheduled independent sleeping [6] is a uniform distribution method in which nodes do not have any knowledge of neighboring nodes' sleep status, and all nodes set their power control status independently. The protocol used for this is called Random Independent Sleeping (RIS). The nodes are programmed with a specific length duty-cycle, T_{slot}, which is divided into an active period ($p*T_{slot}$) and a sleep period (T_{slot}-$p*T_{slot}$). Upon activation of the scheme, all nodes wait a randomly distributed amount of time before starting their duty cycles in order to ensure random coverage. Intensity of monitoring is can be altered by changing the values of T_{slot} and p to increase or decrease the probability that a given node is active at a given time.

(2) Neighborhood Cooperative Sleeping

Neighborhood cooperative sleeping [6] is the other uniform distribution method for sleep planning. It is different from pre-scheduled independent sleeping in that it takes neighbor nodes' sleep status into account. The protocol used to implement this method is Probing Environment Adaptive Sleeping (PEAS). Under PEAS, each node is given a probing rate λ and a sleep time duration T_s, which it uses to determine how long to sleep and when to send probing messages to other nodes. Upon sending a PROBE message, a node listens for any active nearby nodes which will send a REPLY message. If a REPLY is received, the node goes back to sleep, but if no REPLY is received, the node assumes no other nearby nodes are active and begins surveillance. The desired intensity of surveillance coverage can be achieved by changing the power level of PROBE messages to create a larger or smaller radius.

The drawback to PEAS is that there is no mechanism for balancing the load of the surveillance task across the network. Each active node remains in surveillance node until it runs out of power and fails. This results in prematurely dead nodes that decrease the sensor network's density, reducing coverage over the long term and possibly creating routing problems. Gui and Mohapatra [6] propose an extension to PEAS, called PECAS, that would include a *Work_Time_Dur* value at each monitoring node to allow it to calculate when it should go back to sleep. The value of *Next_Sleep_Time* would be incorporated into REPLY messages so idle nodes would know when the active node is scheduled to stop performing surveillance. At that point, any nearby sleeping nodes should wake up and probe again. A random offset would be added at each node to reduce the likelihood that multiple nodes in the same area will wake up, probe, and become active at the same time.

(3) Planned distribution

Planned distribution [6] requires each node to know its position as a set of coordinates, so that the sensor net forms a virtual mesh or grid. For each virtual horizontal line *i* and vertical line *j*, where the distance between lines is l_G and δ is a distance tolerance to allow for random placement of sensors, sensors which have x coordinates in the range of $[i * l_G \pm \delta]$ or y coordinates in the range of $[j * l_G \pm \delta]$ are tasked with surveillance. This creates a grid where each uncovered area is a square. The burden of surveillance can then be spread across the entire network through the use of time slots to create a moving mesh, although this introduces the additional requirement of time synchronization. At the end of every time slot, each line is offset by 2δ, wrapping around as needed.

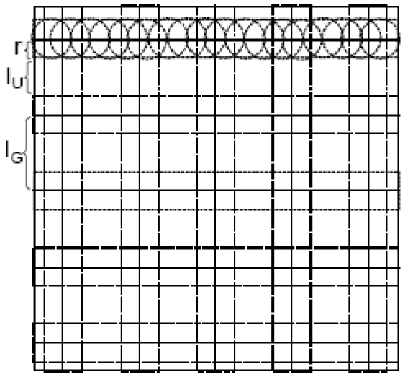


Fig. 2 Planned distribution grid with distance between lines l_G, uncovered area l_U, and sensor radius r [6].

b) Tracking State Wakeup

Sleep planning methods result in a few nodes performing surveillance duties for the entire network. It is therefore necessary for these nodes to be able to alert surrounding nodes when it is time to enter the tracking state [6].

Gui and Mohapatra [6] propose a scheme for proactive wake-up of nodes in which there are four levels of readiness: Tracking, SubTrack, Prepare, and Waiting. In this scheme, the ordinary traffic exchanges among tracking nodes are marked as tracking packets. Nodes that receive tracking packets are placed into the SubTrack state, where they attempt to acquire the target. If a node acquires a target, that node is placed in the Tracking state and begins actively collaborating with other nodes. For this system, the Tracking state radius is the sensor range *r*, and the SubTrack radius is the transmission range *R*. To further increase the range, nodes that receive tracking packets can transmit PREPARE packets. If this is done, the radius for PREPARE mode is r+2R, compared to r+R for SubTrack mode.

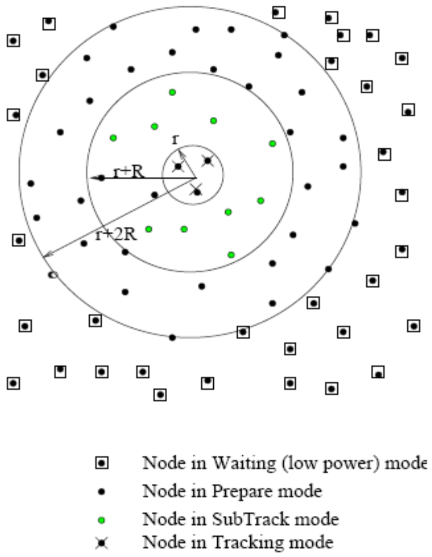


Fig. 3 Readiness states in tracking mode as a function of sensing range r and transmission range R [6].

B. Communication Phase

Radio units used by sensor networks consume energy at a rate that is several orders of magnitude larger than the processing and storage equipment of a typical mote [3]. Power management in the communication phase involves techniques to minimize this cost, incorporating radio management, batching, aggregation, and middleware improvements.

1. Radio Management

Radio management refers to MAC-level power management techniques that keep the radio transceiver powered down as much as possible. Radios in sensor networks draw similar amounts of power whether they are actively transmitting or simply listening to the channel, [3] so techniques to minimize unnecessary radio monitoring had to be developed. These include polled operation, scheduled operation, and triggered operation.

a) Polled Operation

With polled operation [4], the radio is turned off a significant percentage of the time, awakening frequently to sample for traffic. Most sensors that employ this method have a duty cycle of between 1-2% [4], meaning the radio is powered down 98% of the time.

The constraining factors for polled operation are the radio startup and sample time. For polled operation to be effective, the radio must be able to detect the preamble of a transmission, which is typically several bytes in length [4]. The longer the radio startup and sample times are, the longer the preamble needs to be. Longer preambles reduce channel capacity and cost more energy to transmit. However, in sensor networks that do not need send data very frequently, it can be more efficient overall to set a longer radio sampling period through software, with a corresponding increase in the length of the preamble, since this overhead cost will not be borne often [4].

b) Scheduled Operation

Scheduled Operation [4] involves scheduling ahead of time when transmissions will occur, which saves power because the channel does not need to be monitored the rest of the time. The two implementations of scheduled operation are S-MAC and 802.15.4. S-MAC works by coordinating sleep periods during which transmission is illegal and power does not need to be consumed [2]. 802.15.4, on the other hand, is a form of TDMA [4]. TinyDB, which is commonly used to query sensor networks, is aware of scheduled operation protocols and is able to save power by turning off the network stack when it is not needed[4]. Scheduled operation incurs a small overhead cost because of the complexity of scheduling and synchronization.

c) Triggered Operation

Triggered operation involves using a low-power secondary radio to monitor the channel and trigger the main radio when needed, allowing for purely asynchronous communications without the need for polled or scheduled operation [3]. This type of power management is mostly a matter for academic research because of the difficulty in developing a suitable low power or zero power secondary radio [4].

Triggered operation requires the use of a wake-up call containing a code or address for the target [3]. This method requires the secondary wake-up radio of each mote to be operated continuously, which means it is continuously drawing power. In order for this method to result in any net energy savings, the power draw of the wake up radio must be less than 50 µW [3]. No such radio currently exists, however, Le-Huy and Roy [3] outline a design for a 20 µW radio that combines a zero-bias Schottky diode envelope detector with an address decoder to read an 8-bit wake-up address [3].

2. Communication Batching

Since low powered listening techniques such as Polled Operation require a lengthy preamble for each transmission, it is useful to send as many packets as possible at once in a packet train [4]. This batching reduces overhead costs at the cost of transmission delay.

3. Aggregation

Many sensor net applications require queries that return aggregate data [4]. In networks with a hop count > 1, it is helpful if each node can perform the aggregation function before passing the data on to the next hop. On-board computation is usually far more energy-efficient than communication, so aggregation represents in a net benefit for power consumption by reducing the number of messages for a relatively small cost in power used for computation. An ordinary sensor network with n number of hops and no aggregation requires $n^2/2$ messages to transmit sensor data in response to a query, but aggregation at each hop reduces this to n messages [4]. In the case of Count, Min, Average, and Histogram queries, aggregation will significantly lower the amount of data that needs to be transmitted; but aggregation does not result in significant savings for Median or Count Distinct functions[4].

4. Middleware Improvements

Time synchronization is one area where middleware services can save energy. Time synchronization is very important to sensor net applications, but it costs significant amounts of energy to send and receive periodic time synchronization messages every few seconds. Middleware can eliminate this need by combining time-stamped packets with an API layer to perform time synchronization and make the necessary database adjustments after an event has already occurred, saving energy [4].

Routing is another area where middleware can help save energy. Routing protocols should take the remaining energy of nodes into account, and route through nodes that have a higher remaining power level if possible [4]. This helps balance the energy costs of transmission across the entire network. Routing protocols should also prefer paths that require the least number of transmissions.

The Trickle algorithm [7] is used in sensor networks running TinyOS to maintain code updates across the network. In Trickle, each node periodically broadcasts an information summary unless it has recently received an identical summary from a neighbor. If a node receives a summary older than its own, it will broadcast its own summary along with update packets. This method of code distribution is more energy-efficient than multicast [4].

C. Computation Phase

Most sensor network applications are tied to microcontrollers and peripherals rather than the main CPU [4]. The types of microcontrollers used include application-specific integrated circuits / processors, as well as general processors using reduced or complex instruction sets [2]. The efficiency and power requirements of each application depend on how well suited the available hardware resources are to the application [2].

TinyOS [4] supports power management through the StdControl interface. This interface calls the Init, Start, and Stop functions for each supported device that is capable of utilizing power control. TinyOS is based on events and tasks; and ordinarily the processor remains powered down until woken by an event such as a communication device interrupt or a timer. Tasks are a method of deferred computation in response to events or other tasks. Tasks are executed in a queue on a first-come-first-serve basis. They can be interrupted by events, but any task resulting from the event will be placed in the back of the queue. When the task queue is empty, the processor and related devices are put to sleep.

D. Storage Phase

Sensor motes typically employ RAM and Flash as their storage mechanisms, and the main storage power costs come from reading and writing to Flash [4]. Large power savings can be had from buffering data in RAM and writing it to Flash as a full page rather than individual bytes. [4] describes this process as being up to 233x more power efficient than writing individual bytes. TinyOS supports buffering as part of the Deluge service [4].

As flash memory becomes cheaper, faster, larger, and more power efficient; it is becoming more effective to store data locally rather than transmit it. Under this model, nodes would behave like a group of small databases, only transmitting collected data when queried [4].

IV. Discussion

The most power intensive operations for sensor networks involve sensing and communication, and the largest gains are to be had in those phases. Computation and storage, on the other hand are relatively cheap, although they are still significant sources of power consumption.

Given the relatively high power cost of communication, it is advantageous from a power standpoint to increase the level of onboard processing and storage in order to reduce the number and size of messages that need to be transmitted. As computation and storage become cheaper and more power efficient, this option will probably become the norm, and sensor networks will be treated like large distributed databases. The only drawback here is that eventually the nodes will be damaged or depleted, and at that point their collected data will be lost.

Efficiency in the communication phase depends on keeping the transceiver powered down as often as possible while still maintaining adequate connectivity. Duty cycling with long preambles is one way to accomplish this, as are the polling mechanisms in S-MAC and 802.15.4. Triggered operation can eliminate the need for long preambles or complicated polling mechanisms, and we may start to see this option become more common in the future if a suitable ultra-low-power transceiver / decoder can be developed for use as a wake-up radio.

Power efficiency in the sensing phase is critical, and besides reducing the requirements of the sensors, the greatest gains are to be had from increasing the efficiency of the surveillance state. Duty cycling is efficient for events that have sufficient predictability or a large enough detection window. Sensor hierarchy can also be used to increase efficiency by allowing the node to make intelligent decisions using data from low-power sensors to determine if external events warrant the use of higher-power sensors.

For target-tracking networks, soft deployment can vastly decrease the power requirements of the surveillance state by spreading the task cost across the network. The physical deployment of a network is constrained by the requirements of the tracking state, which requires many more nodes than the surveillance state, so there is no reason not to let most nodes sleep while a few maintain watch. Soft deployment is also attractive because it allows the tradeoff between vigilance and power use to be adjusted as needs change.

On the computation side and storage side, hardware and software improvements can increase efficiency by small but noticeable amounts. For hardware, the most important thing is to have the correct type of microprocessor (general or application-specific) for the task at hand. Efficient storage is also important, and improvements to Flash memory help with this. For software, TinyOS is an excellent operating system for sensor nodes, as it was built from the ground up with power efficiency in mind. TinyOS already includes many power management techniques, such as writing to Flash a full page at a time or powering down peripherals when not in use, so the application programmer does not need to worry about it.

V. Conclusion

In this paper, we discussed power control methods for sensor networks in terms of four broad categories: sensing, communication, computation, and storage.

For sensing, we covered duty cycling and sensor hierarchy. We also discussed sensor soft deployment for surveillance applications.

Methods of power control for communication included the radio management techniques of polled operation, scheduled operation, and triggered operation, as well as batching and aggregation. Middleware improvements dealt with time synchronization, routing, and code maintenance. For computation, we covered the importance of selecting the right type of microprocessor for the job, as well as the event / task based architecture and power control interface of TinyOS. In the storage phase, we highlighted the importance of writing to memory efficiently, by page instead of by byte, as well as the trend toward local storage instead of immediate transmission of data.

References

[1] Gartner Research, *Hype Cycle for Wireless Networking Infrastructure, 2008*. July 2008, 26-45.

[2] S. Corroy, J. Beiten, J. Ansari, H. Baldus, P. Mahonen, *Energy Efficient Selection of Computing Elements in Wireless Sensor Networks*, 2008 Second International Conference on Sensor Technologies and Applications, August 2008, 312-318

[3] P. Le-Huy, S. Roy, *Low-Power 2.4 GHz Wake-Up Radio for Wireless Sensor Networks*, 2008 IEEE International Conference on Wireless & Mobile Computing, Networking & Communication, October 2008, 13-18

[4] P.K. Dutta, D.E. Culler, *System Software Techniques for Low-Power Operation in Wireless Sensor Networks*, 2005 International Conference on Computer Aided Design, May 2005, 925-932

[5] J. Taneja, J. Jeong, D. Culler, *Design, Modeling, and Capacity Planning for Micro-solar Power Sensor Networks*, International Conference on Information Processing in Sensor Networks, April 2008, 407-418

[6] C. Gui, P. Mohapatra*, Power Conservation and Quality of Surveillance in Target Tracking Sensor Networks*, International Conference on Mobile Computing and Networking, 2004, 129-143

[7] Y. Yu, L. Rittle, V. Bhandari, J. LeBrun *Supporting Concurrent Applications in Wireless Sensor Networks*, Proceedings of the 4th International Conference on Embedded Networked Sensor Systems, Boulder, CO, USA, 2006, 139-152.

[8] C. Buratti, A. Conti, D. Dardari, and R. Verdone, An Overview onWireless Sensor Networks Technology and Evolution, Sensors journal, August 2009.

[9]F. Fabbri, C. Buratti, R. Verdone, J. Riihij P. M. onen, Area Throughput and Energy
Consumption for Clustered Wireless Sensor Networks. In Proceedings of IEEE WCNC 2009, Budapest, Hungary, 2009.

[10] O. Zytoune, M. El Aroussi, D. Aboutajdine, An energy efficient clustering protocol for routing in Wireless Sensor Network, International Journal of Ad Hoc and Ubiquitous Computing, PP 54-59, Volume 7, Number1, 2011.

Design and Implementation of Wireless Sensor Network based Ubiquitous Greenhouse System

Jeong-hwan Hwang[1], Mi-suk Kim[2], Hyun Yoe[*]
Dept. of Information and Communication Engineering Sunchon National University
Suncheon, Republic of Korea
jhwang[1], mskim[2], yhyun*@sunchon.ac.kr

Abstract—**The WSN(Wireless Sensor Networks) technology is one of the important technologies to implement the ubiquitous society, and it could make many changes in the existing agricultural environment including livestock rearing, cultivation and harvest of agricultural products if such a WSN technology is applied to the agricultural sector. This study attempts to establish ubiquitous agricultural environment and improve productivity of the greenhouse's crops by proposing ubiquitous greenhouse system using WSN technology. In proposed ubiquitous greenhouse system, soil sensors and environmental sensors are installed inside/outside the greenhouse in order to collect environmental information for greenhouse's crop growth such as environmental information, and soil information, and these sensors construct a wireless sensor network each other to collect environmental and soil information in the greenhouse. In addition, CCTVs are installed inside/outside the greenhouse to collect image information in real time for collecting greenhouse and crop image information and preventing dangers such as burglary and fire. Such collected environmental and image information is stored in the server via the gateway, provided to users in real time through various interfaces, and environmental control facilities in the greenhouse could be automatically or manually controlled suitable to optimum growth environment of cultivated crops based on collected information. Farmers may increase production and improve crop quality through this ubiquitous greenhouse system and prepare a data base with information collected from environment factors and control devices of the greenhouse, which is expected to provide information for control strategy of greenhouse operation.**

Keywords-wireless sensor networks, middleware, ubiquitous, greenhouse

I. INTRODUCTION

WSN(Wireless Sensor Networks) is a technology that sensor nodes capable of computing and communication shall be deployed to various application environments so that they can form an independent network, then physical information collected by wireless from the network shall be utilized for monitoring and controlling etc.[1][2] This WSN technology contributes to realizing high productivity, safety and high humans life level through its application to various industries such as distribution, logistics, construction, transportation, military defense and medical service etc.[3]

In particular, it is labor-intensive industry compared to other industry, and when applying WSN technology to agricultural area which lacks IT technology application, added value and productivity of agriculture can be increased.[4][5]

Recently, various studies combining WSN technology with protected agriculture such as greenhouses and stables and precision farming has been conducted,[6][7] and advanced nations including EU and the U.S. have established monitoring system from cultivation environment to production management and distribution in order to secure production of agricultural and stockbreeding products and transparency of distribution routes.[8][9][10]

However, Korea lacks researches applying agriculture and IT technologies as well as tools to collect growth environment information and analyze monitoring data compared to advanced nations, and its cost of output is high but output is low because the level of optimized technology for environmental control is low compared to advance nations.[11]

In order to solve these problems, we would like to propose a ubiquitous greenhouse system applying WSN technology in this study.

The ubiquitous greenhouse system proposed in this study is a system that they install WSN and CCTV in the greenhouse to collect its environmental and visual information, and from which they can monitor the greenhouse via the Web outside the greenhouse and further can control the greenhouse's facilities by manual even in remote place.

In addition, greenhouse facilities could be automatically controlled based on the crop growth environmental value which is already set up and SMS notice service shall be provided to users when dangerous situation occurs.

The proposed system improves productivity by maintaining optimized environment for growth and development through information on environment for greenhouses and growth and development of the crop, and it not only reduces production cost by optimizing management of production components but also provides convenience to producers through wire-wireless remote automatic control of environment for growth and development of crops.

[*] Corresponding Author

The composition of this article is as follows; Session 2 explains the structure of proposed ubiquitous greenhouse system structure and provided service process, and session 3 the results of realizing proposed systems. Finally, it conclusions in the last session.

II. Design of the Proposed Wireless Sensor Networks based Ubiquitous Greenhouse System

A. *System Structure*

The proposed ubiquitous greenhouse system shall be classified to three layers as figure 1, and each layer is composed of the physical layer which consists of environmental sensor, soil sensor, CCTV and other control facilities of the greenhouse, and the application layer which consists of interfaces that support monitoring the greenhouse and controlling service of the crop growth environment, and finally of the middle layer which supports the communication between the physical layer and the application layer, and store the collected information of the greenhouse to data base, provides monitoring and controlling service as well as keeping optimal status of the crop growth environment.

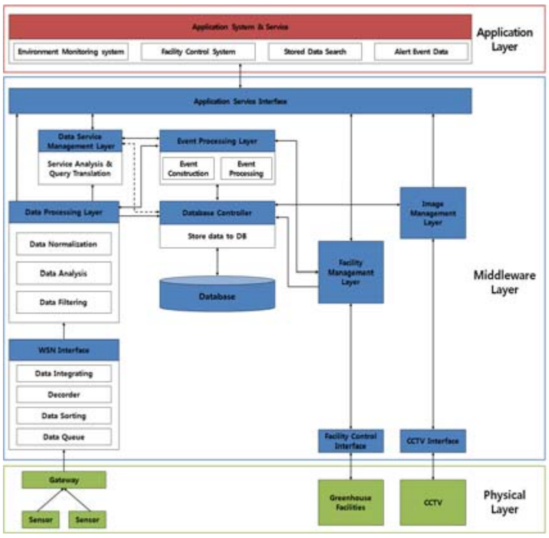


Figure 1. Ubiquitous Greenhouse System Structure

Physical layer is composed of sensor collecting information on external and internal environment of greenhouse and growth of crops, CCTV collecting information on images of greenhouses, and environment control facilities to create the optimized environment for greenhouse's crops.

Sensors are broadly divided into environment sensor collecting information on internal and external environment and growth sensor collecting information on growth of crops. Environmental sensor measure information of internal and external environment of greenhouse such as intensity of illumination, temperature, humidity, wind direction, wind speed EC, pH, and CO2, which affect the growth of crops. Growth sensor measures change of growth and development of crops such as temperature of leaves and parts of stems, weight of plant body, fruit temperature and volume.

CCTV is installed inside and outside of greenhouses and internal CCTV is for collection of information of images of crops and external one for prevention of dangers such as burglary and fires.

Environment control facilities include ventilation and heating systems that can control greenhouse environment which affects the growth of crops such as illumination, temperature, EC, pH, and CO2, systems to keep warm for reduction of energy, systems of controlling curtains to shade the light according to the intensity of light, systems for circulating fans to control the circulation of air inside of facilities, systems to control temperature of hot water, working fluid, and systems to control an artificial source of light according to external intensity of light, and each facility of environment control is controlled by power controller.

Middleware layer has features to collect data occurred in the physical layer such as environmental sensors and soil sensors installed for monitoring the greenhouse environment, and to lower the load on the application program by filtering real time information and provide data which application program requests.

In order to meet all these requirements, the middleware layer has functions to refine, filter and convert the collected information and support functions to recognize the situation and process and control the respective situation information when an event occurred.

The WSN interface layer provides a common interface function to heterogeneous sensors, and offers continuous monitoring and control functions for status of sensor networks.

DPL(Data Processing Layer) plays a role to provide a function to process various queries for data collected from the WSN infrastructure and a real-time management function for sensor information. In addition, the data processing layer employs a data management component, which delivers valid messages to the upper layer and supports queries of various forms by filtering data for reducing system's load that could be generated due to an enormous amount of data collected via sensors.

EPL(Event Processing Layer) is composed of an event construction and an event processing, and the event construction builds up the WSN data received from the lower layer as a simple event. The event processing specifies a reference value through a complex event language defined by the upper application layer and builds up the constructed simple event as a significant event considering the reference value.

DC(database controller) stores data generated in real-time into the database, supports queries of the application layer to use data at the time when the data is needed, and stores and updates the reference data for automatic control and condition notification of greenhouse facilities.

FML(Facility Management Layer) converts the control signal transmitted from the application layer or EPL to a proper typed data format, and transmit it to the greenhouse control facility to control, and transmit the status, operation time and control frequency of the control facility to DC to store them into the database.

IML(Image Management Layer) provides the Web with stream data of images taken from CCTV and classifies them to the greenhouse ID and camera number etc to store into the database.

The database plays a role to store environmental data collected from sensors installed inside/outside the greenhouse, image data collected from CCTVs, conditions, operating time and the number of controls of greenhouse controlling facilities, environmental reference values for automatic control and condition notification into each table.

The application service interface plays a role to deliver and establish the range of services established by the application. And data stored in the DB could be searched and stored by connecting the DB controller for demands of the application. In addition, it supports various queries to constitute flexible connections between applications and hardwares.

Application layer is composed of application services supporting various platforms such as web, PDA, smart-phone, which can provide users with crop growth information monitoring services, greenhouse environment monitoring services, greenhouse image monitoring services, crop growth and development environment control services.

B. Service Processes

Crop growth information and greenhouse environment monitoring services save greenhouse's crop growth and development information such as information of internal and external environment of greenhouse such as intensity of illumination, temperature, humidity, wind direction, wind speed, EC, pH, and CO2, which collected from sensors installed in greenhouses and information of growth and development of crops such as temperature of leaves and parts of stems, weight of plant body, fruit temperature and volume, and it shows this information to producers through GUI. The Figure 2 shows the process of operation of monitoring services for growth and development of crops and greenhouse environment.

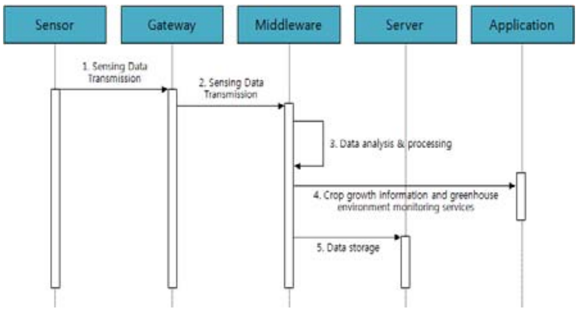

Sensor
Gateway
Middleware
Server
Application
1. Sensing Data Transmission
2. Sensing Data Transmission
3. Data analysis & processing
4. Crop growth information and greenhouse environment monitoring services
5. Data storage

Figure 2. Greenhouse Environment & Crop Growth Information Monitoring Service Operations

Image monitoring services for greenhouses are to provide producers and consumers with images of greenhouses and crops through CCTV installed inside and outside of greenhouses. The Figure 3 shows operation process of image monitoring services.

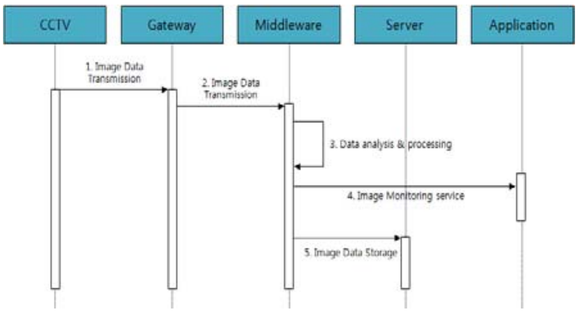

CCTV
Gateway
Middleware
Server
Application
1. Image Data Transmission
2. Image Data Transmission
3. Data analysis & processing
4. Image Monitoring service
5. Image Data Storage

Figure 3. Greenhouse Image Monitoring Service Operations

Greenhouse facility control service is a service which controls the greenhouse control facility automatically in order to keep the optimal crop growth environment based on the collected information from the greenhouse or helps users controlling the facility manually. The Figure 4 shows the process of operating automatic control services of greenhouse environment control facilities.

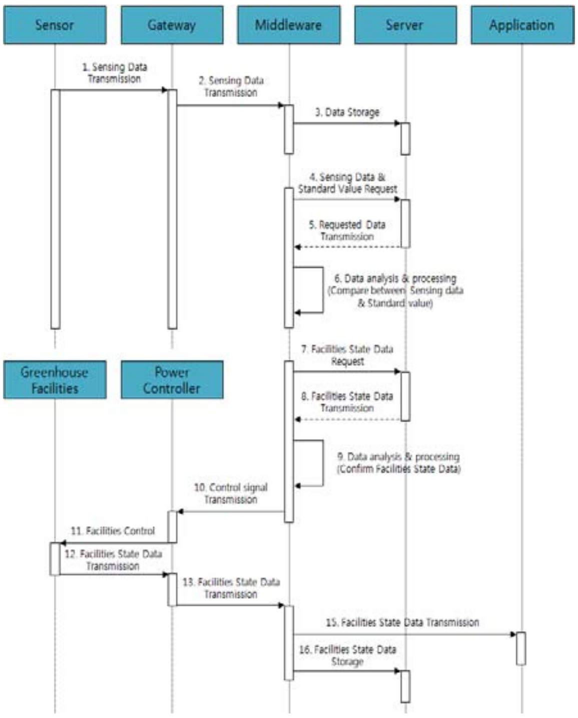

Sensor
Gateway
Middleware
Server
Application
1. Sensing Data Transmission
2. Sensing Data Transmission
3. Data Storage
4. Sensing Data & Standard Value Request
5. Requested Data Transmission
6. Data analysis & processing (Compare between Sensing data & Standard value)
7. Facilities State Data Request
Greenhouse Facilities
Power Controller
8. Facilities State Data Transmission
9. Data analysis & processing (Confirm Facilities State Data)
10. Control signal Transmission
11. Facilities Control
12. Facilities State Data Transmission
13. Facilities State Data Transmission
15. Facilities State Data Transmission
16. Facilities State Data Storage

Figure 4. Greenhouse Facilities Automatic Control Service Operations

The Figure 5 shows manual control services of greenhouse environment control facilities.

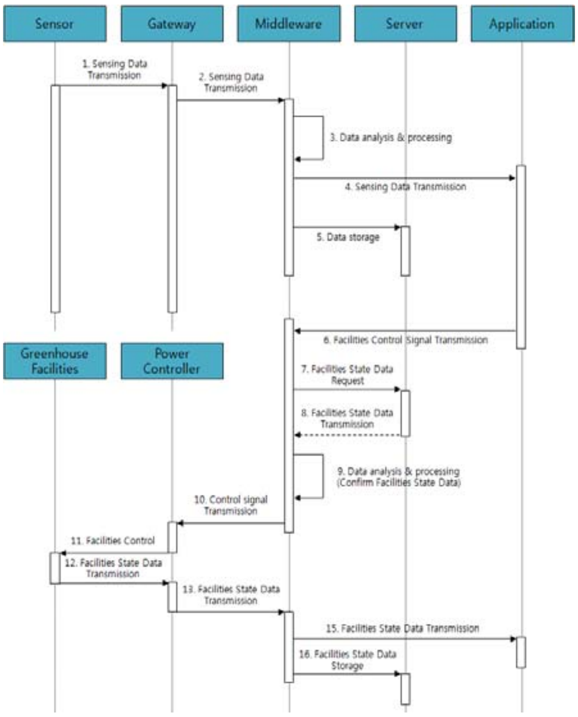

Sensor
Gateway
Middleware
Server
Application
1. Sensing Data Transmission
2. Sensing Data Transmission
3. Data analysis & processing
4. Sensing Data Transmission
5. Data storage
6. Facilities Control Signal Transmission
Greenhouse Facilities
Power Controller
7. Facilities State Data Request
8. Facilities State Data Transmission
9. Data analysis & processing (Confirm Facilities State Data)
10. Control signal Transmission
11. Facilities Control
12. Facilities State Data Transmission
13. Facilities State Data Transmission
15. Facilities State Data Transmission
16. Facilities State Data Storage

Figure 5. Greenhouse Facilities Manual Control Service Operations

Greenhouse situation alarming service is to prevent dangerous situations in advance by informing users of change of weather and conditions of greenhouses and letting them take measures. The Figure 6 shows operation process of greenhouse situation alarming service.

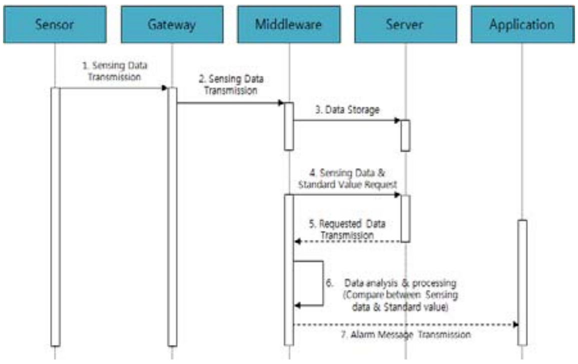

Sensor
Gateway
Middleware
Server
Application
1. Sensing Data Transmission
2. Sensing Data Transmission
3. Data Storage
4. Sensing Data & Standard Value Request
5. Requested Data Transmission
6. Data analysis & processing (Compare between Sensing data & Standard value)
7. Alarm Message Transmission

Figure 6. Greenhouse Situation Alarming Operations

III. IMPLEMENTAION OF THE PROPOSED WIRELESS SENSOR NETWORKS BASED UBIQUITOUS GREENHOUSE SYSTEM

A. Implementation

In order to measure information of internal and external environment for greenhouse, sensors were installed inside and outside of greenhouses as seen in Figure 7, and external weathers such as temperature and humidity of outside and inside of greenhouses, speed of light, and wind speed and direction and internal weathers such as temperature, relative humidity, intensity of light in the upper and lower parts of crops, and the amount of penetration of light.

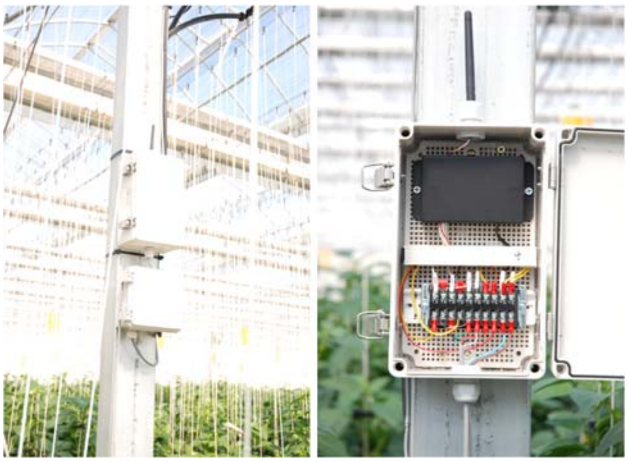


Figure 7. Environmental Sensor

In addition, since management of rooting zone which greatly affects absorption of nutrient solution culture considering the characteristics of greenhouse's crop cultivation, which mainly uses nutrient solution culture is very important, the amount, EC and pH of supplying liquid, the amount, EC, and pH of waste liquer, rate of absorption and temperature within culture medium, and temperature of supplying water and waste liquer, which affect rooting zone of plants inside of greenhouses, were measured by installing sensors to collect information of environment of rooting zone as seen in Figure 8.

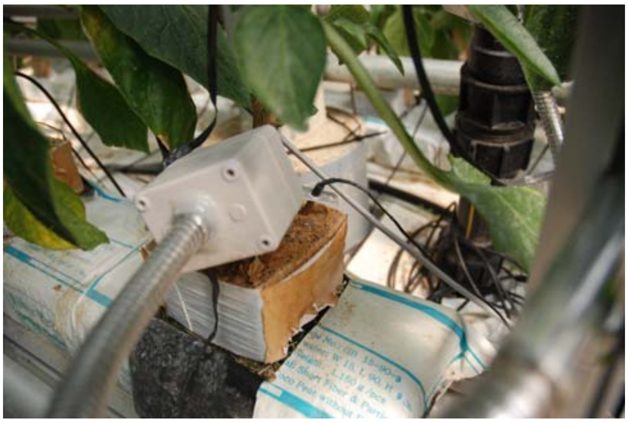


Figure 8. Rooting Zone Sensor

In order to understand information of growth and development of crops, sensors were installed like Figure 9 and temperature of plants inside of greenhouses, temperature of upper and lower parts of stems, temperature and volume of fruits, weight of body of plants, plant heights of body of plants, the amount of water and light that crops absorb, yields, and rate of increase of weight of the body of plants are measured.

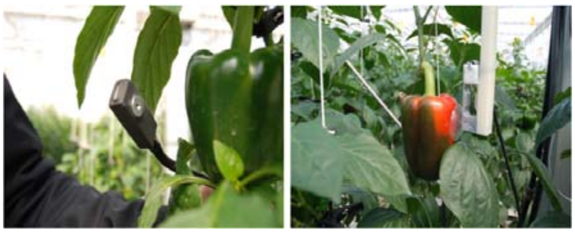


Figure 9. Crop Growth Sensor

In order to create the optimized crop growth environment based on the crop growth information from sensors installed in inside and outside of greenhouses, weather information of inside and outside of greenhouses, and information of rooting zone environment, environment control facilities are installed in greenhouses and in order to control these facilities, Power controller was installed as seen Figure 10.

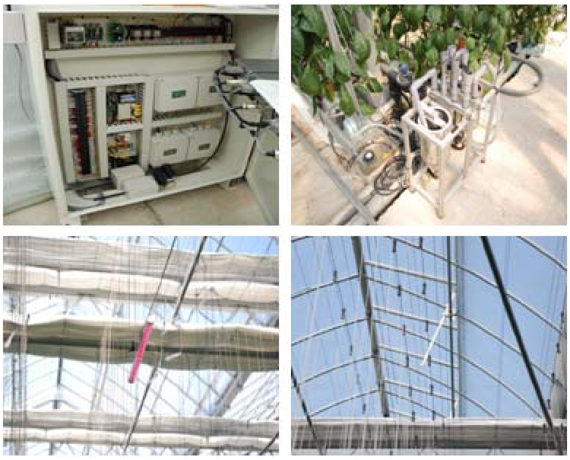
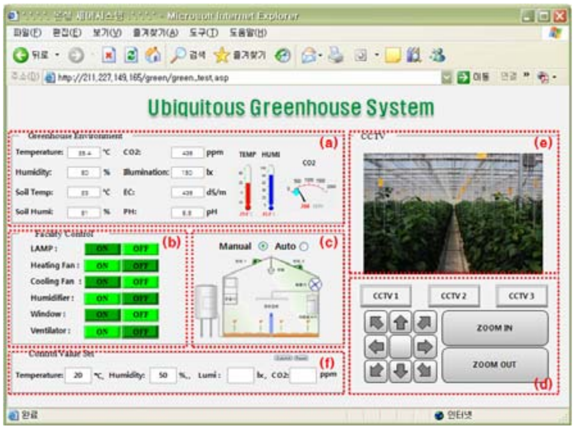


Figure 10. Greenhouse Facilites and Power Controller

In order to monitor and control greenhouses, as seen in the figure 11, GUI for managers is developed as Web environment. WAS uses Tomcat-6.0.20 and database mysql 5.0 which is the safest version among the versions that are currently released.



http://211.227.149.165/green/green_test.asp

Ubiquitous Greenhouse System

Greenhouse Environment (a)

Temperature: [illegible] ℃ CO2: [illegible] ppm
Humidity: [illegible] % Illumination: [illegible] lx
Soil Temp: [illegible] ℃ EC: [illegible] dS/m
Soil Humi: [illegible] % PH: [illegible] pH

Facilty Control (b)

LAMP : ON OFF
Heating Fan : ON OFF
Cooling Fan : ON OFF
Humidifier : ON OFF
Window : ON OFF
Ventilator : ON OFF

(c) Manual ● Auto ○

Control Value Set (f)

Temperature: 20 ℃, Humidity: 50 %, Lumi : lx, CO2: ppm

Figure 11. Ubiquitous Greenhouse System Web GUI

In GUI for manager, the values of sensing which are measured in sensors installed inside and outside greenhouses are appeared in (a), and (b) shows control of equipment in greenhouses and its conditions. (c) expresses the conditions of equipment in (b) as graphics. (d) is the part to control CCTV, (e) the part to show collected images through CCTV, and (f) the part to enter standard values for automatic greenhouse control.

B. *Results*

As a result of applying the proposed system as mentioned above to actual greenhouses, information of environment and images in greenhouses is collected through sensors and image supervision camera, and GUI which is intuitive to users can monitor and control conditions of greenhouses. The Figure 12 is a graph that shows data of environment to growth and development measured by installing the proposed ubiquitous greenhouse system to greenhouse.

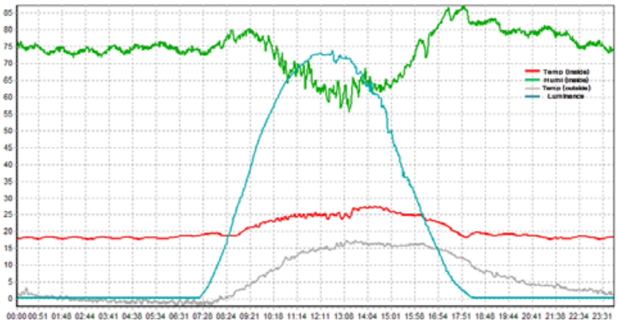


Figure 12. Greenhouse Environment Data Graph (2010.2.23)

IV. Conclusions

This study proposes ubiquitous greenhouse system for comprehensive management of greenhouses requiring precise management of environment for crop growth and development.

The proposed system is composed of physical layer, middleware layer and application layer, and components of each layer collect and manage information of environment for growth and development within greenhouses. Not only the information is delivered to users but remote manual and automatic control in greenhouses also improves users' convenience and productivity, and based on the data of environment for growth and development gained by operating the system, the optimized environment for growth and development of greenhouse's crop is created.

To prove the proposed system, it was tested in greenhouse by installing sensors including sensors for soils, environment, temperature and humidity of leaves and CCTV and was implemented. As a result of the implementation, monitoring and controlling of the environment for growth and development could be conducted through GUI and the results of sensors monitoring related to controlling greenhouses and controlling shows that there is no wrong operation.

Through the study, It is expected that applying the proposed system to greenhouse farms shall help saving labour cost and producing high quality crops, and further obtaining competitiveness of our agriculture

Acknowledgment

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2011-(C1090-1121-0009))

References

[1] Akyildiz, I.F.; Su, W.; Sankarasubramaniam, Y.; Cayirci, E. A survey on Sensor Networks. IEEE Commun. Mag. 2002, 40, 102-114

[2] Chong, C.Y.; Kumar, S.P.; Hamilton, B.A. Sensor networks: Evolution, opportunities, and challenges. Proc. IEEE 2003, 91, 1247-1256.

[3] Pyo, C.-S.; Chea, J.-S. Next-generation RFID/USN technology development prospects. Korea Inform. Commun. Soc. Inform. Commun. 2007, 24, 7-13.

[4] Lee, M.-H.; Shin, C.-S.; Jo, Y.-Y.; Yoe, H. Implementation of green house integrated management system in ubiquitous agricultural environments. J. KIISE 2009, 27, 21-26.

[5] Shin, Y.-S. A Study on Informatization Model for Agriculture in Ubiquitous Era; MKE Research Report; National IT Industry Promotion Agency: Seoul, Korea, 2006.

[6] Jeong, B.-M. Foreign u-Farm Service Model Casebook; Issues and Analysis Report of Korea National Information Society Agency, NCA V–RER-06005; Korea National Information Society Agency: Seoul, Korea, October 2006.

[7] Kwon, O.-B.; Kim, J.-H. A Basic Direction for Building Agricultural Radio Frequency Identification Logistics Information System. M85; Korea Rural Economics Institute: Seoul, Korea, December 2007.

[8] Yoo, N.; Song, G.; Yoo, J.; Yang, S.; Son, C.; Koh, J.; Kim, W. Design and implementation of the management system of cultivation and tracking for agricultural products using USN. J. KIISE 2009, 15, 617-674.

[9] Kim, M.; Son, B.; Kim, D.K.; Kim, J. Agricultural Products Traceability Management System based on RFID/USN. J. KIISE 2009, 15, 331-343.

[10] Hwang J.H, Yoe H., "Study of the Ubiquitous Hog Farm System Using Wireless Sensor Networks for Environmental Monitoring and Facilities Control", Sensors, 10, 10752-10777 (2010)

[11] Jeong, W.J.; Myoung, D.J.; Lee, J.-H. Comparison of climatic conditions of sweet pepper's greenhouse between Korea and The Netherlands. J. Bio-Environ. Contr. Korea 2009, 18, 244-252.

Design And Implementation of middleware for Cattle barn based on Ubiquitous Sensor network

Jiwoong Lee
School of Information and Communication Engineering
Sunchon National University
Sunchon, Korea
Leejiwoong@sunchon.ac.kr

Hyun Yoe[♀]
School of Information and Communication Engineering
Sunchon National University
Sunchon, Korea
yhyun@sunchon.ac.kr

Abstract—**The recent trend in research and development on ubiquitous computing technologies is towards the direction to provide users optimum services suitable to conditions through context awareness, inference and cooperation based on data collected from various sensor nodes To build such the ubiquitous application services easily, a middleware is needed to connect the RFID/USN's hardware with the applications or the enterprise systems. The USN middleware technology is used to filter lots of duplicate data collected from many sensor networks and convert the raw data into meaningful information for users to send it to applications, and provides services to make users could decide contextual information quickly and correctly through the data mining technique and analysis method. Even though it has been presently carried out the studies on such a USN middleware to apply it for various fields but there are very few studies on the middleware suitable to livestock environment which applications of IT technology have not been sufficient relatively comparing to other industries. In particular, for controlled barn, there are many difficulties on user's decision-making for efficient raise cattle due to a number of environmental factors affecting cattle. In order to solve such problems, this paper is trying to propose a USN middleware suitable to livestock environment, which could collect cattle barn environmental information and optimally manage cattle through barn automation. The proposed middleware is composed of a sensor manager, context manager and control manager, which collects a variety of data from heterogeneous sensor networks, processes the collected data into information suitable to user's demand, and sends it to controllers of controlled agriculture, so that it could support users to be provided various application services and make decisions adequate to conditions.**

Keywords-middleware, RFID, USN

I. INTRODUCTION (HEADING 1)

The recent innovation in IT technology is accelerating the fusion between industries. The fusion between IT and traditional industries continuously goes on. The application of ubiquitous technology to livestock, which is a primary industry, is getting expectation that the convergence technology would enhance the added-value and productivity of agriculture [1].

In order to establish such u-agriculture environment successfully, the core ubiquitous technology development optimized to cattle barn, such as sensor hardware, environment hardware and cattle barn environment application service, would be essentially required [2].

To build such the ubiquitous application services easily, a middleware is needed to connect the RFID/USN's hardware with the applications or the enterprise systems[3].

The middleware is a technology to filter lots of data collected from many heterogeneous RFID/sensor devices, process the event data, and then abstract it into meaningful information[3], and to send and process a great number of contexts and data arisen in the ubiquitous environment more efficiently[4].

Even though researches on the USN middleware are currently in progress for various fields, there are very few researches on the middleware focused on application services in agricultural environment that the application of IT technology is insufficient relatively comparing to other industries[5].

In particular, for the controlled cattle barn, the production and the quality of animal is affected by the consistent management of various environmental factors such as temperature, humidity, ammonia, wind speed etc.

affecting animal's growth, and the precision control of environmental control devices including ventilator, windows, heater, lighting, image processor etc, so many difficulties are arisen in producer's decision making. This paper would like to propose an USN middleware suitable to cattle barn environment, which could collect cattle barn environmental information and manage cattle optimally through barn automation in order to solve problems in such controlled agriculture environment.

The proposed middleware is designed to collect and monitor the environmental information from sensors installed in the barn, and to provide the optimum service to the cattle barn application service system by controlling the barn control devices through the corresponding context information processing when a problem is arisen, which helps user could be provided various application services and make a decision suitable to the situation.

This paper is organized as follows. Chap. 2 explains the related researches, Chap. 3 analyzes requirements on the

[♀] Corresponding author.

middleware to design the middleware based on the results, Chap. 4 implements the designed middleware, and finally Chap. 5 draws the conclusion of this paper.

II. Related Works

A. *A design of Cotext Aware Middleware based on Web Service in Ubiquitous Environment.*

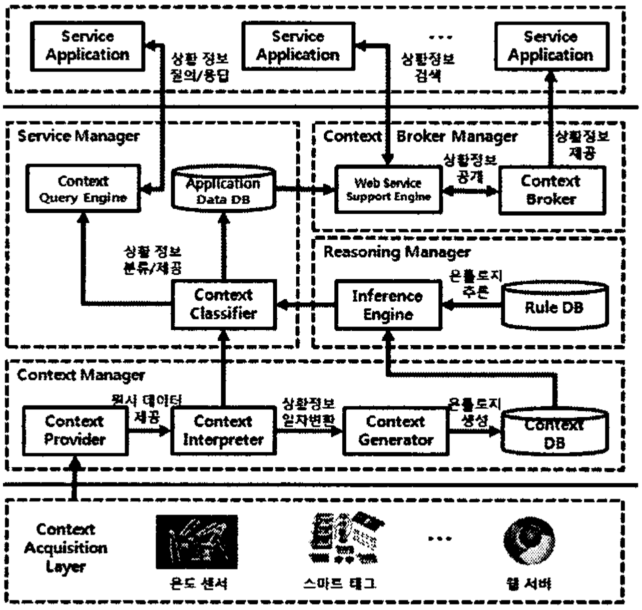


Fig 1. The data flow diagram of Ws-CAM Framework

Context-aware technologies for ubiquitous computing are necessary to study the representation of gathered context-information appropriately, the understanding of user's intention using context-information, and the offer of pertinent services for users. [7] this paper propose the WS-Cam(Web Services based Context-Aware Middleware) framework for context-aware computing. WS-CAM provides ample power of expression and inference mechanisms to various context-information using an ontology-based context model. this also consider that WS-CAM is the middleware-independent structure to adopt web services with characteristic of loosely coupling as a matter of communication of context-information. this paper describe a scenario for lecture services based on the ubiquitous computing to verify the utilization of WS-CAM. this paper also show an example of middleware-independent system expansion to display the merits of web-based services. WS-CAM for lecture services represented context-information itodomaits as OWL-based ontology model effectively, and confirmed the information is inferred to high level context-information by user-defined rules. this paper also confirmed the context-information is transferred to application services middleware-independently using various web methods provided by web services[7].

B. *Implementation of an Application System using Middleware and Context Server for handling Context-Awareness*

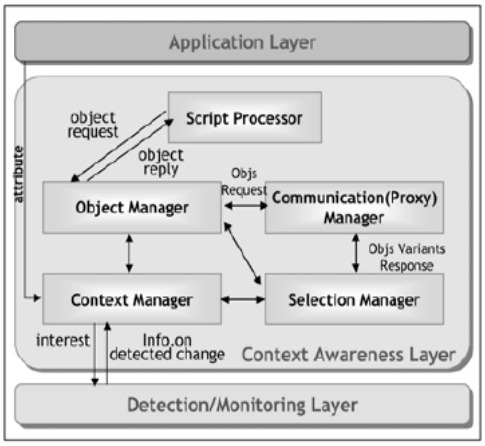


Fig 2. Middleware Structure

Context-awareness is a technology to facilitate information acquisition and execution bysupporting interoperability between users and devices based on users' context. It is one of the most important technologies in ubiquitous computing. this paper propose a middleware and a context server for dealing with context-awareness in ubiquitous computing and implement an application system using them.[8] The middleware proposed in this work plays an important role in recognizing a moving node with mobility by using a Bluetooth wireless communication technology as well as in executing an appropriate execution module according to the context acquired from a context server.[8] In addition, the proposed context server functions as a manager that efficiently stores into a database server context information, such as user's current status, physical environment, and resources of a computing system. [8]

Finally, this application system implemented in this work one which provides a music playing service based on context information, and it verifies the usefulness of both the middleware and the context server developed in this work.[8]

III. Middleware

A. *Middleware Requirement*

To control and manage a cattle barn efficiently, it should be considered the environmental aspect of system and the function of USN middleware.

First, To solve the problem as disease, different depending on the light environment, temperature and humidity environment in the barn. In addition, since producers may suffer a loss due to unnecessary heating bills, in order to cope actively with it[9][10], it is installed the sensors for environmental information including temperature, humidity, ammonia etc., and the control system such as heater, cooler controller, wind speed/wind direction controller, ventilator etc. for the optimum environment.

Second, the fundamental functions of middleware are the multiple query processing of collected services, management of sensing and meta information, creation of context information for sensing information, intelligent event processing required from the application layer[6].

Among them, sensors collect environmental variables (temperature, humidity etc.) to provide adequate services for cattle barn, event processing is carried out to process data for pre-registered conditions if the certain condition is satisfied, and the collected environmental information is compared and analyzed with existing collected data. In addition, a service is provided users to make adequate decisions by creating contextual information through prediction and inference.

The middleware is designed on the basis of such requirements.

B. Middleware Design

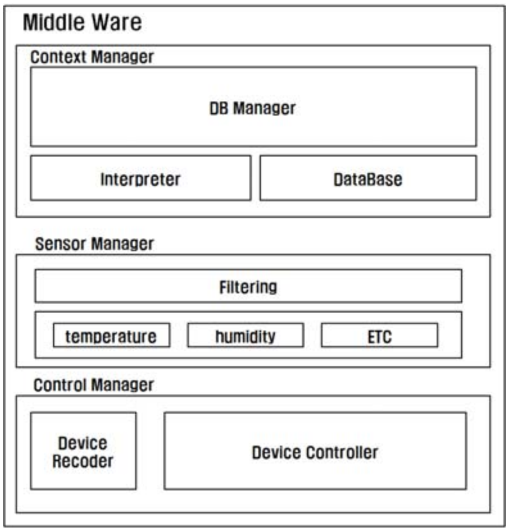


Fig 3. Middleware Structure for cattle barn

Fig. (3) shows the structure of USN middleware proposed in this paper for controlled barn automation, which is composed of a sensor manager (SM), context manager (CM), and control manager (CTM). The SM has a function to collect information taken place in the barn and to take charge of communication between middlewares, the CM has a function to analyze the raw data collected by the SM to convert it into actually usable information and to store and manage information. The CTM controls and manages the barn's device based on information analyzed by the CM.

C. Sensor Manager

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

The sensor manager is a module to deliver environmental information, which takes charge of interfaces between physical sensors and computers. The sensor manager carries out a function to collect information from the sensors including temperature, humidity, ammonia etc. installed in the barn and the control devices such as ventilator, heater etc., sends constant clock signals to synchronize sensors in order to correctly transfer data between sensors and gateways without errors, and removes duplicate data by filtering to send it to the CM since the efficiency is decreased due to lots of data when receiving duplicate data.

D. Context Manager

The context manager could effectively manage the various contextual information to intelligently provide it to users. Such contextual information may be collected from the various sensors installed in the facility, and also collected via the Web such as information of the other facilities or the surrounding area[7]. The context manager takes charge of managing functions to acquire, process, represent, and store information for users and surrounding environment of users obtained from the various sources as above[7]. The context manager is composed of a context interpreter and a context DB manager as the Fig. (4).

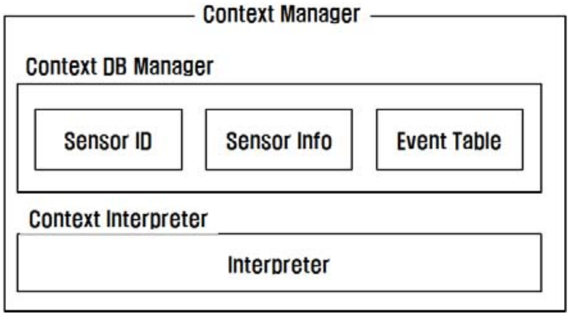


Fig 4. Context Manager Structure

The context interpreter takes charge of converting the raw data collected from sensors into semantics that could be comprehended at the user level. Such converted information is stored in the database through the context DB manager. The context DB manager is comprised of Sensor ID, SensorInfo, and EventTable.

TABLE I. SENSOR ID

ID	Sensor Function
1	*Humidity*
2	*Ammonia*
3	*FAN*

As the Table 1, Sensor ID table is comprised of sensor's ID and sensor's function attribute, and SensorInfo table allocates periods, time, measured values to be collected from sensors and assigns Group ID for each role of sensors.

TABLE II. SENSOR INFO

ID	Location	Sample Cycle	Time	Value	Group ID
1	*4-5*	*250*	*201102170317*	*24*	*3*
2	*6-2*	*250*	*201102170317*	*32*	*2*
3	*3-1*	*250*	*201102170317*	*11*	*1*

The contextual information could be created through the data analysis module, which analyzes conditions of environment in the barn and cattle based on such stored information, and the data mining technique, and the table is constructed as the Table 2 for intelligent event processing required by users. Certain problems occurring in the barn, i.e. many problems that temperature/humidity is too high to disease or the concentration of ammonia becomes Odor Causes, are predefined in the event table as the Table 3, and the barn is controlled if the problem is arisen. The interaction of CM is as follows.

TABLE III. EVENT TABLE

Event	Group ID average Value
Turn on The Fan	*40*
Turn off The Fan	*27*
Turn on The Cooler	*30*
…	…

The Group ID is given according to the sensor's function as the Table 2, the average value of sensor information values collected for each group is stored in the database as the form identical to the Table 4. The barn is automatically controlled if the event condition is satisfied on the basis of this value.

TABLE IV. GROUP ID

Group ID	Average Value
1	*37*
2	*26*
3	*35*
4	*24*

E. Sensor Manager

The control manager is composed of a device controller and a device recorder. The device controller requests the contextual information to control, and the device recorder records the current condition of control device to refer for the next service request.

Exploiting these two functions, the CTM uses the contextual information received by the CM to adequately control various control devices at the locations where the event is arisen, and sends the information to users in case of emergency. scenario for operating

F. Scenario for operating

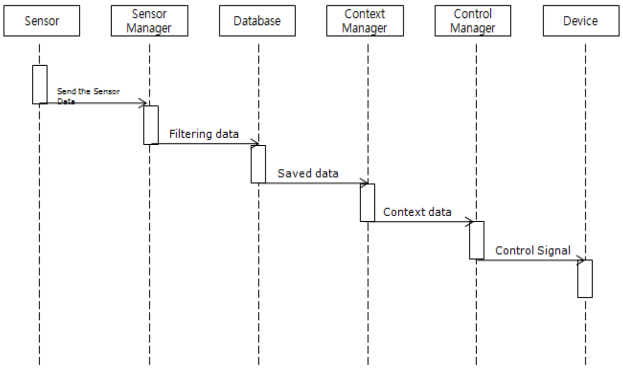
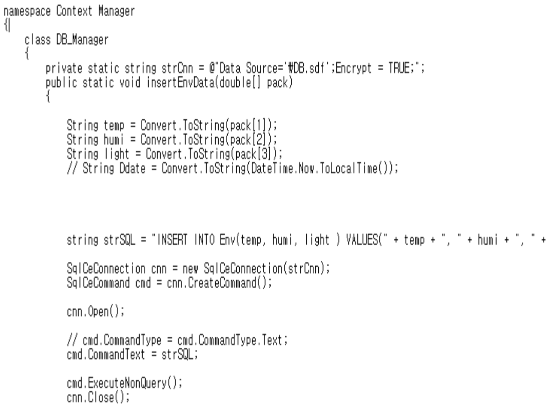


Fig 5. Scenario flowchart

Fig. 5 is the process of entire system. This SM periodically collects the environmental information such as temperature, humidity, ammonia, etc. from the sensor network installed in the cattle barn. The collected information is stored in the database through filtering in order to remove duplicate data. The CM creates the contextual information from the stored information through the data mining and analysis, and uses the contextual information to send the control signal to the CTM through the predefined event manager. The CTM controls the controller in the corresponding area to efficiently operate based on the received control signal.

IV. IMPLRMENTATION

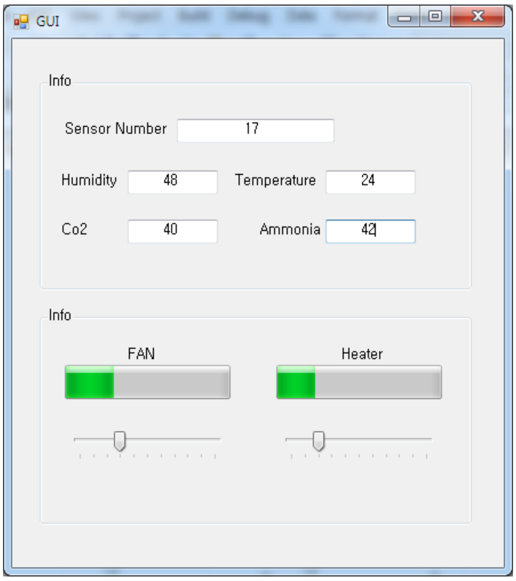
The proposed middleware is aimed at implementing the middleware for the cattle barn suitable to the livestock environment, and data collected through the sensor network is experimented for the event extraction according to the given conditions for the aim. In addition, it is constructed to confirm the results through the GUI implemented by the Microsoft Visual Studio 2005 C#.



```
namespace Context Manager
{|
    class DB_Manager
    {
        private static string strCnn = @"Data Source='#DB.sdf';Encrypt = TRUE;";
        public static void insertEnvData(double[] pack)
        {

            String temp = Convert.ToString(pack[1]);
            String humi = Convert.ToString(pack[2]);
            String light = Convert.ToString(pack[3]);
            // String Ddate = Convert.ToString(DateTime.Now.ToLocalTime());



            string strSQL = "INSERT INTO Env(temp, humi, light ) VALUES(" + temp + ", " + humi + ", " +

            SqlCeConnection cnn = new SqlCeConnection(strCnn);
            SqlCeCommand cmd = cnn.CreateCommand();

            cnn.Open();

            // cmd.CommandType = cmd.CommandType.Text;
            cmd.CommandText = strSQL;

            cmd.ExecuteNonQuery();
            cnn.Close();
```

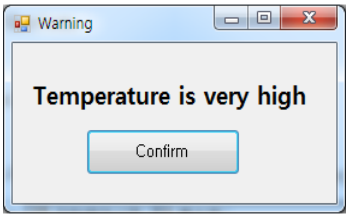
Fig 6. Context Manager Source

Fig. 6 is the CM implemented with the C#, which is part of codes storing the environmental information received from the SM into the database, and Fig. 7 is the GUI to confirm the results of the proposed middleware. Through the GUI in the facility of Fig. 7 (Info), it could be confirmed the environmental information values such as temperature, humidity, ammonia etc. collected from sensors, the intelligent event processing is confirmed through the event notification window as Fig. 8 opened when the contextual information exceeds the reference value, and the performance of middleware is confirmed by controlling various devices such as ventilator, heater etc in the barn through the Fig. 7 (Control).



Sensor Number 17
Humidity 48 Temperature 24
Co2 40 Ammonia 42

Fig 7. GUI



Temperature is very high

Fig 8. Event Notification

V. CONCLUTION

This paper designs and implements the middleware to control the barn according to the contextual information collected from sensors for the cattle barn automation suitable to the livestock. The middleware is composed of the sensor manager, context manager, control manager, which the sensor manager sends various environmental information to the context manager, the context manager creates the contextual information and analyzes the agricultural environment based on the event, and the cattle barn is controlled through the control manager, so it is minimized the problems that could be arisen in the barn. In addition, it is designed to monitor the information collected from sensors to support decision-making in the livestock site. It is expected that the high profit would be given to the farm if the stability and reliability of barn is secured and the collected growth condition of cattle is exploited through the middleware proposed in this paper.

ACKNOWLEDGMENT

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2011-(C1090-1121-0009))

REFERENCES

[1] K.H. Lee, C.M. Ahn, G.M. Park.: Characteristics of the Convergence among Traditional Industries and IT Industry, Electronic Communications Trend Analysis, Vol.23 No.2, pp13--22. (2008)

[2] Meong-hun Lee, Chang-sun Shin, Yong-yoon Jo, Hyun Yoe, “Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments” , Journal of KIISE, Vol.27 No. 6, pp. 21--26. (2009)

[3] Sang Hwan Kung*, Yoon Hee Kang*, Jin Ho Yoo, “ USN Based Middleware Software Design for Agriculture and Stockbreeding “, Proceedings of the KAIS Fall Conference, pp.788-791, 2009

[4] Kwon-jin Lee, Sung Keun Song, Hee Young Youn , “A New Context-Oriented Middleware for supporting Exact Context-Awareness in Ubiquitous Environment ” , Korea Computer Congress, Vol. 33, No.1(D)

[5] Sang Hwan Kung, " The Design of Fungus Cultivating System based on USN “ Korean Institute of Information Technology, pp. 34~41 , 2007.

[6] J.G. Hwang, T.S. Cheong, Y.I. Kim, Y.J.Lee, ETRI, "Trends of RFID Middleware Technology and its Aplications “ Electronics and telecommunications trends, V.20 no.3 = no.93, 2005

[7] Young-Rok Song, Yo-Seob Woo ” A Design of Context-Aware Middleware based on web Services in Ubiquitous Environment” , The Korea Institute of signal Processing and Systems, 10(4) 225-232 1229-9480

[8] Choon-Bo Shim, Bong-Sub Tae, Jae-Woo Chang, Jeong-Ki Kim, Seung-Min Park, “ Implementation of an Application System using Middleware and Context Server for Handling Context-Awareness “, Vol 12 no 1 pp.31-42, 2006

[9] Ji-woong Lee1, Ho-chul Lee1, Jeong-hwan Hwang1, Yongyun Cho1, Changsun Shin1, Hyun Yoe♀1 „Design and Implementation of wireless sensor networks based paprika Green house system “, Communications in Computer and Information Science, Volume 78, 638-646 ,2010

[10] Won-ju jeong, Jeong hyun Lee, Ho cheol Kim, Jong Hyang Bae " Dry Matter Production, Distribution and Yield of Sweet Peper Grown under Glasshouse and Plastic Greenhouse in Korea, Journal of Bio-Environment Control, 18(3):258-265 (2009)

Shorter Delay Transmission of Sensor Data by Intermediate Directional Nodes

Hitoshi Imahori and Hiroaki Higaki
Department of Robotics and Mechatronics
Tokyo Denki University, Tokyo, Japan

Abstract - *In wireless sensor networks, sensor data messages are required to be transmitted along a wireless multihop transmission route with shorter delay. For avoidance of corruption of sensor data messages, the RTS/CTS control is introduced in wireless LAN protocols which solves the hidden terminal problem among 2-hop neighbor wireless nodes. However, it causes additional transmission delay and various methods have been proposed to apply the RTS/CTS control conditionally. This paper proposes introduction of directional intermediate sensor nodes distributed sparsely in the sensor networks. Directional sensor nodes with directional antennas for space-division multiplexing reduces interferences among sensor nodes. It results in conditional omissions of the RTS/CTS control without collisions of data message transmissions among 1- and 2-hop neighbor nodes. Thus, it is expected for sensor data message transmissions to reduce transmission delay without additional losses of them.*

1 Introduction

Recently, wireless multihop networks are getting one of the most important research and development areas related to information and communication technologies due to their flexibility and low overhead for configuration and maintenance. Wireless sensor networks [2] consist of sensor nodes with sensing and wireless communication modules creating and transmitting data messages and sink nodes to which the data messages are transmitted. Usually, without continuous power supply, technologies for low-power wireless nodes are mandatory. Here, wireless multihop transmissions of data messages are critical where data messages are transmitted with help of intermediate sensor nodes. Since wireless communication is intrinsically broadcast-base, wireless signals transmitted by neighbor nodes interfere with each other. Various wireless LAN protocols such as IEEE 802.11 [11] introduce collision avoidance mechanism, CSMA/CA and RTS/CTS for interference between exposed and hidden nodes, respectively. However, the latter causes longer transmission delay due to contentions among 1- and 2-hop neighbor nodes. This paper proposes introduction of directional wireless nodes to wireless sensor networks which is expected to achieve shorter transmission delay by space-division multiplexing without time-overhead caused by RTS/CTS.

2 Related Works

A wireless sensor network is a wireless multihop network composed of wireless sensor nodes and sink nodes. Data messages are transmitted from a sensor node to a sink node along a wireless multihop transmission route which is a sequence of intermediate sensor nodes. Multiple source sensor nodes might achieve sensor data simultaneously and multiple data messages might be transmitted concurrently to a sink node. For transmissions of data messages between neighbor sensor nodes (and between a sensor node and a sink node), a wireless LAN protocol such as IEEE 802.11 and Bluetooth is applied. For avoidance of corruption of data messages by collisions among wireless signals, it is required for each intermediate sensor node to occupy its wireless signal transmission range temporarily. CSMA/CA and RTS/CTS are widely available mechanisms for avoidance of collisions caused among exposed nodes and hidden nodes.

CSMA/CA seems mandatory under an assumption that all wireless sensor nodes transmit control and data messages through the same channel, i.e., by using wireless signals with the same wavelength. On the other hand, though RTS/CTS is one of reasonable time-division solutions for collisions among hidden nodes, it requires higher communication overhead. That is, it causes longer transmission delay and lower efficiency. Figure 1 shows a sequence of control and data messages for a data message transmission in IEEE 802.11. Here, 11.2–31.7% of required transmission time 911–2590 μs is for avoidance of collisions by RTS/CTS. Thus, RTS/CTS is adaptively applied only in cases with data messages longer than RTSThreshold in IEEE 802.11 specification since the shorter the data message is, the lower the expectation of occurrences of collisions are. In [1], another adaptive RTS/CTS mechanism has been proposed for reduction of communication overhead. Since the expectation of collisions among hidden nodes depends on a number of hidden nodes of a sender node, RTS/CTS is adaptively applied based on a number of neighbor nodes of a receiver node.

The reason why such mechanisms have been proposed is that the problem of collisions among intermediate nodes in a wireless multihop transmission route is in-

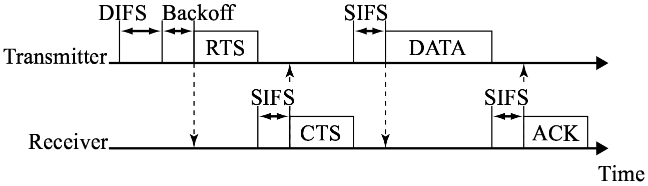


Figure 1. RTS/CTS in IEEE 802.11.

evitable. According to widely-accepted disk model [10] for wireless sensor nodes with omni-antennas, 2-hop previous- and next-hop nodes of each intermediate node are always hidden nodes. In [7], RH2SWL is proposed for space-division solution of intra-route collisions among hidden nodes by routing and transmission power control. Here, under an assumption that each intermediate node controls its transmission power to the minimum to reach its next-hop node, a wireless transmission route in which each intermediate node is included in a wireless transmission range of its previous-hop node and is not included in that of its next-hop node is detected for data message transmissions. Though it achieves high data message throughput without RTS/CTS, multiple channels are required for applying it to a large-scale wireless multihop networks [6].

3 Proposal

As sensor data messages are required to be transmitted without losses due to high retransmission overhead and with shorter transmission delay, in order to reduce communication overhead for RTS/CTS, this paper proposes another space-division solution by introduction of wireless nodes with directional antennas. This paper assumes that a directional sensor node has 6 directional antennas with the same transmission range without overlaps and the same total coverage of an omni wireless node as shown in Figure 2 [5]. Each directional antenna in a sensor node independently transmits and receives control and data messages of the others. Thus, it transmits and receives data messages simultaneously through different antennas and it also receives data messages simultaneously through different antennas.

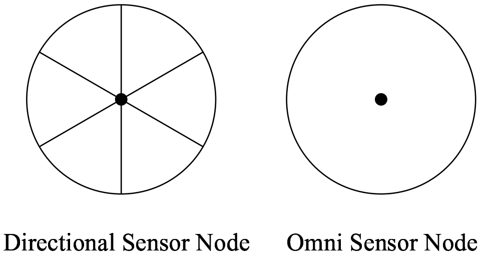


Figure 2. Directional Sensor Node.

As in [5], there have been various proposals for introduction of directional wireless nodes in wireless multihop networks; however, most of them are for efficient broadcasting (flooding), i.e., for transmission of a data message from a source node to all the other nodes in the wireless network. This is because flooding is one of the most important mechanisms in wireless multihop networks since many fundamental network services such as routing, e.g., AODV [9], DSR [3] and TORA [8], location services, directory services and so on depend on flooding of control messages. In addition, due to intrinsic nature of broadcasting in wireless networks, it is possible for wireless nodes to fail to receive the flooded control messages by collisions among them; this is called a broadcast storm problem. On the other hand, in this paper, we adopt directional nodes to sensor networks in which data messages are transmitted along wireless multihop transmission routes and are gathered to a sink node; that is, transmission direction is opposite to in the broadcast storm problem.

As discussed in [4], production, configuration and maintenance costs for directional wireless nodes are higher than the conventional omni ones. Thus, this paper assumes that only a part of wireless nodes devises directional antennas and the others are with the conventional omni antennas. Until now, vast number of ad-hoc routing protocols have been proposed and many routing protocols for wireless sensor networks have also been proposed especially with consideration of battery consumption. In this paper, our discussion is based on the following naive routing based on periodical flooding of control messages initiated by a sink node:

- On first receipt of copies of the routing control message, each intermediate sensor node entries the sender sensor node as its next-hop for transmission of sensor data message destined to a sink node and also broadcasts it[1].

By using this routing protocol, each directional sensor node has the following properties:

[Property 1] Among its 6 directional antennas, each directional sensor node adopts only 1 antenna for transmissions of sensor data messages destined to the sink node and the antenna is never used for receipt of sensor data messages.

[Property 2] Omni nodes whose next-hop is the same directional node and included in the wireless transmission range of the same directional antenna are always exposed with each other [2].

According to these properties, previous-hop nodes of a directional node is not required to apply RTS/CTS since there are no hidden nodes about it. In addition, in accordance with Property 1, no collisions occur at previous-hop sensor nodes of a directional node, i.e., 2-hop previous-hop sensor node of a directional node does not need RTS/CTS for collision-free transmissions of data messages if no pairs of the 2-hop previous-hop sensor nodes are hidden nodes

[1] In a directional node, a copy of the routing control message is simultaneously transmitted through all its directional antennas.

[2] Because of the $\pi/3$ central angle in a sector wireless range of directional antennas.

with one another since data messages transmitted by the directional node never reach its previous-hop nodes. Even if there exist 2-hop previous nodes exposed to each other, collisions among them are avoided only by CSMA/CA without RTS/CTS.

Therefore, by introduction of directional nodes, wireless sensor networks reduces RTS/CTS according to the following property:

[Property 3] The following sensor nodes omits RTS/CTS without collisions by introduction of a directional node (Figure 3):

- Previous-hop nodes.
- 2-hop previous-hop nodes in case that all of them are exposed.

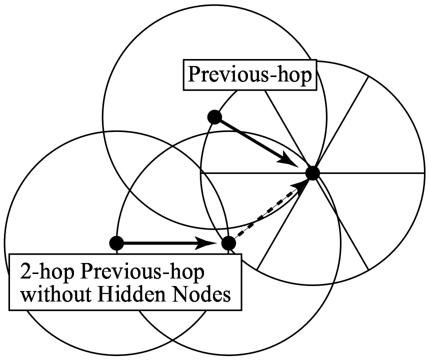


Figure 3. RTS/CTS-Free Transmissions by Directional Node.

4 Evalution

This section reports the results of simulation experiments to reveal the effect of our proposal. As discussed in the previous section, introduction of directional nodes justifies omissions of RTS/CTS absolutely in previous-hop sensor nodes of the directional nodes and conditionally in 2-hop previous-hop ones. The degree of reduction of RTS/CTS depends on ratio of directional nodes in a wireless sensor network and tradeoff between performance improvement, i.e., reduction of transmission delay of sensor data messages, and costs for directional nodes should be considered in reality. Here, the relation between the directional node ratio and the degree of reduction of RTS/CTS is evaluated in simulation experiments.

As shown in Figure 4 200–700 wireless sensor nodes with a 100m transmission range are randomly distributed in a 2,000m×2,000m square area in which a sink node is located at its center. Among of them, randomly selected 0–25% nodes are directional nodes and the others are omni ones. By using the naive routing protocol discussed in the previous section, wireless transmission routes from every sensor node to the sink node are configured and evaluate the ratio of sensor nodes allowed to transmit data messages without RTS/CTS according to Property 3.

Figure 5 shows the simulation results. In cases with directional nodes, RTS/CTS-free nodes linearly increase and the effect is higher in sparser distribution of nodes. This is because there exists more hidden node free 2-hop neighbor nodes of directional nodes in sparse wireless networks. As discussed above, for introduction of directional nodes, additional costs should be considered in reality.

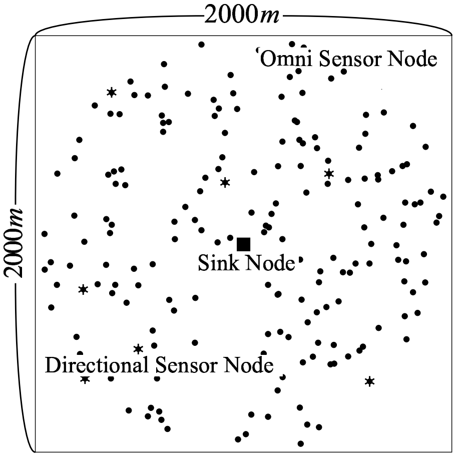


Figure 4. Simulation Field.

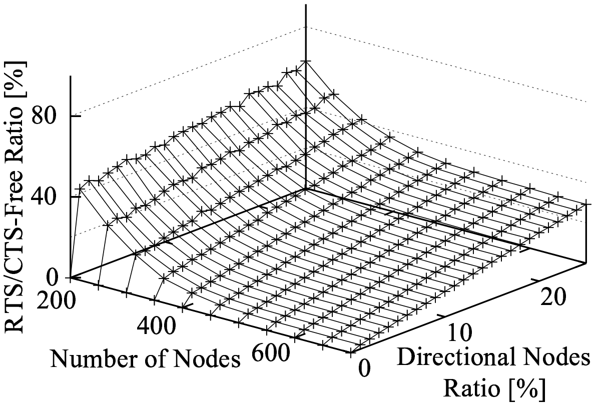


Figure 5. Ratio of RTS/CTS-Free Sensor Nodes.

5 Concluding Remarks

For shorter transmission delay in wireless sensor networks, this paper proposes an introduction of sensor nodes with directional antennas. Due to independence of each directional antennas, less pairs of 2-hop neighbor nodes have the hidden node relations and RTS/CTS for collision avoidance with high communication overhead is omitted absolutely in previous-hop nodes and conditionally in 2-hop previous-hop nodes of directional nodes. Simulation experiments show the effect on reduction of RTS/CTS requirements and it is expected to reduce end-to-end transmission delay of sensor data messages. We have assumed that the directional nodes are randomly located and the routing protocol works independently of the location of the directional nodes. In future work, we investigate a routing protocol for shorter transmission delay based on the randomly located directional nodes and induce better locations of directional nodes in intentional network configuration.

References

[1] Akimoto, M., Shigeyasu, T. and Morinaga, N., "A New Method for Mitigating Transmission Overhead by Adaptive RTS/CTS on the Basis of Existence of Hidden Terminal," Proceedings of the 18th IPSJ DPS Workshop, pp. 137–142 (2010).

[2] Culler, D.E. and Hong, W., "Wireless Sensor Networks," Communications of the ACM, Vol. 47, No. 6, pp. 30–33 (2004).

[3] David, B., David, A. and Hu, Y.C., "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)," RFC 4728 (2007).

[4] Furukawa, T., Bandai, M., Yomo, H., Obana, S. and Watanabe, T., "Multi-Lobe Multicast Using Directional Antenna for Network Coding in Multi-Rate Ad Hoc Networks," Proceeding of the 14th IPSJ DICOMO Symposium, pp. 164–170 (2010).

[5] Kathiravan, K., Thamaral, S. and Reshm, R., "Efficient Broadcasting in MANETs Using Directional Antennas," Ubiquitous Computing and Communication Journal, Vol. 2, No. 2 (2007).

[6] Matsumura, S. and Higaki, H., "Extension of RH2SWL for Collision-Free Data Message Transmissions by Subsidiary Channel in Wide-Area Wireless Multihop Networks," Proceedings of the 11th IEEE International Wireless Communications and Networking Conference, CD-ROM (2010).

[7] Numata, Y. and Higaki, H., "Power Controlled Routing in Wireless Multihop Communication for Higher End-to-End Bandwidth," Proceedings of the 3rd IEEE International Conference on Wireless and Optical Communications Networks, CD-ROM (2006).

[8] Park, V. and Corson, M., "Temporally-Ordered Routing Algorithm (TORA) Version 1 - Functional Specification," Internet Draft, MANET Working Group, draft-ietf-manet-tora-spec-04.txt (2001).

[9] Perkins, C.E. and Royer, E.M., "Ad hoc On-Demand Distance Vector (AODV) Routing," RFC 3561 (2003).

[10] Seada, K., Helmy, A. and Govindan, R., "Modeling and Analyzing the Correctness of Geographic Face Routing under Realistic Conditions," Ad Hoc Networks, Elsevier, vol. 5, pp. 855–871 (2007).

[11] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Standard IEEE 802.11 (1997).

A Cognitive Approach to Link Optimization Utilized in Wireless Sensor Networks

Lun Zhang
School of Transport Engineering
Tongji University
Shanghai, China
Lun_zhang@tongji.edu.cn

Jia Mei Wang
School of Transport Engineering
Tongji University
Shanghai, China.
Jiamei.wang@tongji.edu.cn

Bi Sheng Fang,
School of Transport Engineering
Tongji University
Shanghai, China
Bisheng.fang@tongji.edu.cn

Abstract—**A Game theoretical algorithm ,which is titled as GTL(Game-theoretical Total Link) , is addressed in this paper and applied in the optimization of LINK issues in Wireless Sensor Networks. Through choosing an appropriate payoff value, all nodes are assumed to play games and to do benefit to both itself and all the neighbors. Eventually, all nodes performance in the status of optimized energy-efficiency under the condition that all nodes are in the coverage and able to communicate with each other . Hereafter, experiments shows that GTL makes great improvements by 20% compared to classical CTR algorithm in the energy consumption.**

Keywords-Wireless Sensor Networks, Topology Control, Game Theory

I. INTRODUCTION

As for the issues of *Topology Control* for Wireless Sensor Networks, some algorithms such as CBTC (Cone Based Topology Control)[1][2] are widely utilized to keep the connectivity and coverage by controlling the nodes' radiating radius. However, most of the algorithms consider network connectivity rather than energy saving, subsequently inducing the low energy-efficiency of the nodes. Meanwhile, there emerges various algorithms of Topology Control to meet the demand of mobile networks, resting networks and other uncertain networks instead of the motionless network[3][4][5]. Thus, a cognitive algorithm based on game theory is incorporated in this paper to optimize nodes' radiating and transmitting range, so as to save energy of nodes in the premise of no influence to integrity of the network.

In most cases, a minimal transmitting radius r is calculated to meet the demand of network connectivity, and consequently to use this value as CTR(Critical Transmitting Range) [6][7] to ensure all nodes in communication. Nevertheless, a great deal of energy are wasted and redundancy owing to the equal radiating radius. Although, this way ensures the network connectivity, as well as improves the network robustness in most cases, avoiding the great energy waste is still a problem. Thereupon, a kind of dynamic algorithm for the radius set is addressed in this paper.

II. MODAL

CTR means that there is a unique radiating radius r for each node to keep the network working as a whole, by which all nodes are able to communicate with each other. Normally, r is the longest edge of the LMST(Local Minimum Spanning Tree) in the network.

Compared to CTR, GTL(Game-theoretical Total Link) is able to save more energy without loss of connectivity and robustness. By modeling a payoff function, all nodes choose various radiating radius to change the coverage area periodically and intelligently. accordingly, the Game Theory is incorporated to design this cognitive system[8]-[12]. For details, by calculating the payoff value of each node and the whole network, each node chooses the optimal energy controlling strategy. The main compositions of this game are described as follows:

- The set of players, denoted as N . In this game, nodes are players, which are denoted as i , and $i \in N$;
- The set of strategy, denoted as S . In this game, every player has three strategies, which are increasing the energy level, decreasing the energy level and maintaining the energy level.
- The set of payoff by each player's choice, denoted as P ,which means the payoff of every player.

All nodes make decision through choosing their energy strategies, and change their energy levels until game finishes.

A. PAYOFF VALUE CONFIGURATION

- Connectivity

$$Connectivity = \frac{N_c}{N} \tag{1}$$

Where, N_c indicates the amount of the nodes in the biggest cluster in the network, N means the quantity of the nodes in the whole network. If all nodes are in communication, the connectivity is 1.

- Neighbor Coverage Angle

Neighbor Coverage Angle(NCR) indicates the maximum angle between two radials which originate from the reference node. In figure 1, α is the NCR.

- Neighbors Expectation

In the premise of keeping the NCR of the reference node unchanged , Neighbors Expectation is defined as the minimal amount of the neighbors as the energy consumption of the node decreases. Weak robustness would be caused by the fact that number of the neighbors excesses to the lower bound of Neighbors Expectation.

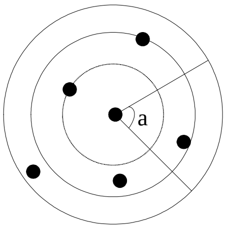


Figure 1. Neighbor Coverage Angle

B. Assumption

- The initial value of all nodes in communication is 1.
- Each reference node gets benefit in accordance with the energy consumption, quantitatively describes as $-\alpha \times r$, in which, r is the radiating radius of nodes with a certain energy level. α is weight parameter.
- Reference node get benefit $\beta \times (n-k)$ from its neighbors. Here, n is the amount of neighbors, k is the neighbor expectation , β is a small valued weight parameter with
- Payoff of the whole network equals to the sum of payoff of all nodes.
- Thus, the payoff of a node in current energy level is:

$$Payoff = 1 - \alpha \times r + \beta(n - k)$$

III. SIMULATION

A. Algorithmic flow

Figure 2 shows the flowing chart of the arithmetic simulation. Through repeated games, the whole network will eventually get a Nash Equilibrium, which is the optimal solution.

B. Simulation Result

Here, both GTL and CTR algorithms are simulated within the background of a random generated Wireless Sensor Networks,. In Figure.3, 50 nodes are deployed randomly. Each node is displayed as being surrounded by a circle with certain grey scale ,which denotes the radiating radius . Accordingly, the grey scales will superimpose when the energy circles overlap.

With CTR algorithm, energy levels of all nodes are equally set by experience. In Figure 3, deeply darken area means more coverage coincidence ,which means more energy waste correspondingly. Thus , the S/N ratio (Signal to Noise ratio) will decrease when more coverage circle overlap. Therefore, CTR algorithm performances great waste of energy due to the coverage overlapping, although it would assure the connectivity and robustness of the network.

It would be worse if nodes are deployed densely. As being shown in Figure3, there're hardly any nodes in the areas with the deepest grey scale. Nevertheless, the nodes in the brim always possess the same energy levels with others in order to meet the demand of CTR, so as to result in senseless energy consumption. Thereupon, these overlapping areas should be reduced because they can't achieve extra payoff.

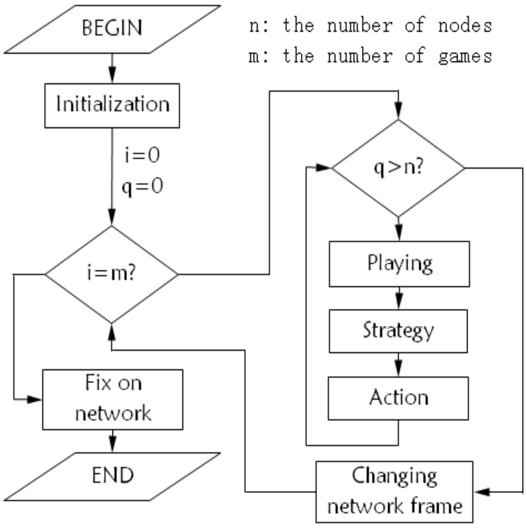


Figure 2. Flow Chart of the Simulation

Note: playing means that each node calculates the payoff values of various strategies;Strategy means the node chooses the strategy with maximal payoff value;action means that nodes change their radiating radius by controlling the energy levels .

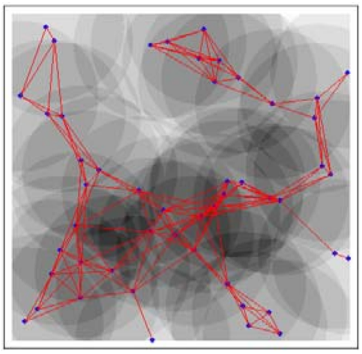


Figure 3. Simulation Result Of CTR

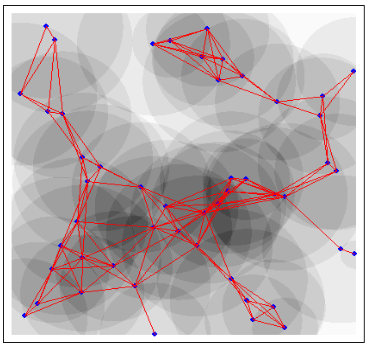


Figure 4. The Final Network Frame Of GTL

Figure 4 shows the result of GTL algorithm after t playing the game .

In Figure4, for the purpose of energy saving, most nodes decrease their energy level after repeated game playing on the premise of that all nodes keep in communication. The energy levels of nodes in the brim get decreased simultaneously. Meanwhile, Neighbor Expectation of each node is also under control.

The energy consummation of CTR and GTL are shown in Table.1 respectively.

TABLE I. THE COMPARISON OF CTR AND GTL

Algorithm	energy summation	optimized percentage
CTR	1148.3449	17.75%
GTL	944.5095	

The Comparison of the optimization result between CTR and GTL in the networks with nodes ranging from 5 to 50are presented in Figure.4 and Figure.5.Hence, it is obtained that the improvement is around 20%. However, the percentage trends to be stable at 20% with the increasing amount of nodes, which is described as:

$$\lim_{N \to \infty} E\left(\frac{\sum_{N}^{CTR} re - \sum_{N}^{GTL} re}{\sum_{N}^{CTR} re}\right) = \varphi$$

Where, re represents the range of Energy Consumption. In most cases, $\varphi = 0.2$.

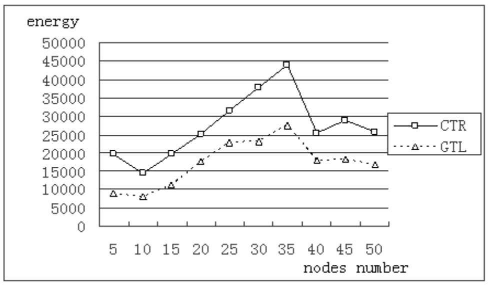


Figure 5. The comparison of CTR and GTL

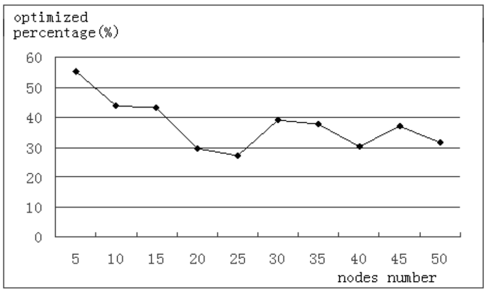


Figure 6. The Optimized Percentage Of GTL

C. With Resting Strategy

In fact, topology structures of most networks are dynamic with cycled sleeping strategy or compose of mobile nodes. Hereby,10 nodes are randomly chosen for sleeping under the random rest timing strategy. Figure 7 shows the simulation result. After nodes play game with GTL algorism , the topology turns into the structure which is shown as in Figure 8.

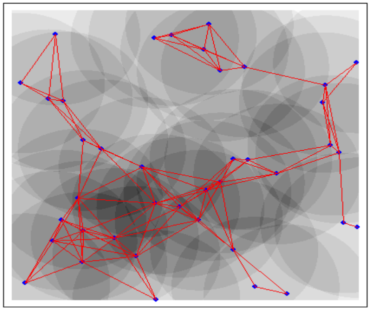


Figure 7. Network Frame of CTR in resting networks

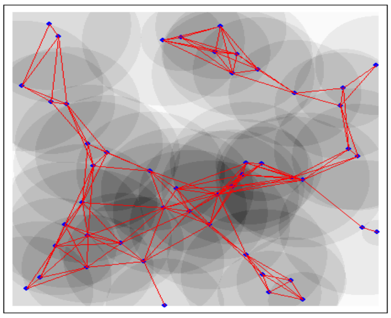


Figure 8. The Network Frame After GTL

In Figure.7 , the sleeping of 10 nodes' do not change the network connectivity and thus did not make any effect on the network integrity. After 10 rounds of games with GTL for random network which selects ten nodes randomly for sleeping, the result is as Table.2.

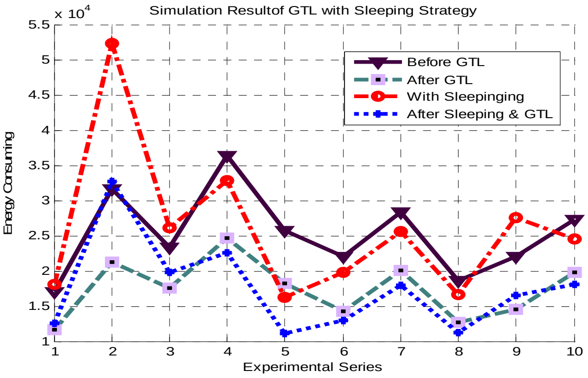


Figure 9. Simulation Result of Energy Saving

In Figure 9 and Figure 10, it shows that GTL algorithm leads to more energy saving in sparse networks than dense networks. For example, the whole energy consumption improves from 31664.42 to 52285.5225 due to thesleeping of 10 nodes in experiment 2.

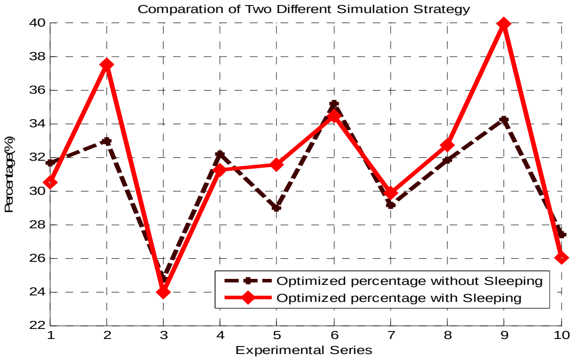


Figure 10. Energy Saving of Ten Experimental Series

TABLE II. THE EXPERIMENT OF GTL ABOUT NETWORKS WITH CYCLED SLEEPING STRATEGY

Ex. No.	Before GTL	After GTL	optimized percentage1
1	17044.405	11651.3375	31.64
2	31664.42	21216.6953	33
3	23352.02	17572.2923	24.75
4	36440.08	24721.4618	32.16
5	25727.585	18272.1837	28.98
6	22053.25	14299.6961	35.16
7	28373.54	20110.4402	29.12
8	18562.405	12653.0409	31.84
9	22102.825	14537.9718	34.23
10	27238.69	19774.1245	27.4
Average	25255.922	17480.92441	30.83
	After resting	After resting & GTL	optimized percentage2
1	18045.9045	12545.1229	30.48
2	52285.5225	32689.4971	37.48
3	26073.801	19824.6042	23.97
4	32796.072	22559.3658	31.21
5	16220.637	11106.5965	31.53
6	19847.925	13009.0599	34.46
7	25536.186	17907.2914	29.87
8	16706.1645	11242.9792	32.7
9	27529.38	16532.7508	39.95
10	24514.821	18128.386	26.05
Average	25955.64135	17554.56538	31.77

In fact, in the periodical sleeping strategy , sleeping nodes transfer more work to other living nodes, which could properly make extra energy waste. Consequently, GTL might be used to balance the energy oscillation

IV. CONCLUSION

CTR algorithm does assure the robustness and the connectivity of the network ,but with relatively high energy consumption. The entire energy consumption of the whole network increases due to the unique pre-set of the radiating radius for all the nodes. However, algorithm of Cognitive Intelligence is incorporated in this paper to make the self-organize more efficiently.

The algorithm of GTL uses Game Theory to set the energy range of each node thus to made nodes control their energy consumption flexibly according to the topological changing. Via the experiments, it obtains that at least 20% energy would e saved compared to the classical algorithm of CTR.

ACKNOWLEDGMENT

The work of this paper is supported by National Science Foundation of China.(Project No.: 50408034) and Shanghai Educational Foundation for Innovation(Project No.: 11ZZ27).

REFERENCES

[1] Li L,Bahl P,Wang Y,and Wattenhofer R, Analysis of a Cone-based Distributed Topology Control Algorith, for Wireless Multi-hop Networks, *Proc.ACM PODC01*,Newport,RI, 2001,pp.264-273

[2] Wattenhofer R, Li L,Bahl P and Wang Y, Distributed Topology Control for Power Efficient Operation in Multihop Wireless Ad hoc Networks. *Proc. IEEE Infocom 01*, Anchorage,Alsaka,2001,pp.1388-1397

[3] Miguel A. Labrador , Pedro M. Wightman ,*Topology Control in Wireless Sensor Networks: with a companion simulation tool for teaching and research*, Springer, ,March 1, 2009

[4] R. Rajaraman ,Topology Control and Routing in Ad hoc Networks: A Survey. *SIGACT News*,33:60-73, June 2002.

[5] I. F. Akyildiz and I. H. Kasimoglu, Wireless sensor and actor networks: research challenges, *AdHoc Networks*, vol. 2, no. 4, pp. 351–367, October 2004.

[6] Paolo Santi The Critical Transmitting Range for Connectivity in Mobile Ad Hoc Networks , *IEEE Transactions on Mobile Computing* ,Volume 4 , Issue 3 ,May 2005.

[7] Penrose M. The Longest Edge of The Random Minimal Spanning Tree. *The analysis of Applied Probability*,1997, 7(2),pp340-361

[8] J. Neel, R. M. Buehrer, J. H. Reed, and R. E Gilles. Game theoretic analysis of a network of cognitive radios. *Proceedings of the 45th Midwest Symposium on Circuits and Systems*,vol. 3, August 2002, pp. 409-412.

[9] J. E. Hicks, A. B. MacKenzie, J. A. Neel, and J. H. Reed. A game theory perspective on interference avoidance. *Proceedings of the Global Telecommunications Conference (Globecom)*, vol. 1, 2004, pp. 257-261.

[10] J. H. Reed, and R. R Gilles. Using game theory to analyze wireless ad hoc networks.*IEEE Communications Surveys & Tutorials*, Volume: 7, Issue: 4,, Fourth Quarter 2005,: 46- 56

[11] L.Ronnie M.Johansson, Ning Xiong and Henrik I.Christensen. A Game Theoretic Model for Management of Mobile Sensors. *Proceedings of the Sixth International Conference of Information Fusion*, 2003,Vol.1,: 583- 590

[12] P. Michiardi, R. Molva, A Game Theoretical Approach to Evaluate Cooperation Enforcement Mechanisms in Mobile Ad Hoc Networks, WiOpt 2003, INRIA Sophia-Antipolis, France, March 3-5, 2003.

[13] Rahul Garg, Abhinav Kamra, and Varun Khurana. A game-theoretic approach towards congestion control in communication networks. *Computer Communications Review*, 32(3):47-61, June 2002 .

Assessment Approaches to Wireless Sensor Networks Utilized in Urban Mass Transit

Lun Zhang
School of Transport Engineering
Tongji University
Shanghai, China
Lun_zhang@tongji.edu.cn

Bi Sheng Fang,
School of Transport Engineering
Tongji University
Shanghai, China
Bisheng.fang@tongji.edu.cn

Jia Mei Wang
School of Transport Engineering
Tongji University
Shanghai, China.
Jiamei.wang@tongji.edu.cn

Abstract—**In the realistic background of Urban Mass Transit, the paper addresses the applications of Wireless Sensor Networks with emphasis on the integrated evaluation approaches to the applications. Assessment strategies are originally designed, being involved in the indicators and criteria of fundamental functions, adaptability, applicability, Quality of Service and Advancement. Hereafter, evaluation methods are taken out on subjective and objective respectively, with which Rough Set Theory, Fuzzy Set, and Markov Chain are used for quantification and standardization.**

Keywords-Wireless Sensor networks;Assessment Strategy; Rough Set; Markov Process;Urban Mass Transit.

I. Introduction

All manuscripts must be in English. These guidelines inc The wireless sensor network(WSN) is hot and cutting-edge research field which is attracted concern mostly, involving cross-disciplinary and high integrated knowledge. It combines sensors, embedded computing technology, modern networking and wireless communication technology, distributed information processing and so on. Physical world, computer world and humanity society the three-mode world can be communicated by the information which is real-time monitored, perceived and collected by means of various integrated micro-sensors, and transmitted in wireless way, then send to the user terminals through wireless multi-hop ad hoc networks mode. It has widely applied prospect in the military, environment science, healthcare, space exploring, industry sensor, security monitor and traffic control et al since the wireless sensor network has low cost, self-organizing, small size and can be disposed easily[1]-[5].

The railway transportation is a high reliable complex system; besides it is also a huge scale, refining profession division system, and it requires coordination in many aspects. The intelligent WSN applied into urban rail transit has irreplaceable advantages. Not only the WSN can improve nodes function and be disposed more easily than the traditional sensor networks based on field-bus, but also it can decrease the cost compared with the traditional field-bus technology.

But we find that the WSN has some morphological characteristics in energy limitation of node, large quantity and dense layout of the nodes, random disposal and micro-size nodes et al which raise a great challenge in software and hardware design of the network and nodes. The WSN shows strong dynamicity, ductility, generality and compatibility et al, and it can transport a large amount of all kinds of information, but the stability and reliability of whole network is impacted by the strict resource limitation, asymmetry of flow, redundancy of data, dynamic of network et al which results in the feature of network such as network heterogeneous, vulnerability and polymorphism of data to affect it applied in the practice.

The railway transportation system is developed based on the “train”, “electric machine”, ”construction”, “electricity” technology. The information network of railway guarantees the reliability, safety, effectiveness of system and service, and furthermore, it raises higher requirement of the collection, process, transfer and fusion of data. The compatibility and scalability is required between data processing system and other system as network management system, information transfer system, environment monitor system, along with rapid expansion of railway network. Daily increasing passengers, service improvement demand, technology development and railway developing trend-intelligent, automation, foundation, standardization call for highly unified function system and secure, reliable, efficient, scalable network system.

Currently, the most application of WSN in railway is building the urban rail transit self-organization network monitor and control system. According to the function it can be simply divided into: FAS (Fire Alarm System), TSIS (Travelling Security Inspect System), SCADA (Supervisor Control And Data Acquisition), BAS (Building Automation System), which distribute along the railway, station and shunting yard, and then related information is send to control center via all-kinds of communication network. The WSN must be specified by particular standard to meet the system demand since the WSN is integrated into current system.

The evaluation index of sensor network is not only the standard of assessing sensor network, but also the optimal object of the sensor network design. Standardizing the technology creates the benefits of applying the technology into monitor the urban rail transit, and resolves the problems in wireless communication, train location, status inspection, real-time monitor and so on. Subsequently it also plays a scientific evaluation and guide role in widely applying the senor network with the development process in the urban rail transit.

II. 2. EVALUATION CRITERIA

A. Performance Evaluation of WSN

It can definitely improve the reliability by integrating the sensor network into the urban rail transit monitor and guarantee the coordination operation of system for safe and efficient operation of urban rail transit system. The performance evaluation refers to following three aspects on account of above situation:

(1)Achievement of performance evaluation—requires completing different function in different applied environment such as: monitor in serious disaster; monitor of situation of electricity equipment; circumstance monitor; train operation status monitor, the evaluation index is classified into 2 types:

One is data index of estimation error of property, probability of event monitor, false detecting rate or missing rate of object, accuracy of location, accuracy and error of perception.

The other is index of network extendability, integrity, stability.

Advanced network technologies are applied to system architecture design and develop of railway monitor which set up integration monitor system by relative monitor subsystem. They can improve process efficiency and data accuracy, realize real-time monitor, and reduce risk, ratio of accident caused by delay and error data, since subsystems carry out collection and analysis of raw data.

(2)Communication performance evaluation—is communication ability among nodes; requires transferring more effectiveness information within less time; underlines cooperation operation among nodes; achieves real-time monitor and process in rail transit monitor system. The evaluation indices are real-time perception ability, communication efficiency, data quality, error tolerance ability, time latency et al.

The more reliability and timeliness of data transmission to ensure safety in train operation is critical in monitor system which demands more efficiency and quality, less time delay and transmission error code in system nodes. It is essential to safeguard higher error-tolerance ability, and the whole network can work as an entirety even though in the condition of emergent events, because one node or more could communicate failure easily in the network, and sometimes transmission link interrupts in the WSN operation.

(3)Energy consumption control evaluation—achieves the most communication and function within the least nodes. And the evaluation indices are lifecycle of network and nodes.

The narrowest bottleneck in the WSN is energy consumption in application. The independent node’s power system is supported by limitation cell stored in it. By means of technology, We can improve utilization and balance the network lifecycle and nodes lifecycle, try to cut down the energy consumption of node and equalize energy consumption of network to extend the lifecycle to come to Pareto optimal under the condition that limited by the system demand and itself condition and each node has to work in the condition of finite cell.

B. Adaptability evaluation

Adaptability has two types: one is system adaptability; another is the environment adaptability.

System architecture design and develop of railway monitor use advanced network technologies, set up integration monitor system by relative monitor subsystem and combine sensor technology, embedded computing technology, modern network technology, wireless communication technology and distributed information processing technology et al. These technologies infusion need higher adaptability (or compatibility). There are two types in system compatibility: compatibility between current rail transit monitor and management system, and compatibility between current rail transit monitor and new technology. It must to consider that how to control effect of manpower, resource and finance caused by system rebuild to least level or direct WSN to a positive orientation.

It presents higher request to the technology (network protocol, network architecture, and route and so on), which can support the topology of network, environment of channel and service mode while those dynamically vary with nodes move, keep uninterrupted communication and low error rate, and be adjust to different condition and status’s variation by WSN set up by self-organization form.

C. Practicality evaluation

There are great problems need to be resolved in the process of developing WSN into practice, such as practice applied route, protocol, energy consumption. Practicality evaluation indices in this paper include: service demand, economy, operability, reliability, efficiency et al.

Fire alarm system (FAS) flood monitoring sensors, fire detection sensors, hazardous gas monitoring sensor, rock fall monitoring sensors and seismic monitoring sensors; electrical equipment condition monitoring sensors of SCADA; Environment Monitoring System (BAS) and the vehicle running state (axle temperature, pressure, track, vehicle positioning, etc.) monitoring system, these subsystems’ functions can be implemented by various sensors and processing system installed along the route, environment *and* the station.

D. QOS (Quality of Service) evaluation

The QOS is defined as “The quality promise of information transmission and share between the network and users, and among the communication users in the network— to satisfy users’ demand, the sensor network is necessary to supply enough resources to achieve a certain performance”.

The WSN has the following features as special self-organization network. (Six features correspond to demands of QOS in table 1)

TABLE I. QOS REQUIREMENTS BASED ON THE NETWORK FEATURE

Feature	Requirement of QOS
Strict resources limitation	Mechanism should be simply enough to avoid large-calculated and energy- consumed algorithm
Asymmetry of flow	Certain mechanism should be considered in design.
Redundancy of data	Data redundancy and time should be balanced combined certain request.
Dynamic of network	Increase difficulty of ensuring QOS and need to be resolved by technology and algorithm.
Balance between energy and time delay	Balance between energy and time delay to guarantee the QOS in real-time service
Support variety service	Supply different routes correspond to different data flow

There are three application types “drive based on event, drive based on query, sustained uninterrupted transmission application” in WSN application according to the data transmission mode. The WSN could optimize maximally for the system application and function demanded by service as application-specific network, but also supported by corresponding requirement of QOS.

The QOS guaranteed by the WSN is no more the conception end-to-end of tradition network, and bandwidth, packet loss rate is no longer the main objective concerned by single sensor node as well, but essential service quality of transmission data which emphasizes mission-critical when sudden events happen in a set of sensor nodes. So information timeliness, packet transmission reliability, efficiency of data, and multi-dimensional external environmental factors are considered to be the measures of QOS.

E. Advance of technology evaluation

Advance of technology evaluation is category of value evaluation, and is a comprehensive evaluation of new technology and its application in new domain.

The WSN is a wireless multi-hop ad hoc network of none infrastructure in advance, self-organization and reconfiguration. The WSN is new information collection system within enough wireless senor nodes distributed in local area. WSN technology is whether to reduce construction cost to a great extend, combine each rail transit monitor system into a correlation, resource share, and coordinated entirety, develop the most transportation ability of urban rail transit and satisfy the society demand, which is imperative to take into account in the research and construction period.

III. 3. EVALUATION STRATEGY

There two types of evaluation methods in the light of the evaluation content above: objective evaluation and subjective evaluation as figure 1:

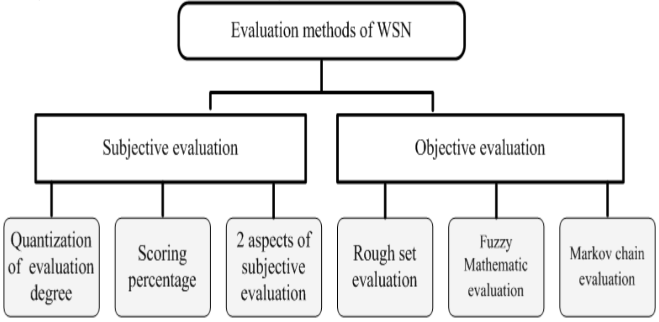

Evaluation methods of WSN
Subjective evaluation
Objective evaluation
Quantization of evaluation degree
Scoring percentage
2 aspects of subjective evaluation
Rough set evaluation
Fuzzy Mathematic evaluation
Markov chain evaluation

Figure 1. Classifications of methods of assessing WSN

A. Subjective evaluation

The subjective evaluation means to evaluate with experts’ knowledge, experience and judgment. Generally speaking, the ratio of subjective evaluation should not excess 1/3.

The method of subjective evaluation is mainly about:

(1) Quantize evaluation indices into “excellent”, “good”, “medium” and “bad” four degrees, each degree represents certain scores. The corresponding scores are calculated by computer system or agency based on evaluation degree given by experts of assessment which is according to their judgment. (As table 2)

TABLE II. EVALUATION TABLE TYPE 1

Performance evaluation	excellent□	good□	medium□	bad□
Adaptability evaluation	excellent□	good□	medium□	bad□
Practicality evaluation	excellent□	good□	medium□	bad□
……				

(2) Score in centesimal system. The corresponding scores are calculated by computer system or agency based on scores given by experts of assessment which is according to their judgment.

(3) List all the evaluation content which is evaluated by experts to tick acceptable index. (As table 3)

TABLE III. EVALUATION TABLE TYPE 2

good general performance	√	Good performance actuality	√
		Good communication performance	
		Good energy consumption control	√
Good adaptability			
Good practicality		√	
……			

B. Objective evaluation [6][7][8]

(1)Rough set evaluation

a. establish evaluation hierarchy (as table 4)

TABLE IV. HIERARCHY OF THE NETWORK EVALUATION

First Class	Second Class Index	Detailed Contents
Function Evaluation	Energy Consumption Evaluation	Energy Efficiency
		Lifecycle
	Transmission Functional Evaluation	Time Delay
		Apperceive Precision
		Expansion Capability
		The Power Of Fault-Tolerant
	Functional Achievement Evaluation	Robustness
		Reliability
Adaptability Evaluation	Systematic Adaptability	Integrity
		MTBF/MTTF
	Environmental Adaptability	Diversity of Topology
		Diversity of Sensing
		Diversity of Transmitting Path
		Diversity of interference
Applicability Evaluation	Demand For Services	Coverage
		Links and Routing
		Others …
	Economic Advisability	Cost
		Time Efficiency
		Spatial Efficiency
	Operability	Installation
		Configuration
		Deployment
	Reliability	Fault Tolerant
		Security
	Efficiency	Time Efficiency
		Spatial Efficiency
		Energy Efficiency
Qos Evaluation	Real Time Features	Quickly Response
		Quickly Self-Recovery
	Transmission Reliability	Information Security
		Information Completeness
	Transmission Efficiency	Rates
		Capacity
		Others…
	Multidimensional External Environmental Factors	Temperature Limitation
		Moisture Limitation
		Electromagnetic Restriction
	Other	Others…
Advancement Of Technology Evaluation	Society Value	Extensibility of New Tech.
		Adaption Of New Technology
	Economic Value	Cost/Effective
		Others…
	Technology Value	Stability of New Tech.
		Sustainability of New Tech.
		Feasibility of New Tech.
		Others…

b. fix weight with rough set

Weight can be fixed by the concept of property importance in the rough set. Procedure as following:

① From the lowest level of Index, KRS (Knowledge Representation System) is established for its parent target. The condition property set C consists of all sub-indexes, and the parent targets are the decision attributes D. Suppose $C = \{a_1, a_2, \cdots a_i, \cdots a_n\}$

② Deal the KRS with numerical processing, and remove duplicates.

③ Obtain the $POS_{C-ai}(D)$ of condition properties

④ Calculate the importance of condition attributes: $P = r_C(D) - r_{C-ai}(D)$

⑤ Process normalization, namely define $P_A = \sum_{i=1}^{n} P_i$, and then calculate $W_i^0 = P_i / P_A$ which is the index weight of sub-index a on its parent.

c. Calculate comprehensive weight

The comprehensive weight of indices in all levels can be calculated start from the previous level and top-down, after obtaining weight of each index on their higher level index weight separately. Calculated by the formula as (1)

$$W_j = \sum_{j=1}^{m} a_j b_{ij} \quad (1)$$

a_j : Weight of first level index relative to the evaluation objective.

b_{ij} : Weight of second level index relative to the first level index

d. Fix membership of each index

We analyze the real feature of each index, fix membership function of each index, substitute the parameter value X_i and standard value of assessment objective into the membership function, and calculate the membership $\mu_A(X_i)$.

e. Comprehensive evaluation

Linear weighting method can be used to calculate a comprehensive evaluation of each index.

$$T = \sum_{i=1}^{n} W_i \mu_A(X_i) \quad (2)$$

(2) Fuzzy Mathematic evaluation

Fuzzy comprehensive evaluation is affected by many factors and makes a comprehensive assessment as a very effective method of multi-factor decision-making. It characterized by the evaluation results that are not absolutely positive or negative, but in a fuzzy set to represent.

Specific process: the evaluation of objective factors are considered as the composition of the fuzzy set (called the factor set U), then we set the assessment levels of these selected factors, composite the fuzzy sets of assessment (called evaluation set V). We calculate membership of every single factor of all assessment level (referred to as fuzzy matrix), distribute rest on weight of each factor in evaluation objective, and the quantitative evaluation of solution value obtained by computing (called fuzzy matrix synthesis).

In accordance with the fuzzy analysis, we evaluate the WSN network.

① Suppose factors set U : $U = \{u_1, u_2, \cdots u_9\}$

We pick U_1 (WSN performance evaluation), U_2 (WSN adaptability evaluation), U_3 (WSN practicality evaluation), U_4 (WSN QOS evaluation), U_5 (WSN advance of technology evaluation) to be the major indices reflect WSN performance.

② Define assessment set $V : V = \{v_1, v_2, v_3, v_4\}$

We set v_1 : excellent, v_2 : good, v_3 : medium, v_4 : bad.

③ We select network of experts familiar with WSN to constitute evaluation group to get evaluation matrix.

④ According to expert opinion, we determine the weight set A

⑤ In accordance with the evaluation model and the principle of maximum degree, we make a final evaluation.

(3) Markov chain evaluation

Comprehensive forecast of events not only is able to point out the various possible outcomes of events, but also give the probability of each outcome. Markov chain method is a prediction of the probability of event, which is based on the Markov chain.

Markov chain evaluation method is to separate all indicators of a WSN network into the following classification: excellent, good, medium, qualified, and bad. And then the initial status is defined as status before improved to show as following:

$$A(0) = (A_1(0), A_2(0), A_3(0), A_4(0), A_5(0))$$

After optimizing the WSN, in the every testing period, the ratio between the total of each indices and total quantity of evaluation serve as status variable (formula 3):

$$A(t) = (A_1(t), A_2(t), A_3(t), A_4(t), A_5(t)) \quad (3)$$

t represents the No. t period in the formula, $t \in N$, furthermore: $\sum_{i=1}^{5} A_i(t) = 1$

It can be obtain from Chapman-Андрей Николаевич Колмогоров formula:

$$A(t+1) = A(t)P \quad (4)$$

P is step transition matrix:

$$P = \begin{Bmatrix} a_{11} & \cdots & a_{15} \\ \vdots & \ddots & \vdots \\ a_{51} & \cdots & a_{55} \end{Bmatrix}$$

Obviously, $P_{ij} > 0, (i, j = 1,2,3,4,5), \sum_{i=1}^{5} P_{ij} = 1$. When $k \to \infty$, $A(k) = A(k+1) = A$. Limit state of $A(k)$ is probability distribution under a stable state homogeneous Markov. In view of the above, the stable vector can be deduced. Solution vector can be resolved from equations $AP = A$ or $(I - P)A = 0$ which is the final result of assessment.

IV. Conclusion

The objective of WSN is to realize calculation at anywhere which has huge potential use value and the prospect of widely using. However, because of the immaturity of wireless sensor network technology and the special nature of their self-organization, it is highly requested on the improvement of node, network protocols, routing, hardware and software. So, the objective evaluation of a network needs to put analysis and research into the practice.

It is the first time to design the framework of evaluation content and research the evaluation strategy in the WSN field, and evaluate the performance of WSN and adaptability in urban rail transit by integrating subjective and objective evaluation. With the further research, the evaluation paradigm requires more supplements and specification to be adapt to technology application and special problems of WSN. Ultimate goal is to search for more targeted application measures through the results of evaluation.

Acknowledgment

The work of this paper is supported by National Science Foundation of China.(Project No.: 50408034) and Shanghai Educational Foundation for Innovation(Project No.: 11ZZ27).

Reference

[1] Akyildiz F, Su W, Sankarasubramaniam Y, Cayirci E. Wireless sensor network: A survey. Computer Networks, 2002,38(4): 393−422.

[2] Romer K. Mattern F. The design space of wireless sensor networks. IEEE Wireless Communications, 2004,11(6):54−61.

[3] Estrin D, Govindan R, Heidemann J, Kumar S. Next century challenges: Scalable coordination in sensor networks. In: Proc. of the ACM/IEEE Int'l Conf. on Mobile Computing and Networking. New York: ACM Press, 1999. 263−270.

[4] Ren FY, Huang HN, Lin C. Wireless sensor networks. Journal of Software, 2003,14(7):1282−1290

[5] Li JZ, Li JB, Shi SF. Concepts, issues and advance of sensor networks and data management of sensor networks. Journal of Software, 2003,14(10):1717−1727

[6] Lun Zhang; Dongxiu Ou, Assessment Strategy with Markov Chain Utilized in Wireless Sensor Networks, Performance, Computing and Communications Conference, 2008. IPCCC 2008. IEEE International , 7-9 Dec. 2008 :418 - 421 ;

[7] Lun Zhang; Yan Lu; Lan Chen, Fuzzy Evaluation for Wireless Sensor Networks Based on Rough Set Theory, Performance, Computing and Communications Conference, 2008. IPCCC 2008. IEEE International , 7-9 Dec. 2008 :406 - 411

[8] Lun Zhang, et al, Evaluating Strategy with Grey Theory Utilized in Wireless Sensor Networks, London, U.K., July, 2008, The World Congress on Engineering 2008;

Deployment of Sparse Sensor-Actuator Network in a Virtual Architecture

Sébastien Faye, Jean Frédéric Myoupo
Université de Picardie-Jules Verne, UFR Sciences, 33 rue Saint Leu, 80039 Amiens France
{sebastien.faye@etud.u-picardie.fr, jean-frederic.myoupo@u-picardie.fr}

Abstract: The use of a virtual architecture in wireless sensor networks provides a powerful and fast partitioning into a set of clusters, each being locatable in space by the sink or base station (BS for short). We first propose a protocol for the detection of empty clusters due to poor distribution of sensors or in low-density network. After this detection, we propose a strategy to allow mobile sensors (actuators) to move to position themselves on empty clusters in order to improve the routing of collected data.

Key Works: Wireless Sensor-Actuator Networks,
Virtual Architecture.

1. INTRODUCTION

The wireless sensor networks (WSN) are from the family of mobile ad-hoc (MANET), but have additional features and constraints: typically, they consist of a wide range of sensors with limited energy capacity. Each sensor is powered from a battery non-rechargeable and non-replaceable ([2]) and has a low capacity in terms of memory, calculation (CPU), and transmission range in wireless. Each sensor is able to harvest a set of data in a certain environment, and transmit it in multi-hop way to a base station (BS) or sink, which may act as instructor of the network. The use of such networks is widespread in many applications: for example we can cite the monitoring of forests, critical infrastructure, or the detection of biochemical agents and in the military industries. Some examples of work can be found through [1], [2], [5], [10].

Technology related to sensors advancing day by day, it is common to see WSN composed of several thousand units to tens of thousands ([6], [12]). In large networks, the sensors can be grouped into clusters based on their proximity to propose a better management and data transmissions to be made, in order to significantly increase the scalability, economy energy, routing, and consequently the lifetime of the network. The structure provided by the use of clusters allows the use of various techniques to improve the quality of a WSN, such as data aggregation which consumes as much as the calculation procedure ([7]).

In this paper, we focus on the concept of virtual architecture developed by Wada et al. in [11]. The latter is created and orchestrated by BS, who is able to split the network into a set of clusters, depending on the strength and direction of the broadcast it can perform. We focus on the case where the network is sparse leaving many empty clusters formed in the virtual architecture, contrary in [11] where the WSN is assumed to be dense. Once detection has occurred, we introduce the notion of actuators (sensors that have the ability to move), to consider positioning strategy of actuators onto empty clusters.

The rest of the paper is organized as follows: Section 2 presents the model of our network and specifically the virtual architecture. The protocol for the detection of empty clusters is described in section 3. The strategy of feeding the empty clusters by the actuators in presented in section 4. Section 5 describes how to move the actuators onto the empty clusters. The simulation results are presented in section 6. And a conclusion ends the paper.

2. VIRTUAL ARCHITECTURE

Consider a set of anonymous sensors distributed around a master node BS. Figure 1 illustrates an example of such a network. In this WSN, each sensor has the ability to collect data, perform calculations, store data temporarily, and to transmit and receive information within a certain range R that does not cover the entire network because of energy constraints and cost. The transmission of information from one sensor to BS is therefore usually done in multi-hop. The BS has the opportunity to transmit information to some powers (to the sensor farther away).

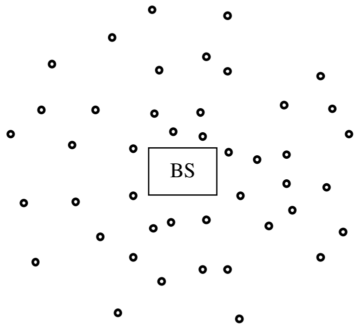


Figure 1. Example of a basic WSN

The virtual architecture proposed in [11] consists of a partition of the WSN into different zones by the BS: the BS has the opportunity to disseminate information from the lower to the higher range in order to create coronas. It has the opportunity to disseminate information in certain directions as [11] to create different angular sectors. The area (i, j) is the intersection of the corona i with the angular sector j. The sensors of the same area are therefore in the same geographical location and form a ***cluster***. This is illustrated in Figure 2. Other works on this type of architecture can be found in [3], [4] and [9].

On this basis, the authors in [11] define a number of mechanisms, such data aggregation, routing, and generally training coordinates. We will not detail these operations here. Interested readers can find them in the original article [11]. Here we focus only on sparse WSN in which some clusters can be empty. In such networks the routing is optimal, and the BS is not able to detect to detect innately these empty clusters.

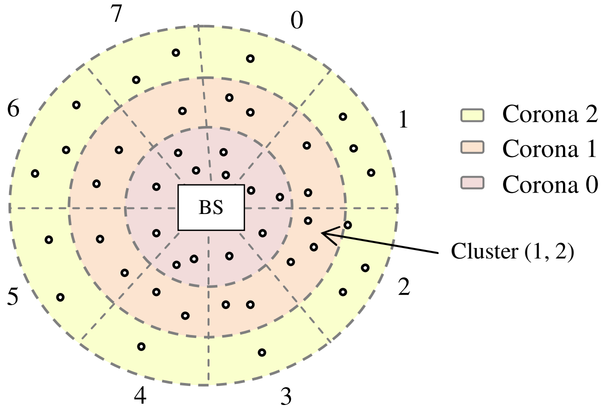


Figure 2. An virtual architecture representation.

Here we propose to equip him with a protocol to detect these empty clusters.

3. DETECTION OF EMPTY CLUSTERS

3.1 Notations

To clarify the result of our paper, we use the following notations below:

- (i,j) : cluster of corona i and angular sector j
- WSN* : the set of sensors-actuators
- ACT : the set of actuators
- s1, s2, … sn : slots time.

3.2 Assumptions

Now we assume that the WSN is clustered as in [11] as in Figure 3 below.

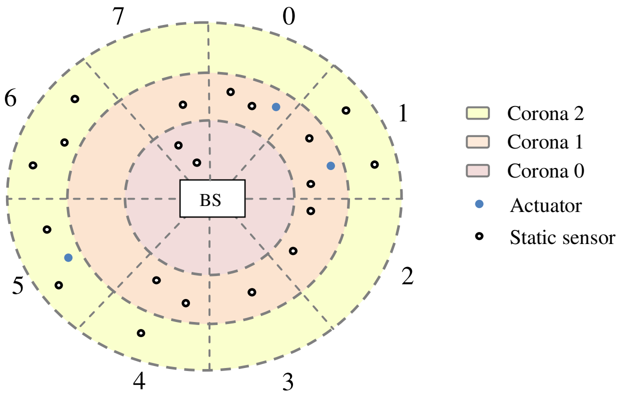


Figure 3. A sparse virtual architecture.

We consider the following assumptions:
1. We assume that the network of WSN* \ {ACT} is static, composed of anonymous immobile nodes once deployed. Optionally, adding and deleting nodes (for fault tolerance) is allowed, but considered rare. In contrast, ACT represents mobile nodes, which have an identity, and whose objective is to improve the connectivity of the network (for example).
2. The network topology is known to any sensor, but the density of nodes allows us to assume that WSN* \ {ACT} is connected, as well as all WSN*.
3. The BS characteristics: BS has the opportunity to disseminate information across the network more or less powerfully (i.e., at different ranges), and at different angular sectors (one-way broadcast).
4. Time is divided into a set of slots s1, s2, ... sn and the local clock of each sensor is synchronized with the one of the BS. In this way, each sensor is awake and listening to the network for each of these slots, but is put to sleep the remaining time to save energy overall network.
5. We consider that the BS and the actuators have a notion similar direction: it is necessary to allow BS to manage the movements of actuators.

3.3 Description of the detection protocol

***Phase1:** Initialization of the detection by the BS*

In the first slot s0 all sensors-actuators are awake. And the BS initiates the construction of the breadth search tree. For this, it broadcasts at range R_{bs} (corresponding to the first corona of the virtual architecture) a message of network discovery noted Build(zone) to all the sensors WSN*. WSN being connected, we are assured that at least one sensor receives the information. Here, Build(zone) = Build((-1, -1)), where (-1, -1) represents the area consisting solely of BS.

***Phase 2 (recursive):** construction the breadth search tree*

This phase is repeated as s0 is not exceeded, and has 4 steps:

Step 1. All actuators that had no heard any message of construction during s0, when receiving the message Build(zone) they simply return a receipt to BS, stating their position for future movements.

Step 2. When receiving the message Build(zone), all static sensors that have never received a message of construction during s0, are able to temporarily (the time slot s0) store Build.zone, corresponding to the cluster where the sensor (sender of the message) is located (father's sensor).

Step 3. When receiving this information static sensors will then perform two actions:

a. If their cluster is different from the one of the message of construction received, then send an acknowledgment backwards to BS, directly or in multi-hop through the cluster Build.zone. This message looks like ACK(zone, parent), zone representing the cluster from which the acknowledgment is sent, and parent information Build.zone registered by the sensor. In receiving this message the BS is able to complete his vision of the tree width. This step can be represented through the figure 4.

b. Continue to spread the message of tree construction by sending the message Build(zone) at a transmission range R (where R is the transmission range of a sensor, which we consider to be homogeneous).

When these first two phases are completed and slot s0 is expired, BS has an overview of the network and the tree width of the zones of virtual architecture installed on it, like the example shown in Figure 4 and figure 5.

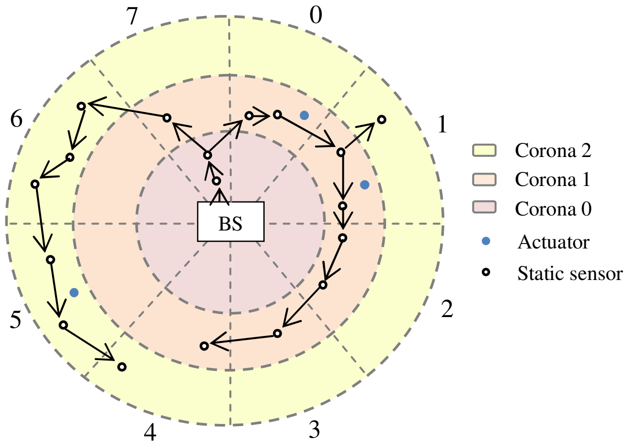


Figure 4. Construction of the breadth search tree.

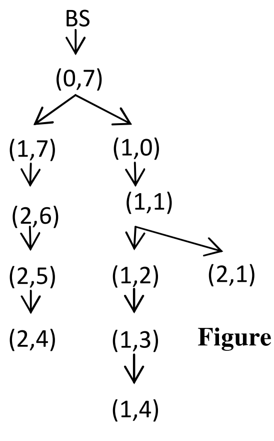


Figure 5. The breadth search tree.

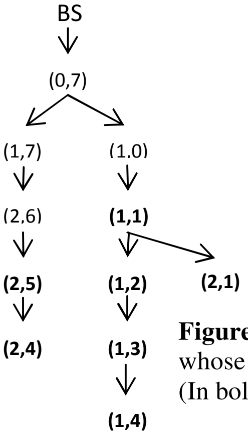
Phase 3: *Detection of empty clusters*
BS being informed of all clusters of the virtual architecture of the network (he is the designer), he is able to detect empty clusters i simply browsing the generated tree.

4. STRATEGY OF FEEDING THE EMPTY CLUSTERS BY THE ACTUATORS

The empty clusters are now known to BS. We detail a strategy that could allow BS to place actuators on these empty clusters in order to gather and route eventual data collected on these clusters. Of course, this strategy is just a sample of application, and the user may choose to apply a priority policy, by choosing for example a part of the actuators for the collection of information in critical areas (empty clusters) that are not equipped with stationary sensors. Our approach consists of two main phases: the first (detailed below) improves the tree width obtained above, in order to reduce the number of hops required for certain nodes to communicate with BS. Second, we may use the remaining actuators to improve communication in some areas with small populations.

Phase 1: Improvement of the breadth search tree
Let H be the height of the graph obtained in section 3. Consider C+1 where C is the number of coronas used by the virtual architecture (example in Figure 4: we have H = 6, and C = 4). The purpose of the first phase is to improve the value of H, reducing it until it reaches - if possible H = C - (requires a sufficient number of actuators). So here we apply a strategy of feeding the empty cluster in terms of the number of actuators available and known to BS.
At this point, two cases arise:

Step 1. If at the start we have H = C, then the application of this phase is not necessary because the routing is optimal we can directly proceed to Phase 2.



L[0] = Ø
L[1] = 1,2,3,4
L[2] = 1,4,5

Figure 6. Detection of clusters whose access can be improved (In bold on the tree)

Step 2. Otherwise, we have to determine the priority empty clusters among all the empty clusters of our virtual architecture. The priority empty clusters are those that we must first feed (by an actuator) in order to improve the height H of the tree. To do this, BS calculates from the tree the C - 1 lists: each list number i denoted L[i] contains the list of corona j ε L[i] corresponding to clusters (i, j) whose access can be improved. A cluster (i, j) to which access can be improved (by positioning correctly an actuator) is such that i+1 is strictly below the level of (i, j) in the tree. This is illustrated through Figure 6.

From this list, we can easily deduce the priority empty clusters: consider the cluster (i, j) such that j ε L[i], *the* *priority empty clusters are the empty clusters (m, j) with m<i.* This is illustrated through Figure 7.

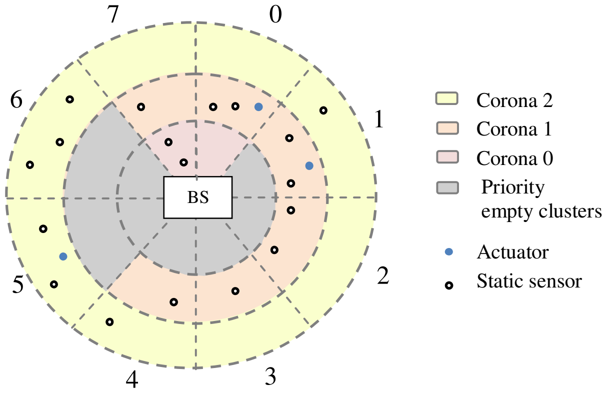


Figure 7. Detection priority empty clusters.

5. MOVING THE ACTUATORS TO THE EMPTY CLUSTERS

It is necessary to provide to an actuator in cluster (i, j), all information that will help it to move correctly to the empty cluster (i2, j2). BS has an overview of the network and knows the parameters used by the virtual architecture. One of its roles is then to calculate and transmit the movements to be performed by each actuator, using simple formulas from trigonometry. Below, we detail the various calculations that BS must do to move an actuator from (i, j) to (i2, j2).

We assume that BS knows the angle α of an angular sector and the scope β of a corona. Let A be a virtual point that we consider to be the center of the cluster (i, j). Let B be a second virtual point considered the center of (i2, j2).

5.1 Computation of the distance Df between A and B:

- Distance from BS to A : $|AD| = \beta \times i + \frac{\beta}{2}$

- Distance from BS to B : $|BD| = \beta \times i2 + \frac{\beta}{2}$

- Let [BS,0) be the half straight line starting from BS and following the left end of the angular 0. The angle formed by [BS,0) and [BS,A] is : $A.R = \alpha \times j + \frac{\alpha}{2}$

In the similar way [BS,B] : $B.R = \alpha \times j2 + \frac{\alpha}{2}$

- Therefore the angle between [BS,A] and [BS,B] is :
R1 = MAX(A.R, B.R) – MIN(A.R, B.R).

- The distance from A to B is given by the following formula:

$$Df = \sqrt{A.D^2 + B.D^2 - 2A.D \times B.D \times \cos(R1)}$$

5.2 Computation of the angle to send A in the direction of B:

- Let [A, 0) the half-line parallel to [BS, 0) but starting from A (assumption (5)). Here we look for the angle between [A, 0) and [A, B], denoted by Rf.

- Here consider only the horizontal axis: Let [0, A] be the segment of length DA from A to the line [BS, 0), perpendicular to [BS, 0). Let [0, B] be a similar segment of length DB, in the case of point B.

$$DA = A.D \times \sin(\Delta A)$$

$$DB = B.D \times \sin(\Delta B)$$

Dx is the angle at x of the rectangular triangle thus formed (formed by BS, x, and the point to the segment [BS, 0]), taking the following values:

```
        If x.R < 90 Then Δx = x.R
        Else If x.R < 180 Then Δx = 180 - x.R
        Else If x.R < 270 Then Δx = x.R - 180
        Else Δx = 360 - x.R
```

- Let us consider now the rectangular triangle of hypotenuse [A, B]. Set Dtmp the length of its adjacent side having the end point A as:

```
# If an point is on the left side of BS and another on its right side
If (A.R > 180 and B.R < 180) or (A.R < 180 and B.R > 180) then Dtmp = DA + DB
# If both A and B are on the same side de BS
Else Dtmp = MAX(DA,DB) – MIN(DA,DB)
```

- We can then easily calculate the angle between the adjacent side of [A, B], say R2.

$$R2 = \cos(\frac{Dtmp}{Df})$$

- Rf is then determined by adjusting R2 depending on the position of A relative to B. Several cases are possible. These cases are summarized algorithmically as follows:

```
If (A.R < 180 and B.R > 180) Then # A on the left BS and B on the right
        If (A.R > 360 – B.R) Then Rf = 270 + R2 # A below and on the right of the line of B
        Else Rf = 180 + R2
Else (B.R < 180 and A.R > 180) Then # A on the right of the line BS, B on its left
        If (B.R > 360 – A.R) Then Rf = 90 + R2 # A above and on the left of B
        Else Rf = 90 - R2
Else (B.R < 180 and A.R < 180) Then # A and B on the left side of BS
        If (A.D > B.D) Then # A on the left of B
                If (A.R > B.R) Then Rf = 90 R2 # A is above B
                Else Rf = 90 – R2
        Else # A on the right of B
                If (B.R > A.R) Then Rf = 270 + R2 # A is below B
                Else Rf = 180 + R2
        Else # A and B are on the right side of BS
                If (A.D < B.D) Then # A on the left of B
                        If (A.R < B.R) Then Rf = 90 + R2 # A is above B
                        Else Rf = 90 – R2
                Else # A is on the right of B
                        If (B.R < A.R) Then Rf = 270 + R2 # Ais below B
                        Else Rf = 180 + R2
```

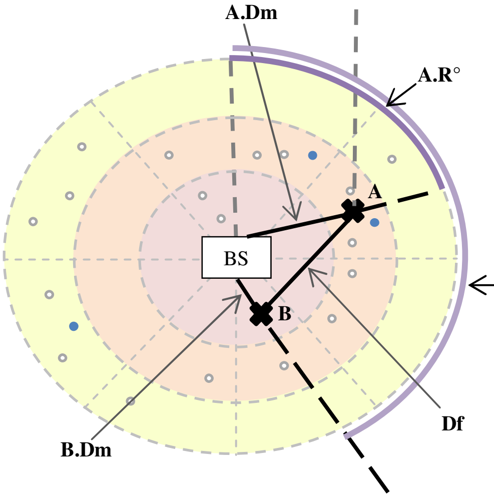


Figure 8.a. Illustration of variables and calculations used for the placement of actuators: calculation the distance between A and B.

These calculations can be illustrated by the example in Figure 8a and 8b, where an actuator placed in (1,1) must go to (0.3): BS sends an instruction to the actuator located at (i, j) to move on a distance Df with an angle Rf° to the north common with BS (assumption (5)).

Example: For α = 45° and β = 30m : A.R = 67.5, A.D = 45, B.R = 157.5, B.D = 15, DA ~= 41.5, DB ~= 5.74, Dtmp ~= 35.8, R2 ~=40°. This yields Df ~=47m and Rf ~=230°.

The actuator then moves on 47m following a path of 230° to the direction north that it shares with BS (assume that the start is the angular sector 0).

Once the actuator placed it sends an acknowledgment to BS containing the area where it is located. BS checks that everything coincides and regularly asks for the actuator's status and position, in order to guide it or replace it if something goes wrong. Conversely, the actuator can send alert to BS messages in case of problems.

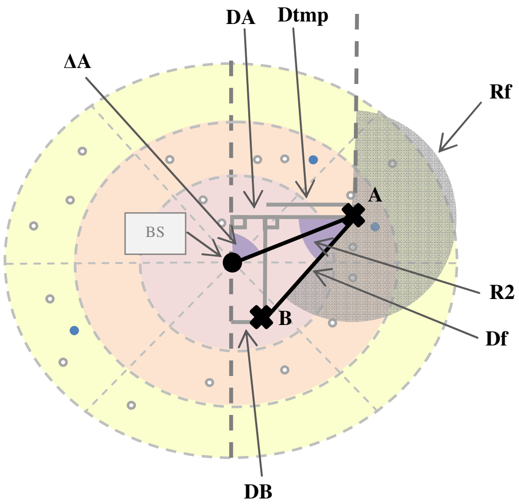


Figure 8.b. Illustration of variables and calculations used for the placement of actuators: calculation of the angle.

6. SIMULATION RESULTS

Here we focus on percentages of empty clusters that may be present in the virtual architecture described above. Our simulations were performed using the software WSNet [13] and is based on circular area of 1km, 4km and 9km in diameter, on which we randomly generated connected network of 500, 1000, 1500, 2000 and 2500 sensors. We take here an ideal case, where BS is located at the center of this circular area. Figure 9 illustrates the simulations obtained from a virtual architecture with 6 coronas and an angular sector of 30 °. Figure 10 illustrates the simulations results obtained from a virtual architecture with 4 coronas and an angular sector of 45 °. Each point of the various curves is the average result of 10 simulations

As we can see on the curves below there are empty some clusters on virtual architecture considered especially in low-density network. If we take the example in Figure 9 where the angular sectors are 30 ° and with 6 coronas, we see almost 80% of empty clusters in the case where 500 sensors are spread over a circular area of 9km in diameter. However, this percentage slows down to 60% in figure 10. So there is a need to manage the existence of empty clusters: the detection and the implementation of strategies with potential sensors and actuators. Furthermore, note that the simulations were performed under ideal circumstances: the area of sensors is circular, the BS is at the center, and we do not take into account environmental constraints. It appears evident that these points are rarely met (the BS can be at the network edge, rectangular area of sensors, ...).

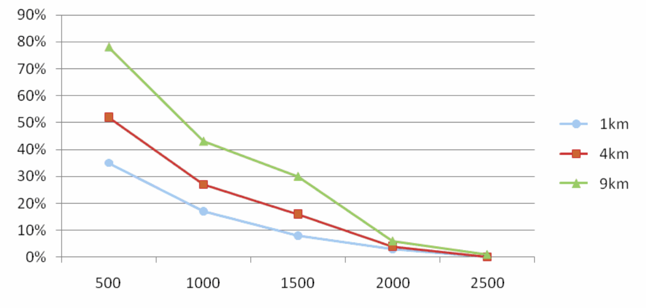


Figure 9. Average percentage of the number of empty clusters based on the number of sensors, for 3 zones of different sizes. Here, we have 6 coronas and angular sectors of 30° each.

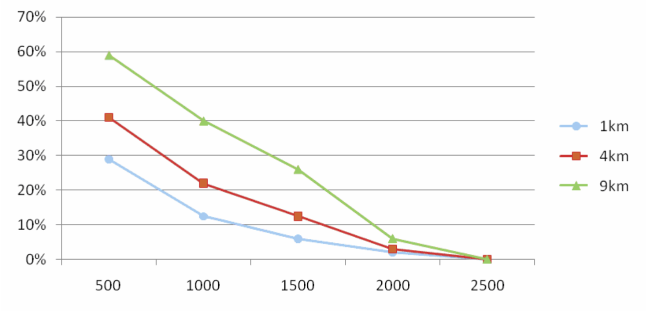


Figure 10. Average percentage of the number of empty clusters based on the number of sensors, for 3 zones of different sizes. Here, we have 4 coronas and angular sectors of 45° each.

7. CONCLUSION

In this paper we have presented a protocol to identify the empty clusters in a virtual architecture of a WSN. Next we have studied few strategies and methods to use actuators, and move on these empty clusters. We have shown how actuators can be guided by the BS to position themselves on desired empty clusters. However we could have also assumed that each actuator is equipped with GPS that should facilitate the management of the movements. Similarly we could have assumed that each actuation is equipped with APS [8] protocol.

In future work, it would be interesting to study such problems with more strong conditions: for example by removing the assumption of connectedness and reducing the density advantage of the network, we could explore the possibility of using the actuator nodes to allow some connected components to reach BS.

REFERENCES

[1] J. Agre and L. Clare, An integrated architecture for cooperative sensing networks, *IEEE Computer* 33(5) (2000) 106-108.

[2] I. F. Akyildiz, W. Su and Y. Sankarasubramaniam. Wireless sensor networks: a survey, *Computer Networks* (38), pp. 393-422, 2002.

[3] F. Barsi, A.A. Bertossi, F. Betti Sorbelli, R. Ciotti, S. Olariu and M. C. Pinotti Asynchronous Training in Wireless Sensor Networks. *Proceedings of the 3rd international conference on Algorithmic aspects of wireless sensor networks,* pp. 46-57, 2007

[4] A.A. Bertossi, S. Olariu, and M.C. Pinotti. Efficient Training of Sensor Networks. *ALGOSENSORS*, pp. 1-12, 2006

[5] C. Intanagonwiwat, R. Govindan and D. Estrin, Directed diffusion: A scalable and robust communication paradigm for sensor networks, in: *Proc. MOBICOM'00*, Boston, MA (August 2000).

[6] J.M Kahn, R.H Katz and K.S.J. Pister, Mobile networking for Smart Dust, in: *Proc. MOBICOM'99*, Seattle, WA, August 17-19 (1999).

[7] C. Karlof, N. Sastry and D. Wagner. TinySec: A Link Layer Security Architecture for Wireless Sensor Networks. *SenSys04*, November 35, 2004.

[8] D. Niculescu and B. Nath, "Ad hoc positioning system (APS),_ *Proceedings of IEEE GLOBECOM*, pp. 2926-2931, San Antonio,TX, November 2001.

[9] S. Olariu, A. Wada, L. Wilson, and M. Eltoweissy. Wireless Sensor Networks: Leveraging the Virtual Infrastructure. IEEE Network, vol. 18,pp. 51-56, 2004

[10] C.-C. Shen, C. Srisathapornphat and C. Jaikaeo, Sensor information networking architecture and applications, *IEEE Personal Communications* (October 2000) 16-27.

[11] A. Wadaa , S. Olariu , L. Wilson , M. Eltoweissy , K. Jones, Training a wireless sensor network, *Mobile Networks and Applications*, v.10 n.1-2, p.151-168, 2005.

[12] B. Warneke, M. Last, B. Leibowitz and K. Pister, SmartDust: communicating with a cubic-millimeter computer, *IEEE Computer* 34(1) (2001) 44-51.

[13] http://wsnet.gforge.inria.fr

SESSION

SYSTEM ANALYSIS

Chair(s)

TBA

Design and Implementation of ZigBee based Vibration Monitoring and Analysis for Electrical Machines

Suratsavadee K. Korkua[1] *Student Member, IEEE*
Wei-Jen Lee[1] *Fellow, IEEE*
Chiman Kwan[2] *Member, IEEE*

1. Energy Systems Research Center, The University of Texas at Arlington, Arlington, TX 76019

2. Signal Processing, Inc. 13619 Valley Oak Circle, Rockville, MD 20850

Abstract — *This paper presents a method to monitor and analyze the vibration of induction machine due to the rotor imbalance. A novel health monitoring system of electric machine based on wireless sensor network (ZigBee™) is developed in this paper. The communication protocol and software design for both wireless sensor network node and base station are also presented in detail. Moreover, the positioning scheme in ZigBee wireless network is also investigated. Based on the receiving strength signal indicator (RSSI), we can determine the distance of the sensed node by applying the distance-based positioning method. Experimental results of the proposed severity detection technique under different levels of rotor imbalance conditions are discussed and show the feasibility of this method for on-line vibrating monitoring system.*

Keywords: ZigBee, wireless sensor network, RSSI, induction machines, health monitoring system, vibration detection

1 Introduction

Predictive maintenance by condition based monitoring of electrical machine is a scientific approach that becomes new strategy for maintenance management [1]-[3]. Traditionally, monitoring system is realized in wired systems formed by communication cables and various types of sensors. The cost of installation and maintenance these cables and sensors are more expensive than the cost of the sensors themselves. To overcome the restrictions of wired networks, using wireless system for monitoring is proposed. ZigBee is a new wireless networking technology with low power, low cost and short time-delay characteristics. Based on ZigBee network communication technology, the system can deal with the various operating parameters of the remote transmission, real-time data collection, and real-time health monitoring system [4]-[5]. Moreover, ZigBee wireless technology enables to access the location of each node under the network with several types of positioning algorithms. Anyway, this study is based on the distance of the location algorithm.

Most recent research has investigated the electrical machines fault detection technique primarily based on the Motor Current Signature Analysis (MCSA) [6]-[7] with various DSP and signal processing techniques [8]-[9]. Since the rotor imbalance is mechanical related problem, the possibility of an analysis of machine vibrations is obvious [10]-[13].

This paper proposes and develops a ZigBee based wireless sensor network for health monitoring of induction motors. By observing the RSSI value and applying the distance-based positioning method, we can estimate the distance of the data collector node where fault happened. The vibration signals obtained from monitoring system are then processed and analyzed with signal processing techniques. In order to predict the level of severity of rotor imbalance, the vibration detection techniques with suitably modified algorithms is used to extract information for induction machine health diagnostic.

2 Wireless sensor network system

2.1 The proposed wireless machine health monitoring system

The proposed wireless health monitoring system is shown in Fig. 1. In this proposed wireless sensor system, vibration signal from three-axis accelerometer are recorded and stored at the base station. Signal analysis is used to extract detailed information for induction machine health diagnostic.

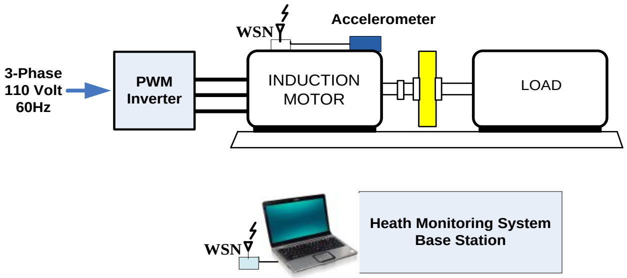


Fig. 1 Schematic diagram of the wireless machine health monitoring system

2.2 ZigBee™/IEEE802.15.4 Standard

IEEE 802.15.4 standard defines the protocol and interconnection of devices via radio communication in a personal area network (PAN). It operates in the ISM (Industrial, Science and Medical) radio bands. The purpose is to provide a standard for ultra-low cost, ultra-low power consumption and low data. ZigBee technology may be used in various applications in industrial controls, embedded sensors, medical devices, and others. Based on these features provided by IEEE 802.15.4/ZigBee, the ZigBee technology is very suitable for our application.

2.3 ZigBee based System Framework

The ZigBee system framework for data collecting system based on wireless sensor network is shown in Fig.2. It is made up of data collection nodes and PAN network coordinator. We are able to set up a network node in several nearby collection points. The nodes can carry out desired functions such as detecting the current/vibration signals, signal quantizing, simple processing, and the ZigBee/IEEE802.15.4 Standard package framing to transmit data to the PAN network coordinator. In the star topology used in this application, every device in the network can communicate only with the PAN coordinator.

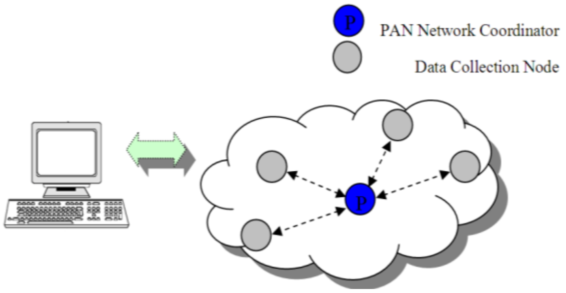


Fig.2 Structure of the wireless sensor network

3 Design of nodes and base station

The hardware framework is illustrated in Fig. 3. The main circuits include the power supply circuit, CC2430/31 external circuit, sensor and signal conditioning circuits, flash ROM and RAM memory, serial port interface, and three LEDs for status indication.

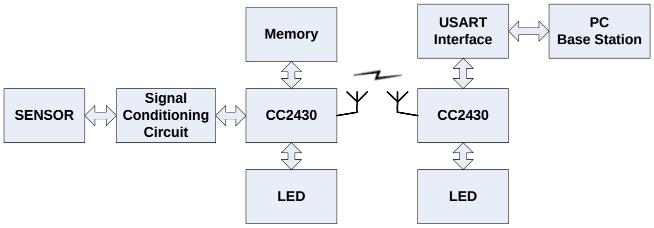


Fig.3 Hardware framework

3.1 CC2430/31 Introduction

The CC2430/31 is a true System-On-Chip [14] for ZigBee/IEEE802.15.4 solutions for 2.4GHz wireless sensor network. It combines the excellent performance of the leading CC2420 RF transceiver with an industry-standard enhanced 8051 microcontroller (MCU), with 128 KB flash memory and 8 KB RAM. Both the embedded 8051 MCU and the radio components have very low power consumption. The CC2430/31 includes 12-bit ADC with up to eight inputs and configurable resolution. There are two powerful USARTs that support serial protocols. Combined with the ZigBee protocol stack from TI, the CC2430 is one of the most competitive ZigBee solutions among industry.

3.2 ADXL330 MEMS Accelerometer

The ADXL330 [15] is a complete 3-axis acceleration measurement system on a single monolithic IC. The ADXL330 has a measurement range of ±3 g minimum. The block diagram is illustrated in Fig. 4. It contains a micro-machined sensor and signal conditioning circuit to implement the open loop acceleration measurement architecture. The output signals are analog voltages that are proportional to acceleration.

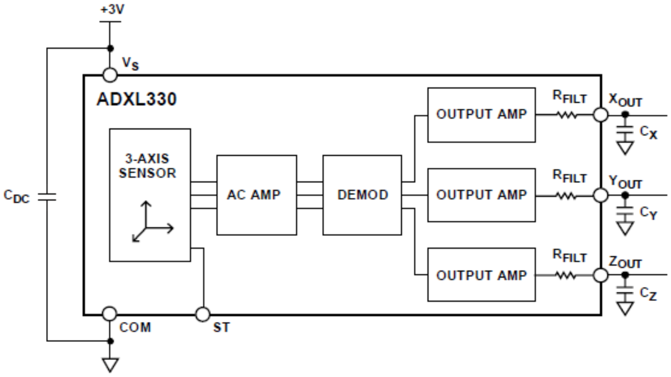


Fig.4 Block diagram of ADXL330 MEMS accelerometer

3.3 Node & Base station

The sensor nodes (shown in Fig. 5) are in charge of collecting information such as vibration signal. All vibration signals are digitized through a 12-bit ADC convert up to 2000 samples per second. Storage unit has 128kb flash memory and 8kb RAM to be chosen. Controlled by the MCU, the data from the ADC can be temporarily stored in the storage unit 8kB RAM and then transmitted to the PAN network coordinator node through the ZigBee module.

The base station includes PAN network coordination (shown in Fig. 5) and a PC. The network coordination primarily takes charge to distribute network address, and notarize the physical address, transmit test data. The network coordination can connect, and send the data and information to PC through RS232 or USB port.

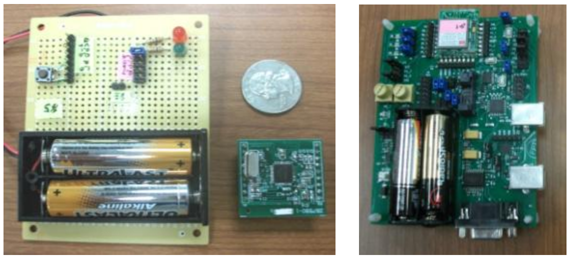


Fig.5 A data collection node board (left side) and a PAN network coordinator (right side) developed in this study

4 RSSI based distance positioning method

In order to determine the location of unknown nodes, most positioning studied in sensor networks is based on the Received Signal Strength Indication (RSSI) of the signal received from nodes with known locations. By applying a mathematical model, the position of nodes can be estimated by analyzing the RSSI signals. In this work, we use distance based localization algorithms. We use CC2431 as a wireless transceiver. ZigBee protocol stack and host-computer program can build a local area wireless positioning system. Basing on the RSSI value, these algorithms can find the distance of the transmitting node. From the test, the localization of sensor network can be conducted in a feasible and efficient way.

4.1 Received signal strength indicator (RSSI)

The received signal strength is a function of the transmitted power and the distance between the sender and the receiver. The theoretical propagation model as the equation below shows.

$$RSSI = -(10n\log_{10} d + A) \quad (1)$$

Where n is signal propagation constant which indicates the decreasing rate of signal strength in an environment, d is the distance between the transmitter and receiver, and A is the received signal strength at a distance of one meter. Given this model, we can first measure the receiving strength in different distances, and then adjust the value of n to let the model fit the actual data.

We use CC2431 as the core of the location. CC2431's location uses received signal strength indicator (RSSI). When CC2430/31 receives a packet it will automatically add an RSSI value to the received packet. After system startup, the location node first sends a certain sequence of RSSI Blast information broadcast. The node has been configured to wait for the completion of positioning at specified intervals. Then the location node sent RSSI request broadcast to the reference nodes. When the request is received, the reference node will sent back its location and the RSSI value.

4.2 Field Test Results

In order to observe the accuracy of the distance-based positioning method, we do conduct experiments at different distance of the location node varied from 0-15 meters. We performed three measurements in an open space environment. By sending 1000 packets from a transmitter, the results of RSSI at different testing distance can be shown in Table 1.

TABLE I: RSSI VALUE AT DIFFERENT DISTANCE TESTS

Range (m.)	Test #1		Test #2		Test #3	
	Packets	RSSI (dB)	Packets	RSSI (dB)	Packets	RSSI (dB)
0	1000	-39	1000	-40	1000	-37
1	1000	-47	1000	-50	1000	-53
4-5	1000	-64	1000	-64	1000	-65
9-10	999	-70	997	-71	995	-72
10-12	997	-79	1000	-76	998	-78
14-15	999	-81	999	-79	999	-85

The RSSI measurements were carried out in a real environment in order to analyze the RF propagation behavior of ZigBee modules in real working condition. To fit the experiment results with the theoretical propagation model, we choose the parameter of A=40 and n=2.5. The result in Fig.6 shows that the distance based positioning method of the proposed ZigBee wireless sensor network can provide satisfying accuracy.

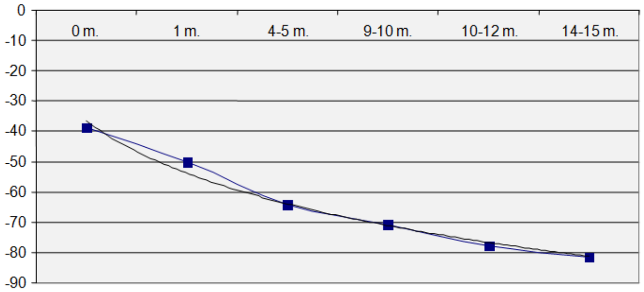


Fig.6 RSSI value (dB) versus Distance: comparison between the averaged tested data (Blue line) with the propagation model (Black line).

5 Determination of rotor imbalance severity

Rotor imbalances are common mechanical faults in induction motors. In general, a mechanical fault in the load part of the overall system can be observed from the variation of the load torque. When a mechanical fault happens, it will result in a rotating eccentricity at the rotating frequency [16]. These faults may also cause speed oscillations that have the effect on the stator current and finally lead to additional undesired harmonic components of power and torque at some particular frequencies in the spectra. From the vibration spectrum analysis, the low frequency harmonics are associated with the rotational frequency and can be distinguished from the lower range of the spectrum. The mechanical vibration due to rotor imbalance is a once per revolution force. Therefore, the harmonics of rotor imbalance can be modeled as an integer multiple of rotating frequency [16];

$$f_{vib} = k.\left[f_s\left(\frac{1-s}{p}\right) \right] \qquad (2)$$

Where f_s is supply frequency, k is the integer number, p is the number of pole pair, and s is the slip. Based on the vibration spectrum analysis, it is normally straightforward to locate the mechanical rotational frequency by monitoring the vibration spectrum and finding the most significant peak in the rotational frequency range expected. In this paper, all the techniques used here for signal analysis and processing have been implemented by MATLAB software. Block diagram of the proposed severity detection of rotor imbalance is shown in Fig. 7.

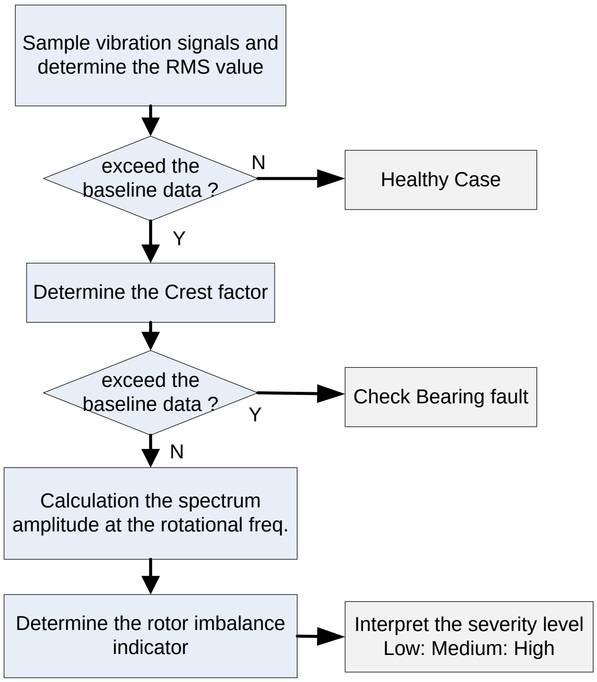

Sample vibration signals and determine the RMS value
exceed the baseline data ?
N
Healthy Case
Y
Determine the Crest factor
exceed the baseline data ?
Y
Check Bearing fault
N
Calculation the spectrum amplitude at the rotational freq.
Determine the rotor imbalance indicator
Interpret the severity level Low: Medium: High

Fig. 7 Block diagram of the proposed severity detection technique of rotor imbalance

The RMS value of the vibration signal is used for primary investigation of the machine health. The RMS values will be used to detect the severity of the abnormal condition.

The crest factor is the ratio of the peak value of the vibration signal to the RMS value. The purpose of this calculation is to give an analyst a quick idea of how much impacting is occurring in a waveform. Crest factor can be used to differentiate the bearing fault and rotor imbalance fault.

Fast Fourier Transform (FFT) algorithm is used to perform discrete Fourier transform (DFT) for all vibration signals. The amplitude of the FFT spectrum at the rotational frequency serves as the rotor imbalance indicator of the proposed monitoring system. Furthermore, the amplitude of this frequency component does reveal the severity level of the rotor imbalance fault.

7 Experimental setup and results

In order to validate the proposed vibration based health monitoring system, the test-bed for mechanical fault was set up. Rotor imbalance was created on a 3-phase, 2-pole and 1h.p squirrel cage induction motor. Induction motor is fed from the variable speed drive at 50Hz. The wireless sensor network is implemented and accelerometer is also integrated in the system for detecting the vibration signals. Vibration signals were collected by using ADXL330 tri-axial accelerometer mounted on the motor housing (in Fig. 8). Axes of acceleration sensitivity corresponding to machine vibration are shown as in Fig. 9.

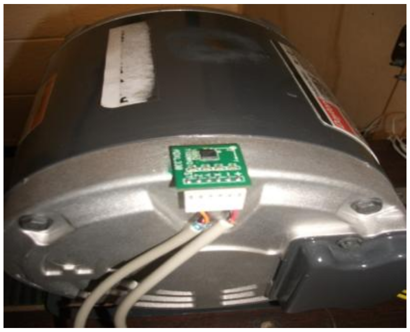


Fig. 8 A accelerometer mounted on the motor housing

The most important element of this test-bed is the flywheel which has holes drilled in it (Fig. 9.) The weights applied to these holes produce imbalance in the flywheel, thereby in the motor. The severity of the fault is determined by the weight.

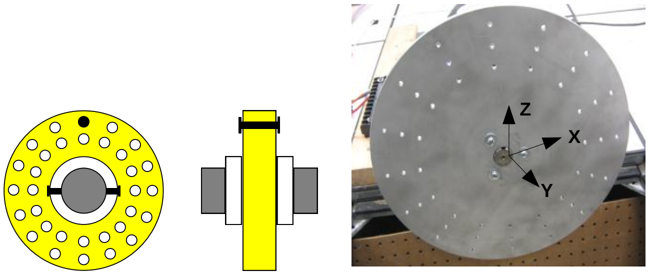


Fig. 9 Fly wheel design used in rotor imbalance test

Rotor imbalance was created using four different weights namely, 5g, 10g, 15 and 20g. No fault data was also collected as baseline data and considered as 0 g. All three-axis vibration data were calculated the RMS values. As shown in Fig. 10, the RMS values are increasing corresponding to the level of imbalance mass. However, the change in crest factor is small as represented in Table 2. It implies that the degree of impacting is relatively small from bearing faults.

TABLE II
CREST FACTOR OF THREE-AXIS VIBRATION SIGNALS FOR DIFFERENT IMBALANCE MASS TESTS

Imbalance mass (g.)	Crest factor		
	X-axis	Y-axis	Z-axis
0 g.	1.457313235	1.7216206	2.18396705
5 g.	1.451261113	1.77381552	1.82287472
10 g.	1.387397149	1.50183948	1.63720601
15 g.	1.303934686	1.40032313	1.40077586
20 g.	1.249204942	1.45098092	1.29577839

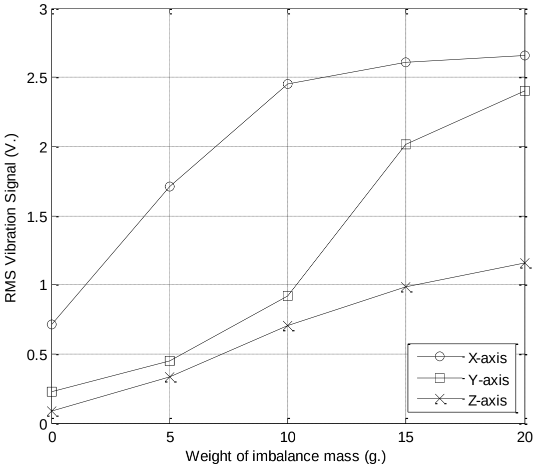


Fig. 10 RMS value of vibration signals from different imbalance mass tests

In order to observe the frequency component amplitude of vibration signals, FFT algorithm is used to perform the vibration signals. The spectrum component amplitude at rotating frequency is shown in Fig. 11. It can be noticed that the rotational harmonic at 50 Hz has a dominant value. In addition, by increasing the weight of imbalance mass, the amplitude of this frequency component will apparently increase in the spectrum. Therefore, the spectrum amplitude can be used to specify the degree of fault for the certain operating condition.

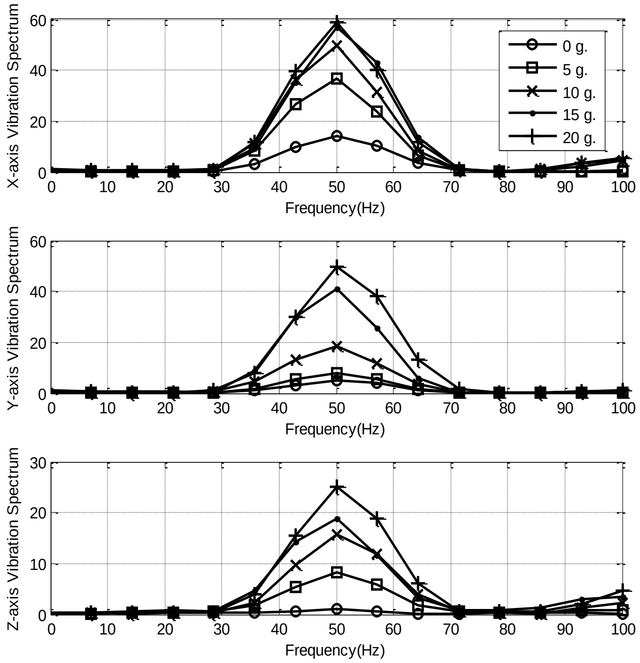


Fig. 11 3-axis vibration spectrum component amplitude at rotating frequency from different weights of imbalance mass

Furthermore, based on the analyzed vibration data and the linear approximation, the relationship between the rotor imbalance indicator and the severity level can be represented in Fig. 12. It is important to note that any estimation is subject to error. However, this relationship can be use as trend to determine the severity level of the fault. For example, the change of rotor imbalance indicator during the time interval T0-T1 can be used to predict the range of the severity level. The result from the prediction will be a very useful part of the condition monitoring system and the estimation on the usable life of the equipment.

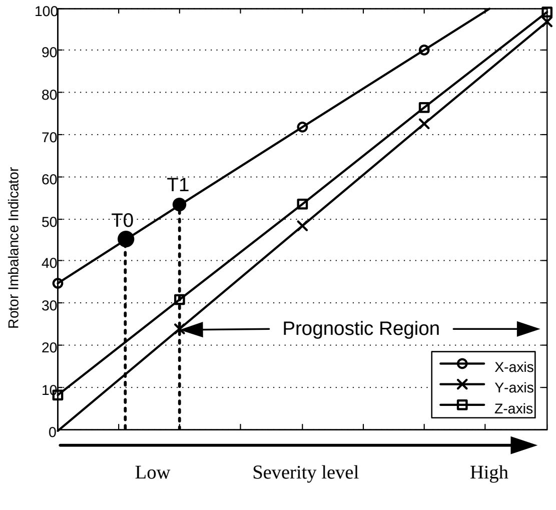


Fig. 12 Rotor imbalance indicator of different levels of imbalance severity

8 Conclusion

The method of rotor imbalance fault diagnostic through vibration analysis has been analyzed and determined on their ability to detect the induction motor abnormalities. Both hardware and software design of a ZigBee based wireless health monitoring system for induction machine is discussed in detail in this paper. By using MEMS accelerometer which is low cost, light in weight, compact in size and consume low power, this leads to the proposed vibration detection method. It is a non physical contact type which is free from electrical hazards. Moreover, it is more flexible because of tri-axial vibration measurement. Vibration signals have been analyzed and determined to detect the mechanical faults. The implementations of analysis technique in time and frequency domain are given. The proposed rotor imbalance detection technique is verified with different level of severity. Rotor imbalance indicator can be used to estimate the range of severity level which is very useful part of the predictive maintenance. RSSI of CC2431 provides the distance between the base station and monitored devices. This feature gives operator the ability to identify the location of the equipment that requires immediate attention. The proposed wireless health monitoring system is tested under various operating conditions and is found to work satisfactorily.

9 References

[1] B. Lu, T. G. Habetler, and R. G. Harley, “A survey of efficiency estimation methods of in-service induction motors with considerations of condition monitoring,” in Proc. 2005 International Electric Machine and Drive Conference (IEMDC), May 2005, pp.1365-1372.

[2] M. E. Steele, R. A. Ashen, and L.G. Knight, “An electrical method for condition monitoring of motors,” International Conf. Electrical Machines Design and Application, IEE Publication No.213, pp. 231-235, July 1982

[3] R. Natarajan, J. L. Kohler, and J. Sottile, “Condition monitoring of slip ring induction motors,” Electric power system, Vol. 15, pp. 189-195, 1988

[4] M. Gao, J. Xu, J. Tian, and F. Zhang, “ZigBee wireless mesh networks for remote monitoring system of pumping unit,” in Proc. of Intelligent Control and Automation, Jun. 25–27, 2008, pp. 5901–5905.

[5] Y.-W. Bai, and C.-H. Hung, “Remote power on/off control and current measurement for home electric outlets based on a low-power embedded board and zigbee communication,” in Proc. of Consumer Electronics, Apr. 14–16, 2008, pp. 1–4.

[6] T. G. Habetler, R. G. Harley, R. M. Tallam, S. B. Lee, R. Obaid, and J. Stack, “Complete current-based induction motor condition monitoring: stator, rotor, bearings, and load,” CIEP 2002 VIII IEEE International Power Electronics Congress, Oct. 2002.

[7] R. Obaid and T. G. Habetler, “Effect of load on detecting mechanical faults in small induction motors,” in Proc. 2003 Symposium on Diagnostics for Electric Machines, Power Electronics and Drives (SDEMPED), Aug 2003, pp.307-311.

[8] R. R. Schoen and T. G. Habetler, “Effects of time-varying loads on rotor fault detection in induction machines,” IEEE Trans. Industry Applications, Vol. 31, pp.900-906, July. 1995.

[9] S. A. Saleh, A. Kazzaz and G. K. Singh, “Experimental investigations on induction machines condition monitoring and fault diagnosis using digital signal processing techniques,” in 2003 Electrical Power Systems Research, pp.197-221.

[10] S. Rajagopalan, J. M. Aller, J. A. Restrepo, T. G. Habetler and R. G. Harley, “Analytic-wavelet-ridge-based detection of dynamic eccentricity in brushless direct current motors functioning under dynamic operating conditions,” IEEE Trans. Industrial Electronics, Vol. 54, No.3, pp.1410-1419, June. 2007.

[11] C. M. Riley, B. K. Lin, T. G. Habetler, and R. R. Schoen, “A method for sensorless on-line vibration monitoring of induction machines,” IEEE Trans. Industry Applications, Vol. 34, pp.1240-1245, Nov. 1998.

[12] Maurice L. Adamis, JR. “Rotating machinery vibration: From analysis to troubleshooting,” 2001

[13] G. S. Maruthi, and K. P. Vittal, “Electrical fault detection in three phase squairrel cage induction motor by vibration analysis using MEMS accelerometer,” in Proc. of 2005 IEEE Power Electronics and Drives Systems (PEDS), Nov. 28–30, 2005, pp. 838–843.

[14] Chipcon TI, “A True System-on-Chip solution for 2.4 GHz IEEE 802.15.4 /ZigBee® CC2430 DataSheet (Rev.2.1) http://focus.ti.com/lit/ds/symlink/ cc2430.pdf ,” 2008

[15] Analog Devices, “Small, Low Power, 3-Axis iMEMS Accelerometer ADXL330 datasheet (Rev. A),” 2006.

[16] C. Kral, T. G. Habetler, and R. G. Harley, “Detection of mechanical imbalances of induction machines without spectral analysis of time domain signal,” IEEE Trans. Industry Applications, Vol. 40, pp.1101-1106, July. 2004.

Vertical Handoff Analysis between WLAN and 3G Networks under an Adaptive QoS Control

Traore Soungalo[1], Li Renfa[2], Zeng Fanzi[3] Humphrey Njogu Waita[4]
College of Information Science and Engineering, Hunan University, Changsha, Hunan, China

Abstract - *It is with no doubt that recent trends indicate that cellular networks based on the 3G standards and wireless Local Area Networks (WLANs) coexist to offer multimedia services to end users. However, while users roam throughout this heterogeneous network, they should be automatically and transparently switched from one access technology to another such as from 802.11 networks to 3G ones, depending upon the availability of resources and QoS requirements, without any user intervention. The transfer of data between the two above named networks presents some challenges and new requirements such as an integrated architecture of WLAN-UMTS and vertical handover support adaptive designed QoS. This paper makes three major contributions. Firstly, we address an integrated architecture for WLAN and 3G networks. Secondly, we propose adaptive QoS components and lastly, this paper focuses on UMTS/WLAN vertical handover analysis under our Adaptive QoS approach. OPNET^{TM} Modeler 11.5 simulation results indicate that the adaptive multimedia framework surpasses the non adaptive approach in terms of lower handoff dropping probability and call blocking probability in performance .*

Keywords: 3G networks, WLAN-UMTS Integrated, Vertical Handover

1 Introduction

Wireless technologies are evolving toward broadband information access across multiple networking platforms, in order to provide ubiquitous availability of multimedia applications. The convergence of heterogeneous wireless access technologies has been envisioned to characterize the next generation wireless networks mostly WLAN and 3G. These two wireless access technologies have characteristics that perfectly complement each other. However, in such converged systems, the seamless and efficient handoff between different access technologies (vertical handoff) is essential and remains a challenging issue. Hence, after some previous researches and for details understanding of the interworking of WLAN and UMTS, Vertical Handoff Analysis between WLAN and 3G Networks under an Adaptive QoS Control has been selected to be the focus of our investigation.

Vertical handover or vertical handoff refers to a network node changing the type of connectivity it uses to access a supporting infrastructure, usually to support node mobility. For example, a suitably equipped laptop might be able to use both a high speed wireless LAN and a cellular technology for Internet access. Wireless LAN connections generally provide higher speeds, while cellular technologies generally provide more ubiquitous coverage. Thus the laptop user might want to use a wireless LAN connection whenever one is available, and to 'fall over' to a cellular connection when the wireless LAN is unavailable. Vertical handovers refer to the automatic fall over from one technology to another in order to maintain communication [1]. This is different from a horizontal handover between different wireless accesses points that use the same technology in that a vertical handover involves changing the data link layer technology used to access the network as shown in Figure 1.

The rest of this paper is as follow: Section 2 addresses some related works. Integrated architecture for WLAN and 3G networks is presented in section 3. We discuss our designed Adaptive QoS model in section 4 while section 5 focuses on the Vertical Handoff Analysis. Section 6 concludes our work and highlights the future work.

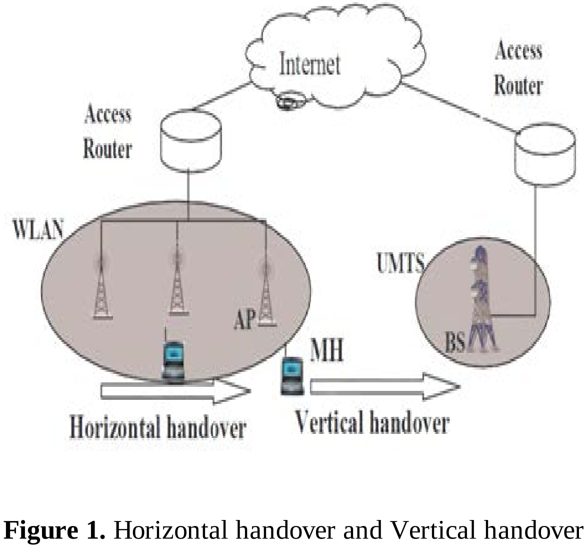


Figure 1. Horizontal handover and Vertical handover

2 Related works

Several researches on vertical Handoff have been postulated from both analytic and simulation-based studies on UMTS/WLAN systems. In this section, we focus on the most relevant research studies that fit in our work.

The design of a network architecture that efficiently integrates WLAN and cellular networks is a challenging task. To enable efficient use of scarce resources provided by the cellular networks while maintaining strong service guarantees, a generic reservation-based QoS model for the integrated cellular and WLAN networks is proposed. This model supports the delivery of adaptive real-time flows for end users while taking the advantage of high data rate WLAN systems as well as the wide coverage area of cellular networks. More specifically, the work analyzes different components of the model and their interactions. The performance of the system was realized by simulation [2].

New metrics for vertical handoff continue to emerge and the use of new metrics makes the vertical handoff decision process increasingly more complex. A generalized vertical handoff decision algorithm that seeks to optimize a combined cost function involving battery lifetime of MNs and load balancing over APs/BSs was proposed [3].

3 Integrated architecture for WLAN 3G networks

ETSI (European Telecommunication standard intitute) has defined two approaches for interworking of WLAN and cellular network; tight-coupled and loose coupled schemes. The architecture of loose coupled and tight coupled network are illustrated in Figure 2. The main difference between tight coupling and loose coupling is wheteher the user’s traffic is delivered througth the core network of UMTS or not [4].

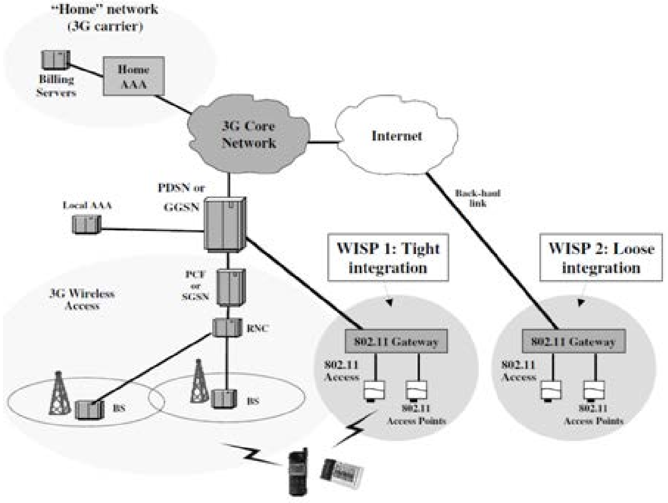


Figure 2. 3G and WLAN integration (tightly and loosely-coupled architectures)

3.1 Tightly-coupled interworking

The rationale behind the tightly-coupled approach is to make the WLAN appear to the 3G core network as another 3G access network. The WLAN network would emulate functions which are natively available in 3G radio access networks. In this architecture, utilized by WISP No.1 in Figure 2, the “802.11gateway”, (802.11 is our WLAN), network element appears to the upstream 3G core as either a PCF, in the case of a CDMA2000 core network, or as an SGSN, in the case of UMTS. The 802.11 gateway hides the details of the WLAN to the 3G core, and implements all the 3G protocols (mobility management, authentication, etc.) required in a 3G radio access network. Mobile Nodes in this approach are required to implement the corresponding 3G protocol stack on top of their standard 802.11 network cards, and switch from one physical layer to the next as needed. All the traffic generated by clients in the WLAN is injected using 3G protocols in the 3G core. The different networks would share the same authentication, signaling, transport and billing infrastructures, independently from the protocols used at the physical layer on the radio interface. However, this approach presents several disadvantages. Since the 3G core network directly exposes its interfaces to the WLAN, the same operator must own both the WLAN and the 3G parts of the network. Today’s 3G networks are being deployed using carefully engineered network-planning tools, and the capacity and configuration of each network element is calculated using mechanisms which are very much specific to the technology utilized over the air interface. By injecting the 802.11 traffic directly into the 3G core, the setup of the entire network, as well as the configuration and the design of network elements such as PDSNs and GGSNs have to be modified to sustain the increased load.

3.2 Loosely-coupled interworking

Like the previous architecture, the loosely-coupled approach calls for the introduction of a new element in the 802.11 network, the 802.11 gateway. However, in this design (WISP No.2 in Figure 2), the gateway connects to the Internet and does not have any direct link to 3G network elements such as PDSNs, GGSNs or 3G core network switches. The user population that accesses services of the 802.11 gateway may include users that have locally signed on, as well as mobile users visiting from other networks. We call this approach loosely-coupled interworking because it completely separates the data paths in WLAN and 3G networks. The high speed WLAN data traffic is never injected into the 3G core network but the end user still achieves seamless access. In this approach, different mechanisms and protocols can handle authentication, billing and mobility management in the 3G and WLAN portions of the network. However, for seamless operation to be possible, they have to interoperate. In the case of interoperation with CDMA2000, this requires that the 802.11 gateway supports

Mobile-IP functionalities to handle mobility across networks, as well as AAA services to interwork with the 3G's home network AAA servers. This would enable the 3G provider to collect the WLAN accounting records and generate a unified billing statement indicating usage and various price schemes for both (3G and WLAN) networks. At the same time, the use of compatible AAA services on the two networks would allow the 802.11 gateway to dynamically obtain per-user service policies from their Home AAA servers, and to enforce and adapt such policies to the WLAN. Since the UMTS standards do not yet include support for IETF protocols such as AAA and Mobile-IP, more adaptation is required to integrate with UMTS networks. Mobile- IP services would need to be retrofitted to the GGSNs to enable seamless mobility between WLAN and UMTS. Common subscriber databases would need to interface to Home Location Registers (HLR) for authentication and billing on the UMTS side of the network, and to AAA servers for the same operations to be performed while clients roam to WLAN networks.

There are several advantages to the loosely-coupled integration approach. First, it allows the independent deployment and traffic engineering of WLAN and 3G networks. 3G carriers can benefit from other providers' 802.11 deployment without extensive capital investments. At the same time, they can continue to deploy 3G networks using well-established engineering techniques and tools. Furthermore, while roaming agreements with many partners can result in widespread coverage, including key hot-spot areas, subscribers benefit from having just one service provider for all network access. They no longer need to establish separate accounts with providers in different regions, or covering different access technologies. Finally, unlike the tightly-coupled approach, this architecture allows a WISP to provide its own public WLAN hot-spot, interoperate through roaming agreements with public WLAN and 3G service providers, or manage a privately installed enterprise Wireless LAN. Therefore, loose-coupled network cannot support service continuity to other access network during handover, thus loose-coupled sheme has long handover latency and packet loss [5].

4 Adaptive QoS model designed

In this paper our proposed QoS framework assumes a packet switching core network based on the UMTS network architecture. However, this holds the same relationship with some GPRS networks or other packet switching cellular systems. It is well known that Internet has some fundamental scalability limitations [6] when it comes to manage individual traffic flows with the approach of resource reservation. Its successor, the prioritization approach addresses the scalability problem at the cost of coarser service granularity. To enable efficient use of scarce resources provided by the cellular networks while also maintaining strong service guarantees, we adopt the reservation based systems [7]. In WLAN the reservation is achieved by using the HCF of the WLAN and in UMTS is achieved by the functionality of Base Station.

4.1 Adaptive QoS components

The framework contains some components defined bellow:

- A Policy Provisioning Module (PPM): The PPM is responsible for mapping actual users QoS profiles with their subscription information and decides the traffic classes for the users' traffic flows. Then these QoS parameters can be handed to CAC module to process.
- Degradation Profile: It allows for negotiation of established QoS connection through the degradation profile, when the user requests to establish a QoS call, some network resources need to be admitted. The requested QoS has to be allocated when the connection is set up. If certain conditions change over the activation time, a negotiation procedure will be called. The traffic class of a connection is defined as C_i , where $C_i \in \{C_1, C_2, \dots C_i, \dots C_K\}$, where K is the number of service classes. The corresponding bandwidth requirement for each class is defined as $B_i \in \{B_1, B_2, \dots B_i \dots B_K\}$, for the sake of simplicity we assume that all the connections in the same class have the same requested bandwidth. Let $D_i \in \{D_1, D_2 \dots D_i \dots D_K\}$ denote the minimum bandwidth request defined in the connection degrade profile. Let $p^i(t)$ denote the degradation probability of class i and $n^i(t)$ the number of connections from class i at time t. Thus the degradable bandwidth at time t can be written as:

$$\sum_{i=1}^{k}(B_i - D_i)p^i\ (t)n^i\ (t) \qquad (1)$$

  We define bandwidth degrade degree BR as the ratio between the amount of bandwidth reduced and the requested bandwidth

$$BR = \frac{\sum_{i=1}^{k}(B_i - D_i)p^i\ (t)n^i\ (t)}{\sum_{i=1}^{k} B_i . n^i(t)} \qquad (2)$$

  The system degradation degree SD is the integration of BR over the period t.
- A Connection Admission Control Module (CAC): The CAC is to admit the number of flows that can be served and allocates bandwidth to them through signaling to all the network nodes along the traffic path. It also needs to maintain the QoS requirements of existing connections.
- A QoS Mobility Management Module (MMM): The MMM decides whether terminals are detached, connected or idle from the network and also monitors those active nodes moving at high speed.

- A QoS Monitoring Module (Monitor): The Monitor continuously measures whether its QoS merits of the QoS enabled mobile nodes have been satisfied. If the established end users'QoS profile is not satisfied, this monitor may pass the state information to CAC or other components to trigger specific actions.

4.2 QoS class mapping over UMTS and WLAN

QoS class usually defines a specific combination of bounds on performance parameters and objectives. Global standardization organizations such as: International Telecommunication Union (ITU), Internet Engineering Task Force (IETF), the 3rd Generation Partnership Project (3GPP), Institute of Electrical and Electronics Engineers (IEEE) recommend using on the concept of QoS classes for the network and the application level. In general, previous studies on QoS mapping between different networks technologies are categorized as mapping between parameters and mapping between classes in each different technology [10].

In this paper to profide a unified QoS traffic classes, the QoS traffic classes from UMTS and WLAN are mapped in a new set of QoS traffic classes called: Broadband conversational (B-conversational), Broadband streaming (Bstreaming), Narrowband conversational (N-conversational), Narrowband streaming (N-streaming), interactive and background. Details of mapping relationships are shown in Table 1.

Table 1. QoS Classes Mapping Table

Class	Integrated Network	WLAN	UMTS
1	B-conversational	Voice	-
2	B-streaming	Video	-
3	N-conversational	-	Conversational
4	N-streaming	Video Probe	Streaming
5	Interactive	-	Interactive
6	Background	Best Effort	Background

5 Vertical handoff analysis

Previous researches are demonstrated that Vertical handoff is the handoff between different networks; Horizontal handoff is the handoff within the same network [8]. This section uses simulation experiments to investigate how the proposed approach can improve the overall QoS for the integrated 3G networks and WLAN networks. Following the assumptions widely used in previous studies [9], the call arrivals in our simulation follow an independent Poisson process and the session time of each connection is exponentially distributed. It is well known that dropping an established communication is worse than rejecting a new call. Therefore cellular systems reserve a guard bandwidth for the handoff calls in order to reduce the vertical handoff dropping probability. The reserved guard bandwidth can be either static or dynamic [11]. The dynamic approach often outperforms the static one at the expense of generating more control overheads. However, the static approach is often attractive in practice owing to its design simplicity. In our simulation, a static guard bandwidth (5% of the system capacity) is employed to deal with handoff calls.

Without loss of generality, the integrated network in the simulation consists of one cellular network and one WLAN hotspot. Since WLAN has higher capacity and cheaper than UMTS, we assume the vertical handoff probability from UMTS to WLAN is 5 times as much as that from WLAN to UMTS. The system capacity for UMTS and WLAN is 2 mb/s and 11 mb/s respectively. The bandwidth requirement for each of four QoS classes $\{B_1,B_2\ ,B_3,B_4\}$ and their acceptable degradation level defined in degradation profile are assumed to be a portion of the system capacity shown in Table 2. The reservation signaling cost before the establishment of each new or handoff connection is set to a fixed value. For the sake of clarity, all the relevant simulation parameters are summarized in Table 2.

Table 2. Simulations parameters

Parameter	Value
UMTS Capacity (U)	2 mb/s
WLAN Capacity (W)	11 mb/s
UMTS to WLAN Vertical Handoff	0.05
WLAN to UMTS Vertical Handoff	0.01
Reservation signaling cost	1%*W
Session time	Exp(50)
Guard Band	5%
{B_1, B_2, B_3, B_4}	{5%*W, 3%*W, 5%*U, 3%*U}
{D_1, D_2, D_3, D_4}	{4%*W, 2%*W, 4%*U, 2%*U}
Simulation Time	1000s

The simulation is carried out under various traffic loads. We compare the proposed approach with non adaptive multimedia services. In the experiments, we set the load to

WLAN and UMTS identical in each single experiment and calculate the overall system performance merits.

Comparative bandwidth utilization supported by the proposed adaptive scheme in the integrated network to that without the adaptive scheme under various traffic loads is shown in Figure 3. Clearly, the utilization for adaptive multimedia connection is better than that for non-adaptive multimedia. When the traffic load becomes higher, the advantage is more evident. The reason why adaptive multimedia can better utilize the system bandwidth is that the proposed scheme allows the network intelligently adjust each admitted QoS connection by its degradation profile and give sufficient amount of resources for the new or handoff calls.

Figure 4 further evaluates the vertical handoff dropping probability in the integrated network. The vertical handoff dropping probability for adaptive multimedia connection is less than that for non-adaptive multimedia. When the traffic load becomes higher, the trend is more evident. It reveals that the proposed approach reduces a great number of vertical handoff dropping calls for the integrated WLAN and cellular system.

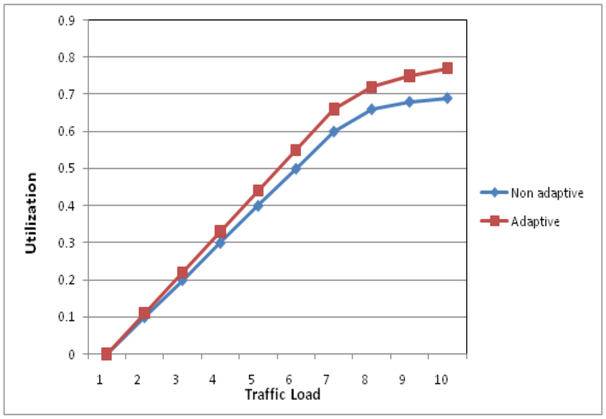


Figure 3. Utilization over traffic load

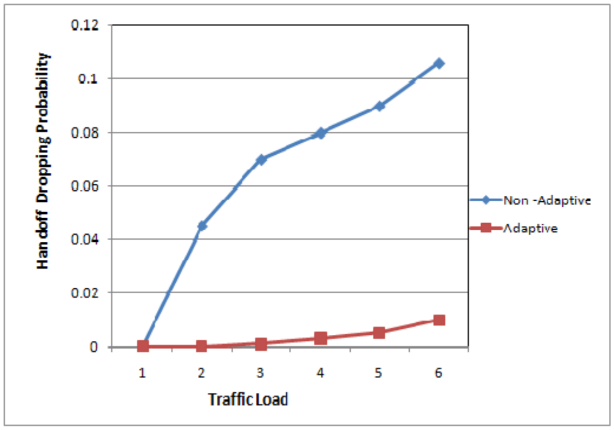


Figure 4. Vertical Handoff dropping probability over traffic load

6 Conclusions

This paper has proposed vertical handoff analysis between heterogeneous networks such as UMTS and WLAN under an Adaptive QoS Control. Through this paper we address the integrated model of cellular and WLAN networks according to the ETSI (European Telecommunication standard institute) designed network architecture that efficiently integrates 802.11 and cellular. Our proposed model supports the delivery of adaptive real-time flows for end users taking the advantage of high data rate WLAN systems as well as the wide coverage area of 3G networks. The simulation results show that the adaptive multimedia framework performance is better than the non adaptive approach in terms of lower vertical handoff dropping probability and call blocking probability while still maintain acceptable QoS to the end users.

As a future and continuation work, the handover operation from WLAN to UMTS needs to be studied further while taking into account more optimization approaches for mobility management schemes. This would provide more QoS features in order to try to annihilate the vertical handoff dropping.

7 Acknwoledgment

We are grateful to the College of Information Science and Engineering of Hunan University; in particular Assamoi Valerie Christian encouraged and supported us.

8 References

[1] Kun Ho Hong, SuK young Lee, Lae Young Kim, and PyungJung Song, “Cost-Based Vertical Handover Decision Algorithm for WWAN/WLAN Integrated Networks”, Hindawi Publishing Corporation EURASIP Journal onWireless Communications and Networking Volume 2009, Article ID 372185, 11 pages, May 2009.

[2] Xin GangWang, Geyong Min, John E. Mellor “Adaptive QoS Control in Cellular andWLAN Interworking Networks ” Mobile Computing and Networks Research Group Department of Computing, School of Informatics, University of Bradford, Bradford, BD7 1DP, UK.

[3] SuKyoung Lee, Kotikalapudi Sriram, Kyungsoo Kim, Yoon Hyuk Kim, and Nada Golmie “Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks”, IEEE Transactions on Vehicular Technology, January 2009.

[4] L. Yu, And V. Leung, “A New Method To Support UMST/WLAN Vertical Handover Using STCP”, IEEE Wireless Communications. Volume 11, Issue 4, pages: 44-51, Aug.2004.

[5] P. Pinto, L. Bernardo and P Sobral, “UMTS-WLAN Service integration at core network level,” ECUMN 2004, pages: 29-39, 2004.

[6] M. Welzl and M. Muhlhauser, "Scalability and quality of service: a trade-off?," Communications Magazine, IEEE, vol. 41, pp. 32-36, 2003.

[7] Xin GangWang, Geyong Min, John E. Mellor , “An Adaptive QoS Management Scheme for Interworking Cellular and WLAN Networks”, CiteSeerX - Scientific Literature Digital Library and Search Engine, 2008.

[8] Ying Wang, Pin Han Ho, Sherman Shen, Shimin Li, Sagar Naik, “Modularized Two Step Vertical Handoff Scheme In Integrated WWAN and WLAN”, high Performance Switching and Routing, 2005. HPSR. 2005 Workshop on pp. 1004–1008, June 2005.

[9] M. El-Kadi, S. Olariu, and H. Abdel-Wahab, "A rate-based borrowing scheme for QoS provisioning in multimedia wireless networks," IEEE Transactions on Parallel and Distributed Systems, vol. 13, pp. 156-166, 2002.

[10] Xiao-hong HUANG Zheng HE, Yan MA and Qiong SUN, “Packet-switch mechanism based on QoS class mapping using flow label for overlay network” The Journal of China Universities of Posts and Telecommunications Volume 17, Supplement 1, pp. 17-23, July 2010.

[11] S. Choi and K. G. Shin, "Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks," IEEE Transactions on Parallel and Distributed Systems, vol. 13, pp. 882-897, 2002.

Throughput Analysis of the IEEE 802.11 Power Save Mode in Single Hop Ad hoc Networks

Pravati Swain, Sandip Chakraborty, Sukumar Nandi, Purandar Bhaduri
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, India 781039
Email: {pravati, c.sandip, sukumar, pbhaduri}@iitg.ernet.in

Abstract—*In the IEEE 802.11 Power Save Mode (PSM) specified for Independent Basic Service Set (IBSS), time is divided into beacon intervals. At the beginning of each beacon interval, each station in the power save mode periodically wakes up for a duration called announcement traffic indication message (ATIM) window. The stations that have successfully transmitted ATIM frame within the ATIM window will compete to transmit data frame in the rest of the beacon interval. The transmission of an ATIM frame depends on the CSMA/CA mechanism specified in the IEEE 802.11 DCF. The probability of a successful transmission of an ATIM frame has a great impact on the performance of IBSS in power save mode. This paper presents an analytical model to calculate the throughput using the success probability of an ATIM frame transmission in ATIM window of fixed size. The simulation results validate the accuracy of this analytical analysis.*

Keywords: IEEE 802.11 standards, power save mode, ATIM frame, Markov chain, throughput analysis

1. Introduction

IEEE 802.11 MAC for wireless LANs is the most used medium access protocol. It defines two methods for channel access, the mandatory Distributed Coordination Function (DCF) and the optional Point Coordinate Function (PCF). DCF is based on carrier sense multiple access with collision avoidance (CSMA/CA) scheme. CSMA/CA uses a binary exponential backoff (BEB) [1] algorithm to avoid collision in the network. A station may proceed to transmit frames if the medium is sensed idle for an interval larger than DIFS (Distributed Interframe Space) period, otherwise it defers the transmission until the medium is idle more than the DIFS period. Then the station generates a backoff time given by:

$$\textit{Backoff time} = \textit{Random}() \times \textit{Slot time}$$

The random value is uniformly distributed over $[0, CW - 1]$, where $CW_{\min} + 1 \leq CW \leq CW_{\max} + 1$, i.e., $CW_{\min}$ and $CW_{\max}$ are the minimum and maximum contention window sizes, respectively. These values are fixed by the physical layer. The backoff counter is decreased as long as the channel is sensed idle and frozen when the channel is sensed busy. After each unsuccessful transmission CW is doubled up to $CW_{\max}+1 = 2^m(CW_{\min}+1)$. The constant m is called maximum backoff stage. For a successful transmission the CW is reset to $CW_{\min}$.

Analytical models have been proposed for the performance analysis of the IEEE 802.11 DCF. Bianchi [2] presents a discrete time Markov model of the IEEE 802.11 DCF with ideal channel conditions. The paper [3] presents a modified version of the Bianchi model, which introduces a fixed retry limit. A number of papers [4], [5], [6], [7], [8] have built upon the original Bianchi model for handling error prone channels, non-ideal transmission channels and capture effect. All these theoretical models are derived for IEEE 802.11 DCF in data frame transmission.

In the IEEE 802.11 power save mode (PSM) for IBSS, time is divided into beacon intervals and each beacon interval divided into two parts, ATIM window and DATA window. In the IEEE 802.11 power save mode for DCF, at the beginning of each beacon interval, each node must stay awake for a fixed interval called the ATIM window. The ATIM window is used to announce any frame pending for stations in power save mode. When a station successfully transmits an ATIM frame within the ATIM window, it can compete to transmit the data frame in the corresponding DATA window. In wireless networks energy resources are considered valuable. Wireless devices usually depend on batteries. The design of “energy efficient" and “energy aware" protocols for wireless networks becomes an important research area. Several MAC protocols have been designed for wireless LANs to minimize the power consumption. The paper [9] introduced a MAC protocol to improve power save in wireless LANs which dynamically chooses an adaptable ATIM window size and different nodes use different ATIM window sizes. The paper [10] proposed a carrier sensing window which is shorter than the ATIM window. However to the best of our knowledge no one has modeled the performance of

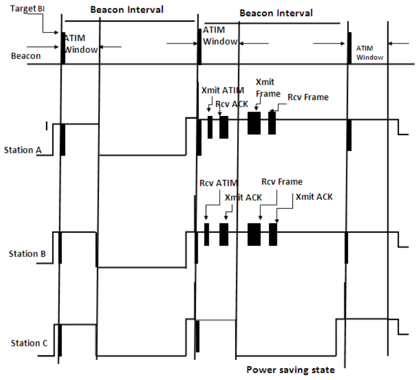


Fig. 1: Power save mode in IBSS [1]

IEEE 802.11 power save mode in IBSS using ATIM frame transmission. This paper presents a discrete time Markov model to calculate the probability that an ATIM frame is transmitted successfully. The throughput of the IEEE 802.11 PSM is then calculated using the ATIM frame success probability. The simulation tool *NS-2* [11] is used to validate the model.

The plan of the paper is as follows. In Section II we first present a brief overview of the IEEE 802.11 PSM. We propose a theoretical model in section III to calculate the throughput using the probability that an ATIM frame is transmitted successfully. Section IV validates the correctness of this model by using simulation. Finally Section V presents the conclusions.

2. The IEEE 802.11 DCF in Power Save Mode

In the IEEE 802.11 PSM there are two different power modes, *power on* and *power save*. In power on or active mode a station transmits or receives frames at any time. We assume nodes are synchronized through a beacon message. Those stations in power save mode wake up periodically to listen to the beacon message and stay awake for an ATIM window period. The transmitter buffers all the broadcast/multicast or unicast frames to the stations in power save mode and announces them in the ATIM window through an ATIM frame. During the ATIM window the control packets are exchanged by the stations to determine whether to go for power save mode or stay awake after the end of the ATIM window for data transmission. The algorithm for the transmission of an ATIM frame is according to CSMA/CA DCF specified in the IEEE 802.11 [1]. For an unicast frame, when a station receives an ATIM frame within the ATIM window, it sends an acknowledgement and stays awake up to the end of the next ATIM window. If no acknowledgement is received the ATIM frame will be retransmitted using the conventional DCF access procedure. If a station is unable to transmit an ATIM frame during the ATIM window, e.g., due to contention, the data frame is buffered and an attempt is made to transmit the ATIM frame during the next ATIM window. If a station does not receive or transmit an ATIM frame during an ATIM window, it may enter the *power save* state at the end of the ATIM window. An ATIM frame or an ATIM-ACK can be transmitted or received only within the ATIM window. A station may discard frames buffered for later transmission to power saving stations if the frame has been buffered for an excessive amount of time. In the IEEE 802.11 standard [1] neither the retry limit nor when to discard the data frame has been specified. As the ATIM window size is very small, the retry limit of seven is not appropriate for ATIM frame transmission. The paper [9] presented a power saving mechanism and has defined the retry limit of three for an ATIM frame transmission and assumed rebuffering of the data frame for at most two beacon intervals.

The power save mode is illustrated through an example. In Fig. 1 station A announces a frame destined for station B by transmitting an ATIM frame during the ATIM window. Station B sends ATIM-ACK to the station A and remains awake for the rest of the beacon interval. Station C goes to *power save* state at the end of the ATIM window, thus saving energy.

3. Modeling and Analysis

3.1 Network model assumptions

To model and analyze the ATIM packet transmission, we have made the following assumptions. We consider n number of stations. We assume a saturation condition, in which each station has packets to transmit at all times. We have assumed an ATIM window of fixed size. The channel is ideal, i.e., there is no hidden terminal and capture [12]. When a station has a data frame in the buffer to transmit it generates an ATIM frame. There is no broadcasting of ATIM frames, only unicasting transmission. If station A successfully transmits an ATIM frame to station B in a beacon interval (BI), then it cannot transmit another ATIM frame to station B in the same beacon interval.

Before every ATIM frame transmission, the station sets the value of CW to $CW_{\min}+1$. For each unsuccessful transmission the CW is doubled up to $CW_{\max}+1$ and for a successful transmission, the value CW is reset to $CW_{\min}+1$. When station A transmits an ATIM frame to station B, the ATIM frame may collide with another ATIM frame sent by another station. In this case the station will retransmit the ATIM frame. The retry limit for an ATIM frame is three within one beacon interval. If ATIM-ACK is not received after three transmissions in one beacon interval, then the data frame is rebuffered for another try in the next beacon interval. An attempt will be made to transmit the ATIM frame for a total of three times. A rebuffered packet can stay in the buffer for at most two beacon intervals. After three beacon intervals if the ATIM frame is not successfully transmitted then the packet is dropped. This algorithm is derived from the idea proposed in [9]. Algorithms 1 describes the ATIM frame transmission. In this algorithm the variable $BeaconNum$ represents the number of beacon intervals.

3.2 System Model

Consider the stochastic process $(s(t), b(t), a(t))$ representing the backoff stage $s(t)$, backoff counter $b(t)$ and backoff layer $a(t)$ (the beacon interval number) at time t. Since we have a discrete model of time, the beginning of two consecutive slots will differ by one time unit. The backoff counter is decremented at the beginning of each slot time. The backoff stage represents the retry number to transmits an ATIM frame within one beacon interval and the backoff layer represents the number of beacon intervals used to successfully transmits an ATIM frame. We have modeled this three dimensional process $(s(t), b(t), a(t))$ with a discrete time Markov chain depicted in Fig. 2, where

$$\begin{aligned} P\{i_1, k_1, a_1 | i_0, k_0, a_0\} = P\{s(t+1) = i_1,& \\ b(t+1) = k_1, a(t+1) = a_1 | s(t) = i_0,& \\ b(t) = k_0, a(t) = a_0\}.& \end{aligned}$$

Assume that p is the conditional collision probability, which is constant for a fixed number of stations and independent of the number of retransmissions. This is the probability p that a frame collides. Consider the probability q that the ATIM window ends in the current slot. This is also independent of the number of frame retransmissions. The non null one-step transition probabilities of the Markov chain in Fig. 2 are presented in the equations (1) in Fig. 3.

Algorithm 1 To transmit an ATIM frame

```
1: $BeaconNum \leftarrow 2$
2: $CW \leftarrow CW_{\min} + 1$
3: $W \leftarrow$ random integer from an uniform distribution
   over the interval $[0, CW - 1]$
4: while $W > 0$ do
5:   if Channel = Idle then
6:     $W \leftarrow W - 1$
7:   end if
8: end while
9: Transmit ATIM frame.
10: if ATIM window ends before ATIM-ACK is received then
11:   $BeaconNum \leftarrow BeaconNum - 1$
12:   if $BeaconNum \geq 0$ then
13:     GOTO 2
14:   else
15:     DROP the ATIM frame.
16:   end if
17: else
18:   if ATIM-ACK is not received after ATIM-ACK
      time out then
19:     $CW \leftarrow 2 \times CW$
20:     if $CW \leq CW_{\max} + 1$ then
21:       GOTO 3
22:     else
23:       $BeaconNum \leftarrow BeaconNum - 1$
24:       if $BeaconNum \geq 0$ then
25:         GOTO 2
26:       else
27:         DROP the ATIM frame.
28:       end if
29:     end if
30:   else
31:     Success
32:   end if
33: end if
```

The first equation indicates that at the beginning of each slot within an ATIM window, the backoff counter is decremented with probability $(1 - q)$. The second equation indicates that at any backoff stage and for any backoff counter value if the ATIM window ends, the protocol tries to retransmit the ATIM frame with backoff stage 0 in the next ATIM window. The third equation indicates successful transmission. The fourth equation indicates either successful transmission or an attempt to starts a new ATIM frame transmission. The fifth equation shows that there is collision at the last try within a beacon interval, so one more attempt will be made to send the frame in the next beacon interval with backoff stage 0. The sixth equation shows that within an ATIM window, if there is unsuccessful transmission at

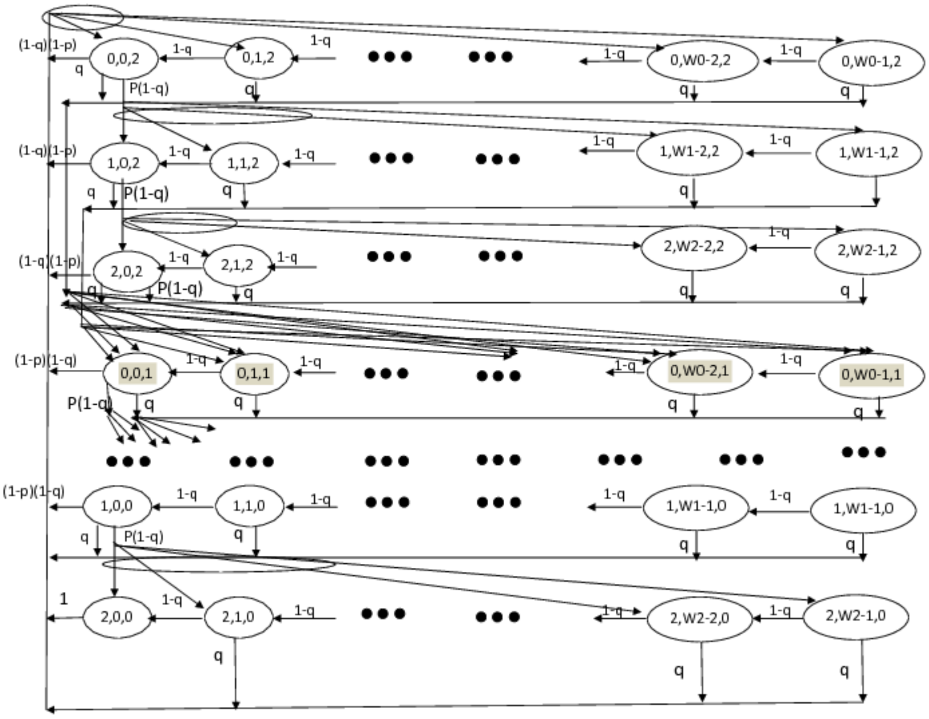


Fig. 2: Markov Model for ATIM frame transmission

$$\begin{cases}
(I) & P\{i,k,a|i,k+1,a\} = 1-q, & i \in [0,2], k \in [0, W_i - 1], a \in [0,2]; \\
(II) & P\{0,k,a-1|i,k',a\} = q, & i \in [0,2], k \in [1, W_0 - 1], a \in [1,2], k' \in [0, W_i - 1]; \\
(III) & P\{0,k,2|i,0,a\} = (1-p) \times (1-q), & i \in [0,2], k \in [0, W_0 - 1], a \in [0,2], \\
 & & \mathit{if}\ \ a = 0, i \neq 2; \\
(IV) & P\{0,k,2|2,0,0\} = 1, & k \in [0, W_0 - 1]; \\
(V) & P\{0,k,a-1|2,0,a\} = p \times (1-q), & k \in [0, W_0 - 1], a \in [1,2]; \\
(VI) & P\{i+1,k,a|i,0,a\} = p \times (1-q), & i \in [0,1], k \in [0, W_i - 1], a \in [0,2]; \\
(VII) & P\{0,k,2|i,k',0\} = q, & i \in [0,2] k \in [0, W_0 - 1], k' \in [1, W_i - 1];
\end{cases} \tag{1}$$

Fig. 3

$$b_{i,k,a} = \begin{cases}
M, & i = 0, k = W_0 - 1, a = 2; \\
M \times \sum_{l=0}^{W_0-(k+1)} (1-q)^l, & i = 0, k \in [0, W_0 - 2], a = 2; \\
N, & i = 0, k = W_0 - 1, a \in [0,1]; \\
N \times \sum_{l=0}^{W_0-(k+1)} (1-q)^l, & i = 0, k \in [0, W_0 - 2], a \in [0,1]; \\
\frac{p(1-q)}{W_i} \times \sum_{l=0}^{W_i-(k+1)} (1-q)^l b_{i-1,0,a}, & i \in [1,2], k \in [0, W_i - 1], \\
 & a \in [0,2]
\end{cases} \tag{2}$$

Fig. 4

backoff stage i, the stage will be increased to $i+1$. The seventh equation shows an unsuccessful transmission, when there is end of ATIM window at the third beacon interval (indicated by $a(t) = 0$) with probability q.

3.3 Model Analysis

Let $b_{i,k,a}$ be the stationary distribution of the above Markov chain, i.e.,

$$\begin{aligned} b_{i,k,a} &= \lim_{t\to\infty} P\{s(t)=i, b(t)=k, a(t)=a\}, \\ & \quad i \in [0,2], k \in [0, W_i - 1], a \in [0,2] \end{aligned}$$

To obtain the stationary distribution $b_{i,k,a}$, we solve the balance equations:

$$\begin{aligned} b_{i,0,a} = \frac{p(1-q)}{W_i} \sum_{l=0}^{W_i-1} (1-q)^l b_{i-1,0,a} \\ 0 < i \leq 2, a \in [1,2] \end{aligned} \tag{3}$$

$$\begin{aligned} b_{0,0,a-1} = \frac{p(1-q)}{W_i} \sum_{l=0}^{W_i-1} (1-q)^l b_{2,0,a} \\ + \sum_{i=0}^{2} \sum_{k=0}^{W_i-1} q b_{i,k,a} \quad a \in [1,2] \end{aligned}$$

The stationary distribution is given by equation (2) in Fig. 4, where

$$\begin{aligned} M &= (1-q)(1-p) \sum_{i=0}^{2} \sum_{a=1}^{2} b_{i,0,a} + b_{2,0,0} + \\ & \quad (1-p)(1-p) \sum_{i=0}^{1} b_{i,0,0} + q \sum_{i=0}^{1} \sum_{k=0}^{W_i-1} b_{i,k,0} \end{aligned}$$

and

$$N = p(1-q) b_{2,0,a+1} + \sum_{i=0}^{2} \sum_{k=0}^{W_i-1} b_{i,k,a+1}.$$

Using the normalization condition for a stationary distribution, the simplified result is as follows:

$$1 = \sum_{i=0}^{2} \sum_{k=0}^{W_i-1} \sum_{a=0}^{2} b_{i,k,a} \tag{4}$$

After some calculation, using equation (3) and (4) we simplify $\sum_{a=0}^{2} b_{0,0,a}$ as a function of the conditional collision probability p, the probability q that ATIM window ends and $CW_{\min}$, the minimum contention window size. We calculate the the probability q using the uniform distribution on the number of ATIM frames that can be successfully transmitted within an ATIM window.

Let τ be the probability that a station transmits in a randomly chosen slot time. This can be obtained as,

$$\tau = \sum_{i=0}^{2} \sum_{a=0}^{2} b_{i,0,a} \tag{5}$$

(6)

As usual the relation between τ and p is

$$p = 1 - (1-\tau)^{(n-1)}. \tag{7}$$

A collision in the channel occurs when at least one of the remaining stations transmit. Let P_{tr} be the probability that there is at least one ATIM frame transmission in the considered slot time. The probability P_{as} that an ATIM frame transmission is successful is given by

$$P_{tr} = 1 - (1-\tau)^n \tag{8}$$

$$P_{as} = \frac{n\tau(1-\tau)^{(n-1)}}{P_{tr}}. \tag{9}$$

This probability value P_{as} gives a gross overview of the number of stations that remain active in the data window if there are n number of stations in the IEEE 802.11 PSM. The total energy saved can be calculated using the probability P_{as}. Similarly the probability P_{as} can be used to calculate the throughput, as it is the probability that a station will stay in power on model in the data window for the data transmission.

3.4 Throughput Analysis

The fraction of time the channel is used to successfully transmit payload bits is called the system throughput [2]. Let S denote the normalized system throughput. In the IEEE 802.11 power save mode, when a station successfully transmits an ATIM frame within the ATIM window, it competes to transmit the data frame in the corresponding DATA window. For simplicity we assume that if a station successfully transmits an ATIM frame within the ATIM window then eventually it can successfully transmit data frames in the DATA window. We calculate the throughput using the probability value P_{as} as follows.

$$\begin{aligned} S &= \frac{E[\mathit{payload\ info.\ transmitted\ in\ a\ slot\ time}]}{E[\mathit{length\ of\ a\ slot\ time}]} \\ &= \frac{P_{as} P_{tr} E[p]}{(1-P_{tr})\sigma + P_{as} P_{tr} T_s + (1-P_{as}) P_{tr} T_c} \end{aligned}$$

$E[P]$ is the average packet payload size (in terms of time unit, e.g., μs). We assume all packets have the same size, so $E[p] = P$. T_s and T_c are the average time the channel is sensed busy because of a successful transmission or a collision respectively, and σ is the empty slot time. Let $H = PHY_{\text{hdr}} + MAC_{\text{hdr}}$ be the packet header and δ the propagation delay. Then

$$\begin{aligned} T_s &= \mathit{DIFS} + H + E[P] + \delta + \mathit{SIFS} + ACK + \delta \\ T_c &= \mathit{DIFS} + H + E[P] + \mathit{SIFS} + ACK. \end{aligned}$$

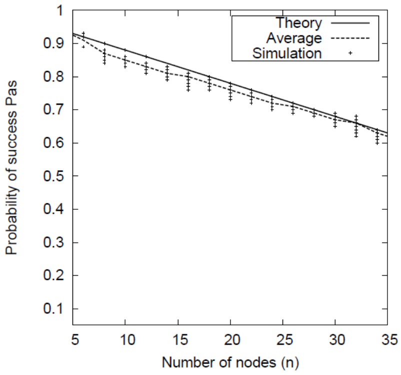


Fig. 5: Probability of Success of an ATIM frame

4. Model validation

For validating our model, we used the simulation tool *NS-2* [11]. The simulation area is chosen such that all stations are within one single hop distance, i.e., the received signal strength is always detectable. We assume the ATIM window period is 20 percent of the beacon interval. We present the throughput for the basic access in IEEE 802.11 DCF in power save mode under the Direct Sequence Spread Spectrum (DSSS) physical layer [1]. The system parameters used in the calculation are listed in Table 1.

Table 1: Parameters used in the calculation

Payload of data packet	1024 bytes
Data	1024 bytes + MAC header + PHY header
ACK	14 bytes + PHY header
PHY header	$192\mu s$
MAC header	28 bytes
Basic rate	1Mbps
Data rate	2Mbps
Slot time	$20\mu s$
SIFS	$10\mu s$
DIFS	$50\mu s$

For a fixed number of stations, we run 10 simulations with different random seed values. The symbol + represents the result of each simulation. Fig. 5 displays the probability of successful transmission of an ATIM frame against the number of nodes. In Fig. 5 the solid line represents the results calculated using the Markov model and the dotted line represents the average value of all 10 simulations for each node. The figure shows that the theoretical and simulation results are close. Fig. 6 presents the throughput against number of nodes. Again the theoretical results match the simulation results. It can be noted that the throughput obtained from our model is marginally less than the one obtained from Bianchi's model due to the the ATIM window overhead of IEEE 802.11 PSM.

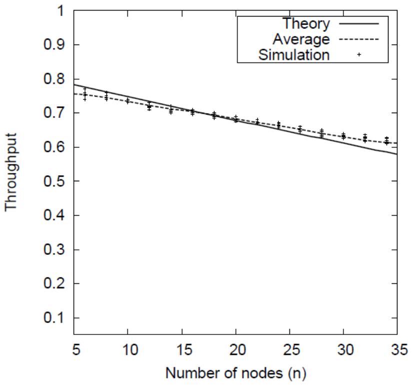


Fig. 6: Throughput of 802.11 PSM with different node size

5. Conclusion

This paper presents an analytical model based on a Markov chain for the transmission of an ATIM frame of the IEEE 802.11 DCF in power save mode. We use the success probability of an ATIM frame to calculate the throughput of the IEEE 802.11 DCF in power save mode. The theoretical results are almost similar to the simulation results in terms of probability of success and normalized throughput.

References

[1] *IEEE Std 802.11-2007, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Edition 2007*, IEEE.

[2] G.Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," in *IEEE Journal on Selected Areas in Communications*, vol. 18, March 2000.

[3] H. Wo, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement," in *INFOCOM*, 2002.

[4] M. Ergen and P. Varaiya, "Throughput analysis and admission control for IEEE 802.11a," in *Mobile networks and Applications 10, 705-716*, 2005.

[5] A. Alshanyour and A. Agarwal, "Three-dimenssional markov chain model for performance analysis of the IEEE 802.11 DCF," in *IEEE GLOBCOM.*, 2009.

[6] T. C. Hou, L. F. Tsac, and H. C. Lia, "Throughput analysis of the IEEE 802.11 DCF in multihop ad hoc networks," in *ICWN, pp.653-659.*, june 2003.

[7] V. M. Vishnevsky and A. I. Lyakhov, "IEEE 802.11 LANs: saturation throughput in the presence of noise," in *IFIP Netw., Pisa, ITALY*, 2002.

[8] F. Daneshgran, M. Laddomada, and M. Mondin, "A model of the IEEE 802.11 DCF in presence of non ideal transmission channel and capture effects," in *IEEE GLOBCOM.*, 2007.

[9] E. S. Jung and N. H. Vaidya, "Energy efficient MAC protocol for wireless LANs," in *IEEE INFOCOM*, 2002.

[10] M. J. Miller and N. H. Vaidya, "Improving power saving protocols using carrier sensing for dynamic advertisement windows," in *IEEE INFOCOM*, 2005.

[11] "Network simulator 2 (ns2), http://www.isi.edu/nsnam/ns," 2009.

[12] K. C. Huang and K. C. Chen, "Interference analysis of nonpersistent CSMA with hidden terminals in multicell wireless data networks in," in *IEEE PIMRC*, 1995.

A Weight based Context Analysis System to Provide a Required UMM Service

Preeti Khanwalkar and Pallapa Venkataram
Protocol Engineering and Technology Unit,
Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore, Karnataka, India

Abstract— *Ubiquitous Computing is an emerging paradigm which facilitates user to access preferred services, wherever they are, whenever they want, and the way they need, with zero administration. While moving from one place to another the user does not need to specify and configure their surrounding environment, the system initiates necessary adaptation by itself to cope up with the changing environment. In this paper we propose a system to provide context-aware ubiquitous multimedia services, without user's intervention. We analyze the context of the user based on weights, identify the UMMS (Ubiquitous Multimedia Service) based on the collected context information and user profile, search for the optimal server to provide the required service, then adapts the service according to user's local environment and preferences, etc. The experiment conducted several times with different context parameters, their weights and various preferences for a user. The results are quite encouraging.*

Keywords: Ubiquitous Multimedia Service; Composite Context; Essential Context-derived Reasons; User Profile Information;

1. Introduction

With the convergence of various networks (wired, wireless, etc.), and devices (PDA, Smart Phone, TV, etc.), the Ubiquitous Computing [7] is becoming a reality. Tremendous growth in the mobile access technologies has enabled services to enter in almost every part of the life. Dramatically, it has stimulated even the demand of multimedia applications to be ubiquitous.

UMMS means provision of different kind of multimedia (audio, video, graphics, etc.) to the user any time, irrespective of location, device he/she carries and without explicit user request. To achieve this system needs to ensure that services must support mobility, interoperability, location awareness, situation awareness, seamlessness, pervasiveness and timely adaptation [5]. UMMS have broad application areas such as remote health care, e-business, ubiquitous learning, on-line entertainments (sports, movies, etc.) and so on.

There exists a strong relationship between the various contexts and types of service required based on it. User would like to have services that self configure themselves in the user's physical environments and integrate seamlessly with their everyday tasks in an intuitive, and non-intrusive way. With contextual information, the system foresee the user's requirements and acts proactively, without any user's explicit interaction. It enables system to determine relevance of service in the user's operating environment, which improves user satisfaction and service quality.

Many researchers have given various definitions of conext [1], [4]. Any environment attributes that is related to the particular application domain can be considered as a context data, like temperature, location, any object, its status, etc. However among all the available contextual parameters some are more critical or have more weightage as compared to other depending on different types of situation. We consider a system with different weights of context values. These weightage can be varied and decided dynamically, while finding the appropriate service for the user. In our work we are concentrating on on-line ubiquitous multimedia services, such as live sports as a case study to discuss the usability of our system. Servers that contains required service are connected to the Internet, and as user activates his/her device, it can connect to a nearby available network (WiFi, GSM/GPRS, etc.) thus can connect to nearby available server via Internet.

1.1 Proposed Idea

We propose a new approach, to make use of context information with different weightage and user profile information for ubiquitous multimedia service identification, which is proactive and adaptive based on various changes in user's surrounding environment. Collected Preliminary Contexts *(PCs)* information, we analyze as a Composite Contexts *(CCs)* which further analyzed as Essential Context-derived Reasons *(ECRs)*. This inferred *ECRs* along with User Profile Information *(UPIs)* is used to identify the required ubiquitous multimedia service. This identified service is optimally trace to cater the user services at the required time.

1.2 Organization of Paper

The remainder of this paper is structured as follows. Section 2 describes some of the existing works, Section 3 presents our approach of context information analysis, Section 4 provides an overview of the system architecture and discusses the functionality of each components, Section

5 presents a case study to show the usability of the system, Section 6 explains simulation and results, followed by conclusion in Section 7.

2. Related Works

Many works states the context awareness and user profile in their applications and also, context deduction and use of composite context for various applications in different ways. In [2] have proposed a personalized context-aware services architecture depending on user contexts that are collected from ubiquitous sensor networks. In [3] an exploration the relationship between context awareness and user modelling, through the design of a context-aware personal assistant is shown. The DUPS (Dimension User Profile System) architecture is explained in [6], to recommend services to user based context. It stores location, time, and frequency information of often used services which enables system to provide more accurate service in less time. None of these system has considered the weights of context parameters to identify the appropriate service requirement.

3. Context Information Analysis

The system uses the various context information based on weight to infer a multimedia service corresponding to a definite circumstances of a user. Following section describes the weight allocation.

The context analysis process is done in three steps, starting with *PCs* acquisition to *ECRs* identification, major building blocks are as shown in Figure 1. Using *PCs* parameters step by step derivation and analysis is described in the following sections.

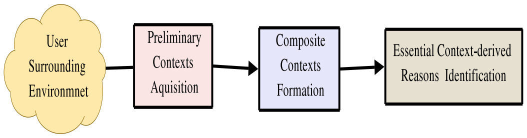

User Surrounding Environmnet
Preliminary Contexts Aquisition
Composite Contexts Formation
Essential Context-derived Reasons Identification

Fig. 1: Context Analysis Procedure used in the System

3.1 Preliminary Contexts*(PCs)* of the System

This represents the context parameters obtained directly from various types of sensors, embedded devices, etc., in the environment. These are common for all users.

We also consider the task and physical environment context which are related to a multimedia application and we categorize them into three sets of *PCs* values: user's physical environment*(PE)*, user's context*(U)* and task context*(T)*. For example preliminary context values of a multimedia application *X*, is given as $PC_X = \{PE, U, T\}$. They further define as follows;

User's Physical Environment(PE): It indicates user's surrounding environment which is characterize as e.g., location(p1), time(p2), temperature(p3), noise, light, available resources, etc. We consider a *PE* context of a user as,

$PE = \{p_1, p_2, \cdots, p_j\}$

User's Context(U): It includes information related to the user, its social and physiological parameters etc., e.g, social relation(u1), people surrounded by(u2), blood pressure(u3), heart rate, etc., Thus we can view user's context as,

$U = \{u_1, u_2, \cdots, u_k\}$

Task Context(T): These are the parameters required to adapt various multimedia contents related to a specific service. Based on the types of service, in what format is it required. e.g. device memory(t1), available networking interface(t2), supporting media format(t3), etc. so we take task context as,

$T = \{t_1, t_2, \cdots, t_l\}$

3.2 Composite Contexts*(CCs)*

It is a context information that can be inferred from *PCs* information. It combines and relates various *PCs* information(same or different types) along with different weightage(based on the services required). Combining various *PCs* information may generate a more accurate understanding of the current situation, rather than taking into account an individual context. In generic terms, for example,

$CC_1 = \langle p1, p2\rangle$; *where* $p1, p2 \in PE$.

$CC_2 = \langle p1, p2, u1, u2\rangle$; *where* $p1, p2 \in PE$ *and* $u1, u2 \in U$.

Steps to formulate *CCs* is given in Algorithm 1, and some of the examples of formulated *CCs* is given in Table 1.

Algorithm 1 Composite Contexts*(CCs)* Formation

```
Begin
Input: PCs and corresponding weights; Output: CCs.
while not end of user session do
    collect PCs from various sources.
    if context changes then
        get $PE = \{p_1 \cdots p_j\}$;
        get $U = \{u_1 \cdots u_k\}$ and $\{T = \{t_1 \cdots t_l\}$;
        assign weights to each PCs based on relevance.
        derive CCs based on context rules.
    else
        wait for context change;
    end if
end while
End
```

3.3 Essential Context-derived Reasons*(ECRs)*

It is derived abstract information. For applications to make context aware decisions, *CCs* information must be further inferred, by considering appropriate weights of *CCs*. This significantly helps in making right decision to identify an appropriate service under certain situation. Various *CCs*

Table 1: *CCs* Formulation

Various PCs involved in CCs Formulation	*Formulated CCs*
CC1-⟨location, time⟩	*Object position at a given time*-It gives the notion, how far a particular service is suitable at some location at a particular time instance.
CC2-⟨noise level, temperature, light, pressure, humidity, available resources⟩	*Ambiance of surrounding environment*- The aggregate of surrounding things and conditions, creates an atmosphere, that influences the user mood.
CC3-⟨location, time, people surrounded by, social relation⟩	*User behavior*- User behavior changes according to place, surrounding people and interrelationship among them.
CC4-⟨ blood pressure, skin respiration, heart rate⟩	*User physical status*- Physiological parameters helps to determines the physical status of the user.
CC5-⟨memory, screen size, resolution, battery power, processing power⟩	*User's device capability*- Every user carries different devices with dissimilar capacities in terms of processing power, memory etc. thus multimedia service has to be tailored according to user device capability.

Table 2: *ECRs* Formulation

Various CCs involved in ECRs Formulation	*Formulated ECRs*
ECR1-⟨Object position at a given time, ambiance of surrounding environment, devices capabilities ⟩	*Service consumer position*- By knowing the ambiance of surrounding environment, devices capabilities*(HDTV, Laptop, etc.)*, and approximate location, one can determine exact position of a user where the service needs to be consume, like- In a home (living room, kitchen etc.), office(cafeteria, conference room etc.)
ECR2-⟨user behavior, user physical status⟩	*User's mood/Emotion*- By understanding user behavior, his physical status, according to people surrounded by, etc., one can recognize user's mood whether he is relaxed, stressed, physically tired, etc.
ECR3-⟨device capability, network conditions⟩	*Service consumption capability*- User may carry several devices with different networks available in his surrounding environment like - Laptop with WiFi, PDA with GPRS, etc. For efficient execution of service system must understand the device and network capability where service needs to be consumed.

information integrates together with some sets of predefined inference functions that results into *ECRs*. One can deduce exact service or location from it. For example exact location like if user is at home, whether he is in kitchen, or living room, etc., an activity of the user like eating while watching TV, or walking in garden or frontyard, etc. In generic terms, for example,

ECR_1 = ⟨ CC_1, CC_2, CC_5⟩,
ECR_2 = ⟨ CC_3, CC_4⟩

Steps to formulate *ECRs* from *CCs* is given in Algorithm 2, and Some examples of identified *ECRs* is given in Table 2.

Algorithm 2 Essential Context-derived Reasons*(ECRs)* Identification

```
Begin
Input: CCs and corresponding weights; Output: ECRs
while not end of user session do
   collect CCs information.
   if context changes then
      get $CC_1$, $CC_2$ $\cdots$ $CC_n$;
      assign weights to each CCs based on relevance.
      derive ECRs based on context rules.
      store inferred ECRs in context history database;
   else
      wait for context change;
   end if
end while
End
```

3.4 Weight Allocation

All the available information within a specific domain may be considered as a context data but not every context data is equally relevent to infer a higher level composite context information. We assume that the *ECRs* and *UPIs* are involved in determining the UMMS. Thus with an assumption that *UPIs* is accurate, and if we consider that *ECRs* is also accurate, system identifies accurate service. If we keep the weight of *ECR* as unity, {w_{ecr_m}=*1*}, we compute the weights of the *CCs* involved in the formation of *ECRs* as per their importance. By assigning weight to each *CCs* according to eq. 1, we get significance of it to formulate *ECRs*.

$$w_{cc_i} = \frac{k_{im} w_{ecr_m}}{n_{CC}} \tag{1}$$

where w$_{cc_i}$ ∈ [0,1], k_{im} is a degree of relevance of CC_i with ECR_m based on various applications, which is decided based on empirical observations after conducting various experiments while finding the appropriate service for the user, and n_{CC} represents number of *CCs* involved in formulating a ECR_m.

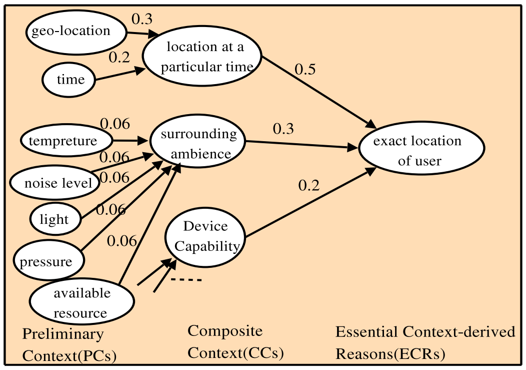


Fig. 2: An Overview of Weight Based Context Analysis

Figure 2 gives an overview of the weight based context analysis scheme. For example, to understand at home, exactly in which room user is in? with geographic location and time we know user is at home, he may be in living room or kitchen, etc., then with the ambiance of surrounding and available devices capabilities, we can determine that user is in living room. In this example, Geo-location and time is important to identify that user is at home, as compared to other context information. As shown in Figure 2, we computed more weightage (0.5), for location at a particular time as compared to surrounding ambiance with weightage (0.3), which in turn has more weightage as compared to device capability with weightage (0.2). Similarly, we assign appropriate weights to *PCs* based on their relevance while formulating *CCs*.

4. UMMS System Architecture

Ubiquitous multimedia Service(UMMS) system architecture is as shown in Figure. 3. We assume that user in a ubiquitous environment carries multiple devices, with multiple networking interfaces, and surrounding environment of the user consists of various sensor based biometric technologies to uniquely identify the user. We are considering this context based multimedia service provision at the application layer to directly provide UMM service functionality to the users. We also assume that the network is connected and equipped with appropriate routing and transport layer protocols for reliable communication.

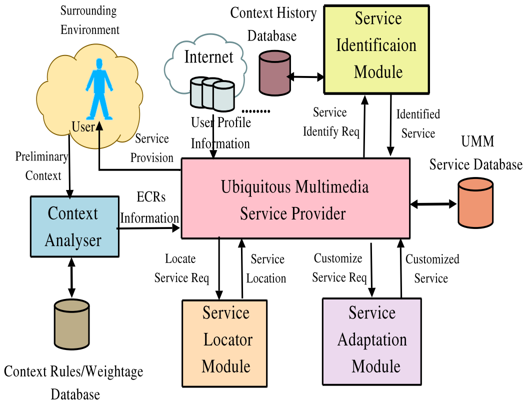


Fig. 3: Ubiquitous Multimedia Service System Architecture

UMMS system consists of three main modules; *1) Service identification module 2) Service discovery module 3) Adaptation module*. We discuss the functionality of these modules after introducing the databases used in the system.

Context Rules/Weightage Database

It contains the logical predefined rules to formulate *CCs* and *ECRs*, that are expressed as conditional and action statements. Conditions are expressed distinctly in the form of boolean expressions, and logical operators like or, equal, more than, and less than, etc. Also it stores the corresponding weights assigned to *PCs* and *CCs* while providing an appropriate service to the user based on the situation. For example- It stores weights of *PCs* like Geo-location as (0.3) and time as (0.2), while determining exact location of a user, to provide a service of on-line cooking recipes. As explained above context analyzer utilizes the information stored in database while formulating *CCs* and *ECRs*.

Context History Database

A persistent storage is needed that includes history-based organization of identified service for a user, corresponding to the specific combination of *ECRs* and *UPIs*. Before any service identification, respective trends in the context history database is evaluated by service identification module. Thus for a known user, it supports service identification module, for retrieving information in much faster way.

UMM Services Database

UMM service provider maintains the database of some of the often used services, to provide access in lesser time and better service in terms of quality. Some of the offered *MM* services and its locations is as shown in Table 3.

Table 3: Some of the multimedia services and its locations

Multimedia Services	***URL/Location***
On-line Games	umm@og.pet.ece.ernet.in
On-line Restaurant Lists	umm@rl.pet.ece.ernet.in
On-line Music	umm@omus.pet.ece.ernet.in
On-line Movies	umm@omov.pet.ece.ernet.in
On-line Cooking Recipes	umm@ocr.pet.ece.ernet.in

User Profile Information(UPI)

User profile information are obtained from unique identification of user or its device. Without a notion of the unique identity, information of a user profile like preference, etc. could not be used for adapting the system. UMM service provider collects *UPIs* from some social networking site. Although the user's profile information is quite steady, and hence his service requirements too, but different types of multimedia services are recommended as a function of his mood, presence of other people, surrounding environment etc.

Service Identification Module

At any time instant, it uses *ECRs* information genrated by context analyzer, along with *UPIs*, to invoke any one of the multimedia service, as represented in eq. 2, which can be interpreted in the form of various definitions like *context based preferred service(CBPS), context based probable(CPS)*, etc., as explain below. In addition *service identification module* checks respective trends in context history database, for a known user. Different understandings of all this information, when inclined to one direction, will results into focused and accurate multimedia service identification. It stores that identified service corresponding to a user and *ECRs* in context history database. At any time instant t = t_1, we have

$$(ECRs(t1) \times UPIs(t1)) \Rightarrow TriggerUMMS(t1) \quad (2)$$

- *Context Based Preferred Service(CBPS): Service position + User's activity + Preference ⇛ CBPS;* If one knows user's present location, current situation and preference, one can judge a strong liking or whether a person is in favor of something. As preference of a user is dynamic, it varies according to different conditions. For example- while travelling in a bus to the college, Paul usually would like to listen music but since he has exam today, he would like to revise his lecture notes through on-line education video lectures on his PDA/Mobile.
- *Context Based Probable Service(CPS): Service position + History ⇛ CPS;* History of a person gives static context information that can be retrieved from a stored database. With this one can have a reasonable basis for establishing presumption under specific circumstances. For example- while travelling, on the way to college, Paul has habit of watching on-line education video lecture, on his PDA/Mobile.
- *Context Based Service Prediction(CBSP): Service position + Preference + History ⇛ CBSP;* Including past patterns and through a logical reasoning one can anticipate the service requirement for the user at any given point of time. If the past history is known one can guess user's multimedia service requirement and based on the preference that can be filtered further according to present conditions. For example- Paul and his friends usually watch mathematical video lectures but, but today since exam is over, so system understands the situation and provides some comedy movie of their liking.

Working of module is as shown in Algorithm 3.

Algorithm 3 Service Identification Logic

```
Begin
Input: ECRs and UPI; Output: Identified Service
while not end of user session do
    collect ECRs and UPIs from UMM service provider.
    formulate different combinations using ECRs and UPIs, e.g. CBPS, CPS, CBSP etc.
    if user is known then
        check respective trends in context history database.
    end if
    determine inclinations of all this information towards one direction.
    if inclined to one service then
        identify the service;
        send it to UMM service provider.
    else
        wait for further change in context information;
    end if
end while
End
```

Service Locator Module

This module is responsible for the discovery of the server. Once service is uniquely identified for a user, it needs to be discovered and fetched from the service provider. If multiple or replicate service providers are available for the same service, optimal service provider is chosen based on user's context like location, device and available network etc. Working of module is as explained in Algorithm 4.

Algorithm 4 Working of Service Locator Module

```
Begin
Input: Identified Service; Output: Service Provider Location
while not end of user session do
    collect the information of required service from UMM service provider.
    discovered and fetch the optimal service provider.
    send the service location to UMM service provider.
end while
End
```

Service Adaptation Module

Dynamic service adaptation is required for the user, so as to the user it appears that data is coming from a unified source. For adaptation of services it is important for a system to understand where the service needs to be consumed. The adaptation module takes adaptation decisions based on user's needs, preferences, device capability and network conditions, and fetched customized service from adaptation proxy, on which various functions like transcoding, content filtering, etc., has implemented. Working of module is given in Algorithm 5.

Algorithm 5 Working of Adaptation Module

```
Begin
Input: Original Service($S_o$), UPIs and ECRs; Output: Customized Service($S_c$)
while not end of user session do
    get $S_o$
    collect adaptation parameters = $\{AP_1, AP_2, \cdots, AP_n\}$ that represents various network, device characteristics and user preferences$\in$\{UPIs, ECRs\} information, from UMM service provider.
    fetch ($S_c$) according to adaptation parameters. $S_c = F(S_o, UPIs, ECRs)$
    Send $S_c$ to the UMM service provider.
end while
End
```

UMM Service Provider

UMM service provider can handle multiple user requests at a time. It collects *ECRs* from context analyzer, and *UPIs* from social networking site one by one, and sends this information to service identification module. UMM service provider gets the identified service from service identification module, It checks for the identified service in its UMM service database, if service is available in the desired format, it directly send it to the user, else send it to the adaptation module for the required customization along with *ECRs* and *UPIs*. If service is not available in the UMM service database, it forwards service locating request to service locator module to discover and fetch optimal service provider based on the context.

If some of the information is misssing then based on other collected context based information, we can only prejudice the service requirement which may or may not be correct. If we have all information available, we can identify a unique multimedia service requirement in a systematic way and one can have focused and accurate service.

We define a service response time of the system is the time required for processing of context data, to identify a multimedia service, for discovery and customization. Service response time is defined as below.

$$\textit{Service Response Time} = t_{ccf} + t_{ecrf} + t_{sid} + t_{sl} + t_{sc};$$

where, t_{ccf} is the time required to formulate *CCs* from simulated preliminary context data, t_{ecrf} is time required to formulate *ECRs*, and t_{sid} to infer *ECRs* and *UPIs* to identify required service, t_{sl} is the time required to locate the service provider, t_{sc} is the time taken to customize the service.

5. Case Study

The working of proposed system is explained using a case study for a live sports ubiquitous multimedia service.

Consider Paul(a college student) is in hostel room. and formulated *ECRs* indicates that Paul is alone in hostel, sitting on a sofa in a relaxed mood, having a Laptop with WiFi connectivity and his profile information reveals that he usually prefers watching sports, shows inclination towards watching a world cup. Thus system understands the situation of Paul, identify his service requirement, according to his current situation and trigger a live match on his Laptop. Following four cases shows, how system understands different situation of Paul and customize service accordingly.

*Case1:*Paul is in a hostel room, having a Laptop with WIFI connectivity. System recognizes his device and network characteristics, and since high bandwidth, high memory, high battery power is available. It sends a high profile Mpeg video stream on his laptop, and Paul starts watching live sports video on his laptop.

Case2: Suddenly his friend called and ask him to come to the department, Paul don't want to loose a match for a single moment. He started driving bicycle. He is carrying a mobile with GPRS connectivity. System understands his context and bandwidth limitations, thus send audio stream so that he can listen match commentry while cycling.

Case3: Paul reach to the department and start moving towards lab to meet his friend, system understands his situation and adapt accordingly, and Paul started watching close-up shots of his favorite player movements on his mobile screen.

Case4: As soon as Paul reach to the lab, he switched on his Laptop and connect to a WiFi. System identifies his environment and again starts sending a high profile Mpeg video stream on his laptop, and Paul can watch live sports video on his laptop.

Timing diagram of various sequence of events is as shown in Fig. 4. As shown in the event sequence diagram service provision is dynamically adapted based on the acquired context information like capability of a device, available resources and accessed network, user preference etc.

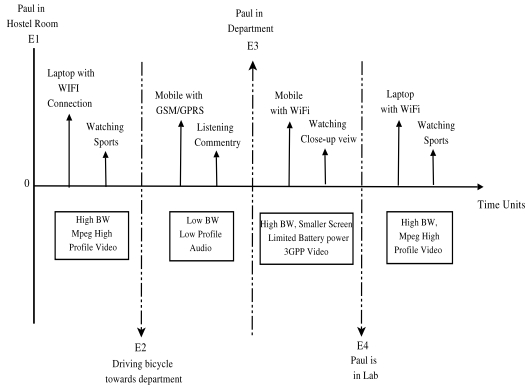


Fig. 4: Event Sequence Timing Diagram of a Case Study

6. Simulation and Results

6.1 Simulation Environment

We have conducted series of experiments in order to evaluate our approach. We have simulated the scheme in a hybrid environment which has WIFI, bluetooth and a GSM network units as shown in Fig. 5. *PCs* are simulated in the given simulation environment and *CCs* and *ECRs* are inferred from them. Further addition of *UPIs* to *ECRs* gives service identification.

Result corresponding to percentage of information available verses cumulative accuracy of service identification is shown in Fig. 6. Cumulative accuracy is the accuracy of a service prediction over a set of 50 users session. As shown in Fig. 6 we consider three different sets of context parameters, based on the weightage. If important sets of context parameters(Context Set3 like time, location, preference, etc.) are missing, it greatly influences inference and accuracy reduces suddenly, similarly some context parameter have less and moderate influences on inference and thus affects service identification accuracy accordingly, plotted with Context Set1 and Set2 respectively.

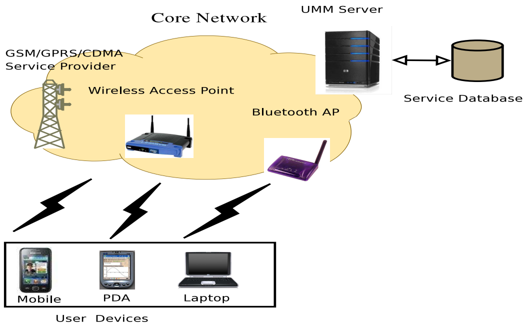


Fig. 5: Simulation Environment of the System

More the user interact with the system, more the system learns from past history and keeps on refining user's profile. System also maintains the database of often used services. As information stored in the data base corresponding to a particular user grows we can identify service more accurately, as usually user's daily routine is fixed. Total service response time will reduce as shown in Fig. 7, and accuracy of the system to provide a service will increase as shown in Fig. 8, with the increase in database of the user.

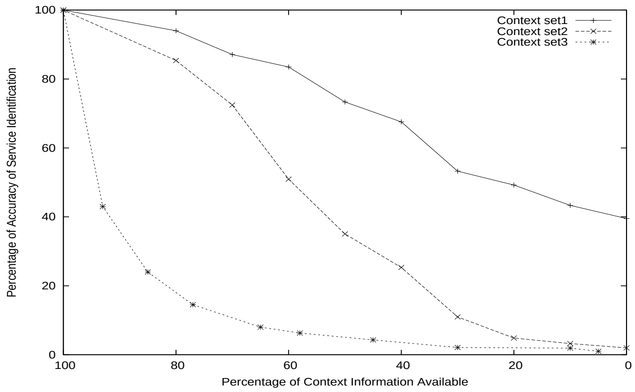


Fig. 6: Context Information Vs. Accuracy of Service Identification

7. Conclusion

This system is designed to support multiple ubiquitous multimedia service access in a ubiquitous environment. The paper suggests a new approach to analyze context information based on weight, in association with user profile information as a base to provide an appropriate multimedia service to the user. This approach is beneficial with respect to ubiquitous environment in providing adapted services to the user proactively, which can be applied to various ubiquitous applications such as ubiquitous museum, ubiquitous smart home, office etc., to improve the quality of user experience. The key feature of the approach is not only in accurate service identification, but also dynamic adaptation according to various situation so as to maximize user satisfaction.

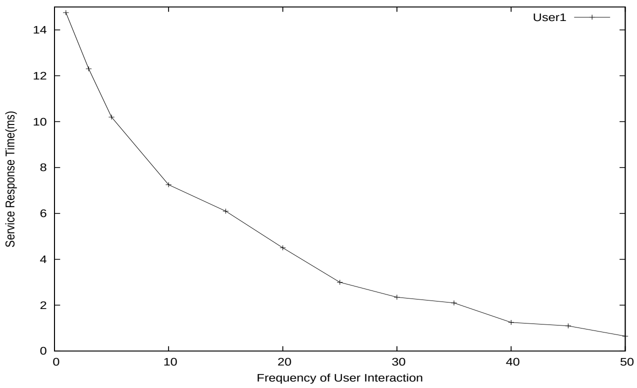


Fig. 7: Service Response Time Vs. User Interaction Frequency

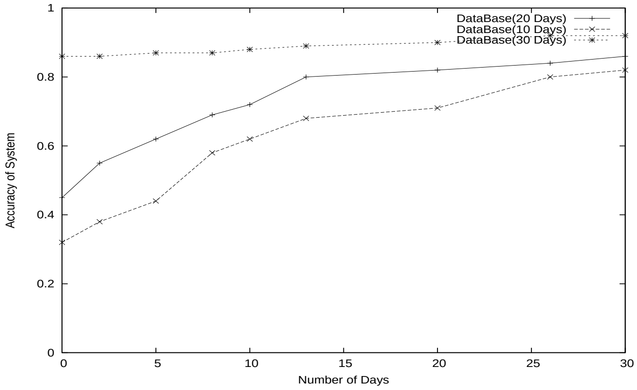


Fig. 8: Number of days Vs. Accuracy of System

References

[1] G. D. Abowd, A. K. Dey, P. J. Brow, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer-Verlag London, UK, September 1999.

[2] J. An, S. Pack, S. An, M. Kim, and Y. Jeon. A novel service architecture for personalized context aware services. In *Proceedings of the 11th international conference on Advanced Communication Technology,Volume-1*, pages 554–559. IEEE Press, NJ, USA, 2009.

[3] H. E. Byun and K. Cheverst. Exploiting user models and context-awareness to support personal daily activities. In *In Workshop on User Modeling for Context-Aware Applications, Sonthofen*, July 2001.

[4] G. Chen and D. Kotz. A survey of context-aware mobile computing research. In *Technical Report TR2000-381*, pages 1–16. Dartmouth College Hanover, NH, USA, 2000.

[5] I. Y. Chen, S. J. Yang, and J. Zhang. Ubiquitous provision of context aware web services. *IEEE International Conference on Services Computing (SCC'06)*, pages 60–68, September 2006.

[6] C. Jang, J. Kim, H. Chang, E. Choi, B. Kim, and G. S. Lee. Method of profile storage for improving recommendation accuracy on ubiquitous computing. In *Second International Conference on Future Generation Communication and Networking*, volume 2, pages 90–94, 2008.

[7] M. Weiser. The computer of the 21st century. In *ACM SIGMOBILE Mobile Computing and Communications*, pages 3–11, July 1999.

Analysis of Processing Parameters of GPS Signal Acquisition Scheme

Prof. Vrushali Bhatt, Nithin Krishnan
Department of Electronics and Telecommunication
Thakur College of Engineering and Technology
Mumbai-400101, Maharashtra, India.

Abstract--- *The primary objective of this research is to analyze the GPS signal acquisition process. To achieve this objective, first several acquisition schemes for the L1 C/A-code are implemented for this research. The acquisition schemes namely circular convolution and modified circular convolution are analyzed in terms of mean acquisition time, acquisition gain and ability to acquire the correct signals. It is observed that the circular convolution scheme provides a better gain but at the cost of processing time and memory whereas the modified circular convolution scheme can be used to reduce acquisition time and memory requirements but the gain is less than that for circular convolution scheme.*

The Global Positioning System (GPS) has become a critical part of the navigation infrastructure not only within the United States but also in other nations around the world. This system was developed by the Department of Defence (DoD) to support the military forces of the United States of America by providing worldwide, real-time positions. GPS can be used for civilian applications even though it was developed for military applications. It consists of a constellation of 28 satellites orbiting around the earth at 20,000 Km above the earth. It provides three dimensional position and velocity anywhere in the world under all weather conditions. The GPS concept is based on satellite ranging. The user estimates time of arrival of the transmitted signals by GPS satellites and uses it to compute its position. A GPS receiver must detect the presence of the GPS signal to track and decode the information from the GPS signal required for position computation. Tracking of the signals is possible only after they have been acquired, so acquisition is the first step in the GPS signal processing scheme. The acquisition process must ensure that the signal is acquired at the correct code phase and carrier frequency.

Keywords: GPS, acquisition, circular convolution, modified circular convolution, Doppler, FFT.

1. GPS acquisition

A GPS receiver must detect the presence of GPS signals to track and decode the information for the position computation. A receiver replicates the GPS signal with code and Doppler. The code phase varies due to the range change between the satellite and the receiver. Doppler variation is due to the relative motion between the satellite and the receiver [1]. The role of the acquisition is to provide a coarse estimate of the code phase and the Doppler to the tracking loops. The satellite motion induces a Doppler within ±5 KHz from the GPS L1 frequency [2]. User dynamics and clock drift introduce an additional Doppler in the GPS signal. The acquisition Doppler search range should be expanded to include these uncertainties to enable proper acquisition. The code phase search range extends from 1 to 1023 chips (of the C/A-code). The acquisition process searches the signal for a particular value of the code phase and Doppler frequency over a certain period of time called the predetection integration time. The acquisition time is determined by the predetection integration period and the number of cells (obtained from code phase and Doppler range) to search. The GPS receiver can compute visible satellites from approximate knowledge of the receiver position, the GPS time and the almanac which reduces the number of satellites to be searched and speeds up the TTFF. There have been various acquisition methods developed to acquire GPS signals and two of them are discussed below.

1.1 Circular convolution (FFT method)

In this method, the signal is transformed from the time domain to the frequency domain using a Discrete Fourier Transform (DFT) [3]. This method uses the correlation property of the Fourier transform. The property states that the correlation of two sequences in the time domain is the same as the inverse Fourier transform of the convolution of the Fourier transform of the two sequences. For a particular Doppler bin, the correlation of the two sequences performed at all code phase shifts is the same as the inverse Fourier transform of the product of the Fourier transform of the two sequences. Thus, this method reduces the acquisition search range to one-dimension.

The cells are searched in parallel by taking the FFT of the incoming and the local signal which reduces the acquisition time. The steps involved in this scheme are [3]:

1. Collect the sampled IF signal for the desired coherent integration period: x(t)
2. Take the FFT of the input signal: X(F)

3. Generate the local PRN code for the same coherent integration period and modulate it with the carrier (IF + desired Doppler) and sample it at the same sampling frequency: y(t)
4. Take the FFT of the local signal: Y(F)
5. Perform convolution in the frequency domain: Z(F) = (conjugate X(F)) * Y(F)
6. Transform the convoluted signal in the time domain: z(t) = IFFT(Z(F))
7. Compute the absolute value of the signal z(t), where z(t) represents the correlation of the input signal with the local signal for that Doppler and all possible code phase shifts.
8. Find the peak of the absolute value of z(t) and compare it against the noise threshold. If the peak is greater than the detection threshold, a signal is present. The detection threshold gives an indication of the noise power present. If a signal is not detected, the procedure is repeated for all possible Doppler values. The detection threshold is optimally based on the noise spectral power density and the allowable probability of false acquisition.

1.2 Modified circular convolution

This method is same as the circular convolution method except for the length of the FFT which is reduced by half [2]. The C/A-code and P-code are transmitted in phase quadrature with each other on the L1 frequency. Hence most of the C/A-code information is contained in the in-phase part of the GPS spectrum. The second half of the spectrum contains little signal information. Hence, this method takes only half the spectrum and performs the correlation [2]. The use of half of the spectrum results in a lower number of FFT points. This reduces the FFT processing time and the acquisition time. There is a loss of 1.1 dB determined from simulation analysis, which is due to a loss of the signal information in the other half of the GPS spectrum [2].

2. Acquisition implementation

The acquisition process is used to detect the presence of a signal and provide coarse estimates of the code phase and Doppler to the tracking process. It exploits the autocorrelation and cross-correlation properties of the GPS PRN codes to acquire the signal. A block diagram of the acquisition process is shown in Figure 1. All the blocks except the acquisition detector and the acquisition manager are common to the tracking process. The acquisition and tracking processes form the core blocks of the correlator in a GPS receiver. Different modules in the acquisition process are discussed below.

Acquisition manager: This module manages the various blocks of the acquisition process and specifies the parameters of operation to each block. It decides the PRN to be searched and the predetection integration time for each cell search. It also specifies the Doppler and code phase range to be searched for the corresponding PRN along with the parameters to compute the detection threshold for acquisition.

Local carrier signal generator: This module is used to generate the carrier to match the frequency of the incoming IF signal. It generates a carrier signal with frequency as the sum of the receiver IF and the Doppler frequency to be searched. It generates both the in-phase and quadrature components of the carrier signal. The Doppler frequency is modified after all the cells for that particular Doppler are searched with no success.

C/A-code generator: This module generates the C/A-code for the desired PRN number. The C/A-code generator should be capable of generating the code for all GPS satellites.

Code shifter: This module is used to shift the C/A-code by the code phase amount to be searched. The code phase should be properly matched with the incoming signal to acquire it.

Combiner: This module is used to combine the signals applied at its input. The carrier signal is combined with the shifted C/A-code to obtain a local replica of the incoming signal.

Sampling module: The incoming IF signal is sampled at an appropriate sampling frequency chosen to avoid the aliasing effect and to reduce processing power. The sampling signal used to sample the incoming signal must match in phase with the local signal. If there is a phase mismatch, there will be incorrect representation of the local signal with the incoming signal which will yield incorrect results.

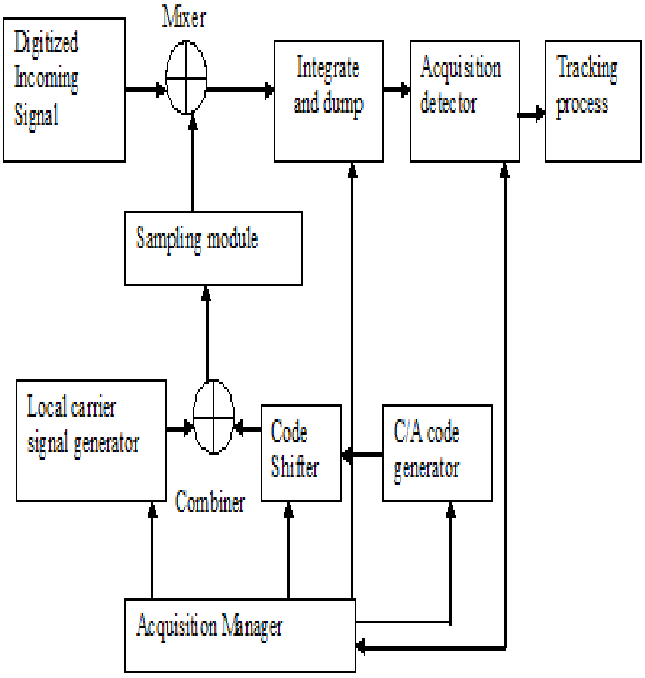


Figure 1: Block diagram of the GPS acquisition process

Mixer: It mixes the incoming signal with a local replica signal to perform carrier and code wipe off. The resulting signal consists of two components with frequencies as the sum and the difference of the two signals. Correlation is performed during the code wipe off which yields a correlation peak. The acquisition detector determines whether the correlation peak is correct. The high frequency component at the mixer output needs to be eliminated and the low frequency component should be processed to determine if the acquisition is a success.

Integrate and dump: This section integrates the mixer output and acts as a low pass filter (LPF) to eliminate the high frequency component. The integrated signal is combined across the integration periods before passing it to the acquisition detector.

Acquisition detector: This module is used to detect the presence of the GPS signal. Noise and detection threshold computation are important part of the acquisition process. It computes the minimum noise level which the correlation peak should exceed to be detected as a signal. It should be optimally chosen to avoid a false lock and to allow weak signal acquisition. A signal is acquired when the correlation peak exceeds the detection threshold and estimates of the code phase and Doppler of the cell under search are passed to the tracking process. If a signal is not detected, the acquisition manager searches the next cell. Once all the cells are exhausted the next GPS satellite is searched and the process is repeated.

3. Acquisition schemes comparison

Time domain correlation, circular convolution and modified circular convolution were implemented in software to analyze the acquisition process. Time domain correlation performs a sequential cell by cell search and is time consuming for the software receiver implementation compared to other two methods. Hence only circular convolution and modified circular convolution methods are compared in this section. Time domain correlation is preferred for a hardware correlator because of its simplicity.

3.1 Details of data set collected and processing methodology

Digitized IF data is required to perform software acquisition and can be obtained by tapping data from a GPS RF front-end or by simulating the GPS signal in software and quantizing it. The GPS signal was simulated in software (using MATLAB) [4] and white noise was added to the signal. The signal bandwidth was kept at 2 MHz and sampled at different frequencies (4, 7, 9 and 12 MHz). These sampling frequencies were chosen at random to verify the proper functioning of acquisition methods. Each data set was generated for one second. Ten data sets were collected for each sampling frequency and thus a total of 40 data sets were collected.

Two different acquisition schemes were used to analyze the performance of acquisition and then combined to improve the acquisition performance. A combination of the MEX (C code compiled in Matlab) and Matlab code was used to reduce the processing times [4].

Table 1: Acquisition parameters used during analysis

Parameter	Values for single satellite data set
Intermediate Frequency (IF)	15.42 MHz
Sampling Frequency (SF)	4, 7, 9 and 12 MHz depending on data set
Start value of Doppler search	-5 KHz
End value of Doppler search	+5 KHz
Coherent integration time	8 ms
Non-coherent integration time	16 ms
False detection probability	5%
Number of PRNs to be searched	32

Table 1 lists the acquisition parameters used to perform the acquisition on the collected data sets.

The IF is at 15.42 MHz which was used to generate the local replica carrier signal. Different sampling frequencies were used to ensure proper functioning of the acquisition process. The acquisition manager uses the specified parameters to determine the Doppler bin using the coherent integration time. The correlation values are used to compute the noise and detection threshold.

4. Results

Acquisition was performed on all the single satellite data sets using both schemes to be verified. The acquisition results from all the data sets were analyzed in terms of the mean processing time, the acquisition gain and the memory required. The results from all the data sets were averaged to obtain an estimate of the above mentioned parameters. The single satellite results were verified with the simulator settings and were found to acquire at the correct Doppler. There were no false locks for the remaining 31 PRNs.

4.1 Mean Processing Time

The processing time was calculated using the time taken by the PC to perform the desired task. The PC used for the analysis was the Intel Pentium 4 processor operating at 2.0 GHz speed and Matlab version 6.5 [4] was used to code the acquisition algorithms. The processing times for all Doppler bins for an 8 ms coherent integration period at different sampling frequencies are shown in Table 2.

Table 2: Processing times for 8 ms coherent integration period

Acquisition scheme	Sampling frequency (time in seconds)			
	4 MHz	7 MHz	9 MHz	12 MHz
Circular convolution	10.00	12.93	15.76	19.33
Modified circular convolution	8.96	11.47	14.27	16.90

The modified circular convolution scheme takes less time than the circular convolution scheme because it uses a half of the GPS spectrum. This reduces the number of the FFT points and thus the FFT processing time. FFT is the most time consuming operation in a software receiver. The FFT and IFFT were performed in Matlab [4] and hence the processing times are in the order of seconds. The processing time increases with an increase in the sampling frequency since the number of samples (i.e. FFT points) is more at higher sampling frequencies for the same duration of time. The processing time also depends on the Doppler search range used for acquisition and increases linearly with an increase in the Doppler range as represented in Table 3.

The Doppler search range increases with an inaccurate receiver clock and high user dynamics. It can be reduced with knowledge of the satellite positions, an approximate GPS time and an approximate user position. Almanac and ephemeris data along with the GPS time can be used to compute the satellite positions. The user position in conjunction with the satellite position is used to compute an approximate code phase and Doppler for that satellite. The acquisition manager uses this information to reduce the search range and acquisition time.

Table 3: Processing time for different Doppler range

Acquisition schemes (time in seconds)							
Sampling frequency							
Circular convolution				Modified circular convolution			
4 MHz	7 MHz	9 MHz	12 MHz	4 MHz	7 MHz	9 MHz	12 MHz
10.00	12.93	15.76	19.33	8.96	11.47	14.27	16.90
14.42	27.01	38.13	50.01	11.70	22.17	34.42	40.82
20.10	38.00	55.00	73.72	16.12	32.06	46.16	61.12

4.2 Processing gain

Acquisition gain is an important factor to determine satellite acquisition. It was computed as a ratio of the correlation peak against the detection threshold. The acquisition schemes should provide as high gain as possible to acquire weak signals. The gains obtained for the two schemes at different signal strengths and sampling frequencies are shown in Table 4.

The gain is nearly the same for different sampling frequencies except for the 4 MHz sampling frequency. A sampling frequency of 4 MHz causes an aliasing effect which introduces a signal loss and results in lower gain. Acquisition gain from the modified circular convolution scheme is about 1-1.5 dB lower than the circular convolution method.

Table 4: Processing gain for 8 ms coherent integration period

Signal power level	Acquisition schemes (gain in dB)							
	Circular convolution				Modified circular convolution			
	Sampling frequency (MHz)				Sampling frequency (MHz)			
	4	7	9	12	4	7	9	12
-120 dBm	21.08	23.03	23.11	23.20	19.84	22.28	22.31	22.16
-125 dBm	16.62	19.62	19.62	19.62	15.75	18.75	18.75	18.75
-130 dBm	10.33	13.33	13.33	13.33	9.56	12.56	12.56	12.56

This is due to the use of half the input signal spectrum to reduce the processing time. The GPS signal information contained in the other half of the GPS spectrum is lost which results in a lower gain. Thus the reduction in the processing time is at the cost of lower gain.

4.3 Memory requirements

One important criterion for choosing the acquisition scheme to implement in an embedded system is the amount of memory required. Memory usage should be as minimal as possible to implement the algorithm across the microprocessors and a DSP where available memory is a constraint. Memory requirements were analyzed at two stages in both acquisition schemes. The first stage is the FFT stage wherein the FFT of the incoming signal and a local signal is taken. The memory locations needed for this stage at different sampling frequencies are given in Table 5. The next stage is the IFFT stage wherein the inverse FFT is taken of the signal resulting from convolution of the two spectrums. The memory locations needed for this stage at the different sampling frequencies are given in Table 6.

These memory requirements were obtained when each sample was stored in a separate memory location. These samples can be packed in bytes to reduce the memory requirements by a factor of eight. The memory required increases linearly with an increase in the coherent integration time. A higher sampling frequency requires more memory as the number of samples is more at higher frequencies for the same duration of time. Hence the coherent integration time and the sampling frequency should be chosen depending upon available system resources.

Table 5: Memory required for 1 ms coherent integration period at FFT stage of acquisition schemes

Acquisition scheme	Number of memory locations			
	Sampling frequency			
	4 MHz	**7 MHz**	**9 MHz**	**12 MHz**
Circular convolution	4000	7000	9000	12000
Modified circular convolution	4000	7000	9000	12000

Table 6: Memory required for 1 ms coherent integration period at IFFT stage of acquisition schemes

Acquisition scheme	Number of memory locations			
	Sampling frequency			
	4 MHz	**7 MHz**	**9 MHz**	**12 MHz**
Circular convolution	4000	7000	9000	12000
Modified circular convolution	2000	3500	4500	6000

4.4 Acquisition plots

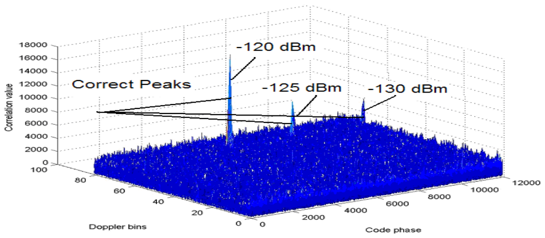
Figure 2 shows the autocorrelation plots (first eight) and the cross-correlation plots (last two) for the two acquisition schemes at different sampling frequencies (SF) and signal power levels.

The plots show that a correlation peak is generated when the phase of the PRN codes match during autocorrelation. Cross-correlation does not yield a peak as observed in the correlation plots. This correlation property of the GPS PRN codes allows proper acquisition of the GPS signal. The signal peak decreases with a decrease in the GPS signal strength which leads to a cross correlation problem for weak signal acquisition.

These results verify the two GPS acquisition schemes. The circular convolution scheme provides a better gain but at the cost of processing time and memory. The modified circular convolution scheme can be used to reduce acquisition time and memory requirements but the gain is less than that for circular convolution scheme. An intelligent acquisition scheme will be to first use the modified circular convolution scheme to acquire the signals with good signal strength in a less amount of time and later switch to the circular convolution scheme to acquire the signals with low signal strength. This was implemented in the software receiver and found to be effective in reducing processing time.

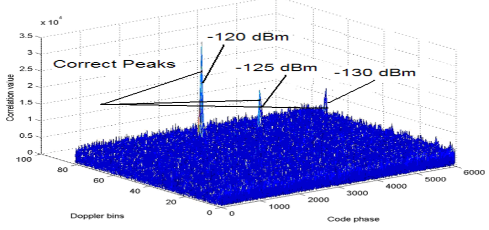
Circular convolution

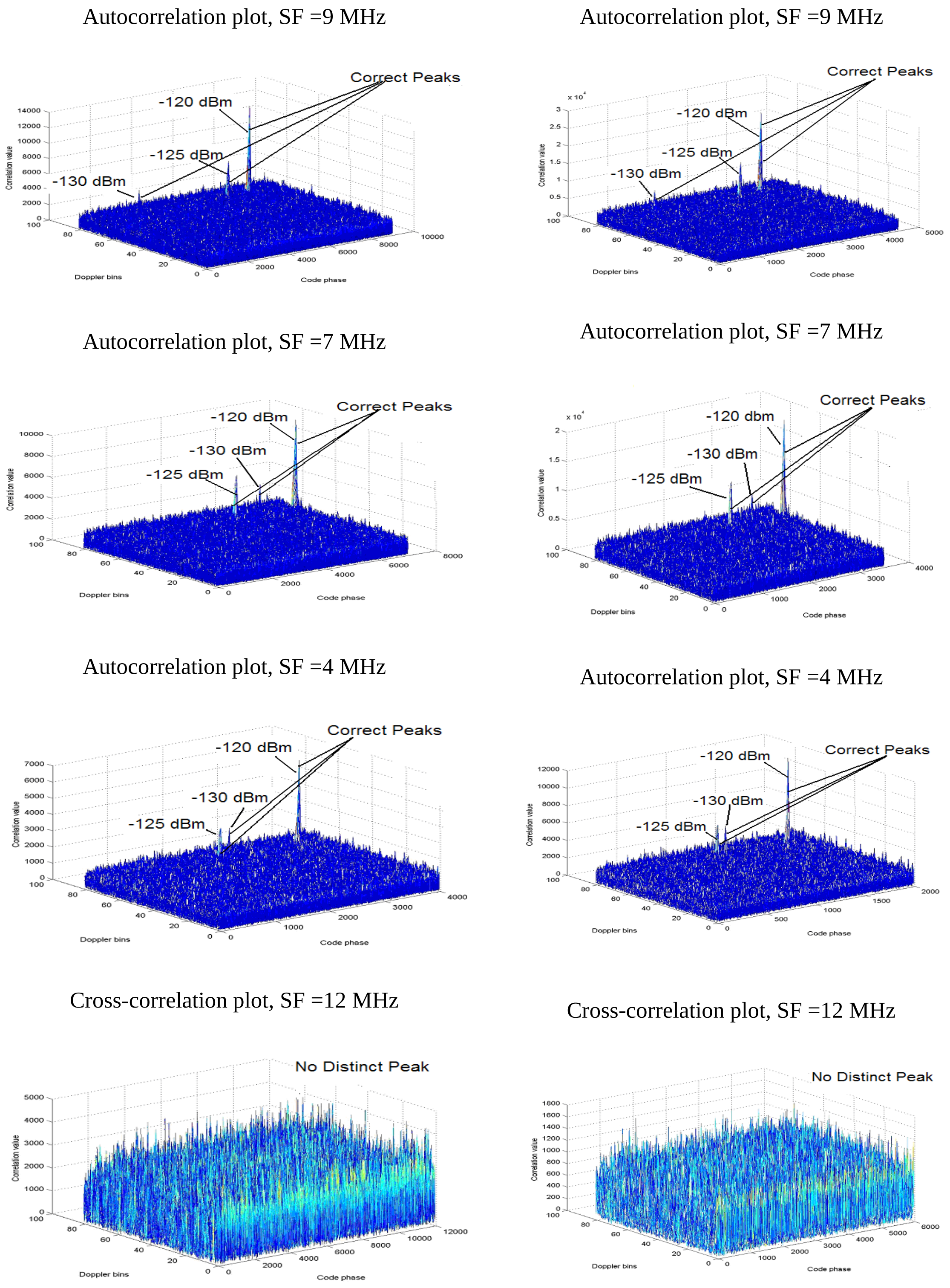
Autocorrelation plot, SF =12 MHz



Modified circular convolution

Autocorrelation plot, SF = 12 MHz





Autocorrelation plot, SF =9 MHz

Autocorrelation plot, SF =9 MHz

Autocorrelation plot, SF =7 MHz

Autocorrelation plot, SF =7 MHz

Autocorrelation plot, SF =4 MHz

Autocorrelation plot, SF =4 MHz

Cross-correlation plot, SF =12 MHz

Cross-correlation plot, SF =12 MHz

Figure 2: Correlation plots for two different acquisition scheme

5. Conclusion

This research investigated the effect of various sampling frequencies on processing speed for GPS signal acquisition. The acquisition schemes were implemented and used to compare different figures of merit for GPS signal acquisition. The conclusion that can be drawn from the result of the research is that processing time increases exponentially with higher sampling frequencies. The modified circular convolution has 50% less processing time for a coherent integration time above 10 ms compared to the other method.

The circular convolution scheme provides about 1.5 dB more gain than modified circular convolution which allows acquisition of weaker signals.

6. Acknowledgement

Every work needs to be planned and executed properly for its success. It gives us immense pleasure to acknowledge our gratitude to all those persons who have been a great source of inspiration. Though it is impossible to give individual thanks to all faculty personnel, we take this opportunity to express our gratitude to them.

We honestly express our thanks to Mr. Sameet Deshpande, System Engineer, Texas Instruments, Bengaluru, India for assisting us on all information to carry out our project work, for always providing valuable suggestions and clarifications whenever needed.

7. References

[1] Kaplan E.D. (1996), *Understanding GPS: Principles and Applications*, Artech House Inc., Norwood, MA.

[2] Tsui Y. and J. Bao (2000), *Fundamentals of Global Positioning System Receivers: A Software Approach*, John Wiley & Sons Inc., New York, NY.

[3] VanNee D.J.R. and A.J.R.M. Conen (1991), *New Fast GPS code acquisition technique using FFT*, IEEE Electronic letters, Vol. 27, No, 2, pp. 158-160.

[4] MATLAB 6.5- the language of technical computing

Analysis of Cooperative Relay-Based Energy Detection of Unknown Deterministic Signals in Cognitive Radio Networks

O. Olabiyi, A. Annamalai

Center of Excellence for Communication Systems Technology Research
Department of Electrical and Computer Engineering
Prairie View A&M University, TX 77446

Abstract— **In this article, we develop an analytical framework for the performance analysis of cooperative amplify-and-forward (CAF) relay-based energy detection. We derive mathematical expression for the detection probability for single cognitive relay system and later extend it to multi-relay system. Our resulting expression is based on the canonical series representation of generalized Marcum Q-function of real order in conjunction with the derivatives of moment generating function (MGF) of signal-to-noise ratio of fading channels. First, we considered single relay based sensing and compared to direct sensing. We found out that the direct sensing performs better than the single relay sensing as expected. Then, combining both direct sensing and relay-based sensing using either of maximum ratio combining (MRC) and square-law combining (SLC) to form a CAF relay-based sensing system greatly enhance the performance. Also, the performance of multi-relay diversity system is analysed; and as expected, increase in relay diversity improves the sensing performance. In all diversity combining cases, we found out that CAF relay sensing based on MRC outperforms the ones based on SLC.**

Keywords— **Energy detection, Cognitive radio, Fading channels, Maximum ratio combining, Square-law combining, Cooperative amplify-and-forward.**

1 Introduction

The emerging technology of cognitive radio has created a paradigm shift in the design of wireless system where radio can now adapt their operating behaviour to take advantage of unused spectrum. One of the main requirements of this system is to ensure that the incumbent (i.e. primary or licensed user) is not interrupted by the activity of these cognitive radios. Therefore the new radio must be capable of determining the presence or absence of an incumbent before spectrum usage. This challenge of identifying unused spectrum has become fascinating topic within research community and is an integral part of IEEE802.22 standard for cognitive radio system. Energy detection has been identified as one of the prospective solutions to this problem due to the simplicity of its implementation.

Urkowitz, [1] first studied the detection of an unknown deterministic signal over a flat band-limited Gaussian noise channel using an energy detector. Since then, lots of work has been published on performance of energy detectors in single channel. [2]- [4] extended the results in [1] to single receiver system in Rayleigh, Rician and Nakagami-m fading channels. The performance of energy detector in diversity system such as Maximal ratio combining (MRC), selection, switch and stay (SSC), equal gain combining (EGC), square-law combining (SLC) and square-law combining (SLC) diversity detectors are considered in [3]-[8] over various fading channels for non-cooperative detection system with each article providing new insight on the characterization and analysis of performance of energy detectors. In [13], we presented new approach based on alternative form of Marcum Q-function and the derivatives of moment generating function (MGF) of SNR. All the mentioned articles only focus on non-cooperative system. In [12], the Authors focused on cooperative spectrum sensing in which the cooperative cognitive radios either makes a binary decision based on local observation or just send the observation value to the common receiver. Most of the articles in cooperative sensing are based on this approach which is basically a hard decision or data fusion problem.

However, recently, [14], [16], and [17] presented a new perspective to cooperative spectrum sensing in which the cognitive radio do not make any measurement or decision, each of them only amplify its received signal in a specific bandwidth and forward it to the fusion center or the cluster head, an operation that completely depicts the non-regenerative amplify-and-forward (AF) relay system. In [16], complementary area under receiver operating curve (AUC) is considered while [14] and [17] focused on the receiver operating characteristics (ROC) of energy detector over cognitive relay system but their analysis is limited to independent and identically distributed (i.i.d) Rayleigh fading channel and its composite fading. In [14], [16] and [17], the Authors claimed that it is analytically difficult to use the "exact" MGF expression in their analysis and therefore resolve to using the MGF of upper bound of harmonic mean signal-to-ratio (SNR) of the relay part of CAF relay system and then use 'exponential-type' contour integral to derive the expression for the detection probability.

The main motivation for this work is to develop an analytical analysis of detection performance of CAF cognitive system. The derived framework is based on the canonical series expression for the detection probability in conjunction with derivatives of MGF of SNR of fading channel for single and multiple relay system. This framework is general and applicable to any fading channel if the kth order derivative of MGF of harmonic mean SNR can be evaluated. For the purpose of brevity we limit our analysis to independent and non-identically distributed (i.n.d) Rayleigh and i.i.d Nakagami-m channel using closed form "exact" expression for the MGF of SNR. This approach greatly simplifies and modularizes the computation of detection probability and to the best of our knowledge; this area of application has not been considered in earlier works. The rest of the paper is organized as follows. In Section II we present the system model and in Section III derived the performance expression for single cognitive relay system. In Section IV we extended this to multi-relay system using MRC and SLC diversity system with or without direct sensing and finally, numerical results and concluding remarks are presented in section V and VI respectively.

2 System model

To be consistent, notations similar to [8] are used as listed below.

$x(t)$	: Unknown deterministic signal waveform
$\eta(t)$	: Noise waveform – White Gaussian random process
s_i	: Unknown deterministic signal waveform
n_i	: Noise waveform – White Gaussian random process
$r(t)$	: Received signal
h	: Channel coefficient amplitude
T	: Observation time interval
W	: One-sided bandwidth
$u = TW$	: Time-Bandwidth product
N_{01}	: One sided noise power spectral density
E_s	: Signal energy over the time interval T
λ	: Energy threshold of the receiver
L	: Number of branches of the receiver combiner
H_0	: Hypothesis 0; no $x(t)$ present
H_1	: Hypothesis 1; $x(t)$ present
χ^2_{2u}	: Central Chi-square distribution with $2u$ degrees of freedom
$\chi^2_{2u}(\epsilon)$	: Non central Chi-square distribution with $2u$ degrees of freedom and non centrality parameter ϵ

Considering a cognitive radio network shown Fig. 1 which consists of a network cluster with N number of cognitive radio (secondary user) nodes, a primary user (PU) and a cluster head (CH) otherwise referred to as cognitive coordinator in [16]. Here, the cluster head serves as the fusion centre and cluster resource manager similar to the architecture in ad hoc wireless networks. However, the cluster head could either be a stationary centralized node as in infrastructure based cognitive radio networks or a distributed coordination as in traditional wireless sensor networks where any the cognitive nodes can become the cluster head.

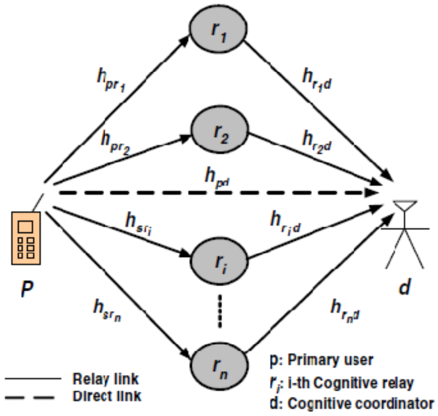


Fig. 1 Illustration of cooperative communication in cognitive radio networks [16].

In this system, the cognitive nodes serve as cooperative relaying node by re-transmitting the PU signal over orthogonal channels to the CH. The main advantage of this is to increase the detection range of a cluster and to afford the cognitive node (CN) more time for transmission since the CNs doest not need to process the data locally. This will eventually leads to increase in the system throughput and longer battery life for the CNs. In our model, the CH can either partake in the sensing activity or not depending on the location.

Considering a single relay system, with a cognitive node, a primary user and the cluster head serving as the fusion center, let $h_{p,d}$ be the channel gain between the PU and CH and $h_{p,i}$ be the channel gain between the PU and the CN i, the received primary user signal at the CN i and CH, d is given by [15]

$$y_{p,d} = \sqrt{P_{p,d}}\, h_{p,d} x + \eta_{p,d} \tag{1}$$

$$y_{p,i} = \sqrt{P_{p,i}}\, h_{p,i} x + \eta_{p,i} \tag{2}$$

where $P_{p,d} = P_{p,i}$ is the transmitted signal power of the PU and $\eta_{p,d}$, $\eta_{p,i}$ are the additive noises introduced between the PU and CH, and PU and CN respectively. Therefore, the relayed signal from the CN to CH is given by

$$y_{i,d} = g_i h_{i,d} y_{p,i} + \eta_{i,d} \tag{3}$$

where g_i is the amplifier gain of the ith CN and $\eta_{i,d}$ is the additive noise introduced between the CN and CH . Since the CN serve as a variable gain amplify-and-forward (AF) relay,

$$g_i = \sqrt{P_{i,d}} \Big/ \sqrt{P_{p,i} \left|h_{p,i}\right|^2 + N_0} \tag{4}$$

where $P_{i,d}$ is the transmit power between the CN and CH and N_0 is spectral density of the additive noise. Eq. (3) could then be expressed as

$$\begin{aligned} y_{i,d} &= g_i \sqrt{P_{p,i}}\, h_{p,i} h_{i,d} x + g_i h_{i,d} \eta_{p,i} + \eta_{i,d} \\ &= \sqrt{P_{p,i}}\, h x + \eta \end{aligned} \tag{5}$$

where $h = g_i h_{p,i} h_{i,d}$ is the effective gain between the PU the CH and $\eta = g_i h_{i,d} \eta_{p,i} + \eta_{i,d}$ is the effective noise at the CH. Therefore the effective SNR at the receiver of the CH is given by [15]

$$\gamma = \gamma_{p,d} + \frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}+1} \simeq \gamma_{p,d} + \frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}} = \gamma_{p,d} + \gamma_i \quad (6)$$

where $\gamma_{p,d} = \left|h_{p,d}\right|^2 P_{p,d} / N_o$, $\gamma_{p,i} = \left|h_{p,i}\right|^2 P_{p,i} / N_o$ and $\gamma_{i,d} = \left|h_{i,d}\right|^2 P_{i,d} / N_o$ are the SNR of the PU to CH direct link, PU to CN link and CN to CH link respectively.

The detection of the existence of the unknown deterministic signal $x(t)$ by the receiver, is a binary hypothesis test and can be expressed as in [13, eq(1)],

$$y(t) = \begin{cases} \eta(t) & : H_0 \quad P_{s,i} = P_{s,d} = 0 \\ hx(t)+\eta(t) & : H_1 \quad P_{s,i} \neq 0, P_{s,d} \neq 0 \end{cases} \quad (7)$$

Without going into the detailed derivation which has been well treated in [13], the decision variable Y under H_0 is a square sum $2u$ Gaussian random variable of $N(0, g_i h_{i,d}+1)$ and follows χ^2_{2u}. Similarly, Y (the channel coefficient amplitude) under H_1 is $\chi^2_{2u}(\epsilon)$, where ϵ is given by $\epsilon = 2\gamma$. Therefore, the detection and false alarm probabilities of the energy detector in AWGN channel are given by:

$$P_d = Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) \quad (8)$$

and

$$P_f = \frac{\Gamma(u, \frac{\lambda}{2})}{\Gamma(u)} \quad (9)$$

respectively, where $Q_u(.,.)$ is the generalised (u^{th} order) Marcum-Q function and $\Gamma(.,.)$ is the upper incomplete gamma function . P_f is the same over any fading channel since there is no γ in (9). In the other sense P_d has to be averaged over different fading channels and diversity combining.

3 Average detection probability over single relay system

The average detection probability over any fading channel is given by [13]

$$\overline{P_d} = \int_0^\infty Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) f(\gamma) d\gamma \, . \quad (10)$$

Using the alternative form of Marcum Q function in [13, (7)] the solution to (10) has been generalized in [13] as

$$\overline{P}_{d\ Gen} = 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k, \frac{\lambda}{2})}{\Gamma(u+k)} \phi_\gamma^{(k)}(s) \Bigg|_{s=1} \quad (11)$$

where $\phi_\gamma^{(k)}(s) = \frac{\partial^k \phi_\gamma}{\partial s^k}$ and $\phi_\gamma(s)$ is the moment generating function (MGF) of different stochastic fading channels and $G(.,.)$ is the lower incomplete gamma function which is defined by $G(a,x) = \int_0^x t^{a-1} e^{-t} dt$. The convergence of this infinite series has been well treated in [13] and it's been shown that few terms are required for four digit accuracy.

In order to compute (11), we need to find the k^{th} derivative of the MGF of the relay part of the SNR, γ_i given in (6). The closed form "exact" MGF of γ_i only exist in literature for i.n.d Rayleigh and i.i.d Nakagami-m fading channels and are given respectively by [21, eq. (20)], [20, eq, (26)]

$$\phi_{\gamma_i}(s) = \frac{16}{3\Omega_{p,i}\Omega_{i,d}(A_1+s)^2} \left[\frac{4\left(\frac{1}{\Omega_{p,i}} + \frac{1}{\Omega_{i,d}}\right)}{(A_1+s)} {}_2F_1\left(3, \tfrac{3}{2}; \tfrac{5}{2}; \tfrac{A_2+s}{A_1+s}\right) + {}_2F_1\left(2, \tfrac{1}{2}; \tfrac{5}{2}; \tfrac{A_2+s}{A_1+s}\right) \right]$$

(12)

where $A_1 = \frac{1}{\Omega_{p,i}} + \frac{1}{\Omega_{i,d}} + \frac{2}{\sqrt{\Omega_{p,i}\Omega_{i,d}}}$, $A_2 = \frac{1}{\Omega_{p,i}} + \frac{1}{\Omega_{i,d}} - \frac{2}{\sqrt{\Omega_{p,i}\Omega_{i,d}}}$, $\Omega_{p,i}$ and $\Omega_{i,d}$ are the average SNR of the PU to CN and CN to CH links respectively, and

$$\phi_{\gamma_i}(s) = {}_2F_1\left(m, 2m; m+\frac{1}{2}; \frac{-s\Omega_i}{4m}\right) \quad (13)$$

where m is the Nakagami-n fading index, and ${}_2F_1\left(.,.;.;.\right)$ is the Gauss hypergeometric function. The alternative expression of (12) could also be found in [19, (7)].

The k^{th} derivative of (12) and (13) could be obtained easily using the identity in [21], [10, 0.430-1], Appendix A and the identity $\frac{\partial^n}{\partial s^n}(A+s)^{-a} = (-1)^n \left(a\right)_n (A+s)^{-(a+n)}$ where $(a)_n = a(a+1)...(a+n-1)$ denotes the Pochhammer symbol. After few algebraic manipulations we obtain,

$$\begin{aligned} \overline{P}_{d\ Ray} = 1 - \sum_{k=0}^{\infty} \frac{1}{k!} \frac{G(u+k, \frac{\lambda}{2})}{\Gamma(u+k)} \Bigg[& \frac{64(\Omega_{p,i}+\Omega_{i,d})}{3\left(\Omega_{p,i}\Omega_{i,d}\right)^2} \sum_{n=0}^{k} \binom{k}{n} \left((3)_{k-n}(A_1+1)^{-(k-n+3)}\right) \\ & x \sum_{r=1}^{n} \frac{U_r}{r!} \frac{(3)_r (\frac{3}{2})_r}{(\frac{5}{2})_r} {}_2F_1\left(3+r, \tfrac{3}{2}+r; \tfrac{5}{2}+r; \tfrac{A_2+1}{A_1+1}\right) \\ & + \frac{16}{3\left(\Omega_{p,i}\Omega_{i,d}\right)} \sum_{n=0}^{k} \binom{k}{n} \left((2)_{k-n}(A_1+1)^{-(k-n+2)}\right) \\ & x \sum_{r=1}^{n} \frac{U_r}{r!} \frac{(2)_r (\frac{1}{2})_r}{(\frac{5}{2})_r} {}_2F_1\left(2+r, \tfrac{1}{2}+r; \tfrac{5}{2}+r; \tfrac{A_2+1}{A_1+1}\right) \Bigg] \end{aligned} \quad (14)$$

where $U_r = \sum_{p=0}^{r-1} (-1)^p \binom{r}{p} \left(\frac{(A_2+1)}{(A_1+1)}\right)^p \sum_{\substack{m=0 \\ m \le r-p}}^{n} \binom{n}{m} \frac{(r-p)!}{(r-p-m)!}$

$$x(A_2+1)^{(r-p-m)}(-1)^{n-m}(r-p)_{n-m}(A_1+1)^{-(r-p+n-m)}$$

$$\begin{aligned} \overline{P}_{d\ Nak} = 1 - \sum_{k=0}^{\infty} \frac{1}{k!} \frac{G(u+k, \frac{\lambda}{2})}{\Gamma(u+k)} \left(\frac{\Omega_i}{4m}\right)^k \frac{\left(m\right)_k \left(2m\right)_k}{\left(m+\frac{1}{2}\right)_k} \\ x \, {}_2F_1\left(m+k, 2m+k; m+\frac{1}{2}+k; \frac{-\Omega_i}{4m}\right) \end{aligned} \quad (15)$$

The performance of single cognitive relay detection is compared to direct sensing and the result is shown Fig. 2 and 3 for i.i.d Nakagami-m channel. Fig. 2 is the complementary ROC curve showing $P_m = 1 - P_d$ against P_f while Fig.3 shows P_d against the average SNR. Both figures show that direct sensing out- performs the relay sensing as expected (similar pattern was observed for symbol error rates [18]) but the motivation for practical implementation is the increase in the detection range or coverage and reliability of energy detection. We will show later that combining cognitive relay detection with direct detection to form CAF relay based sensing will greatly improve the performance.

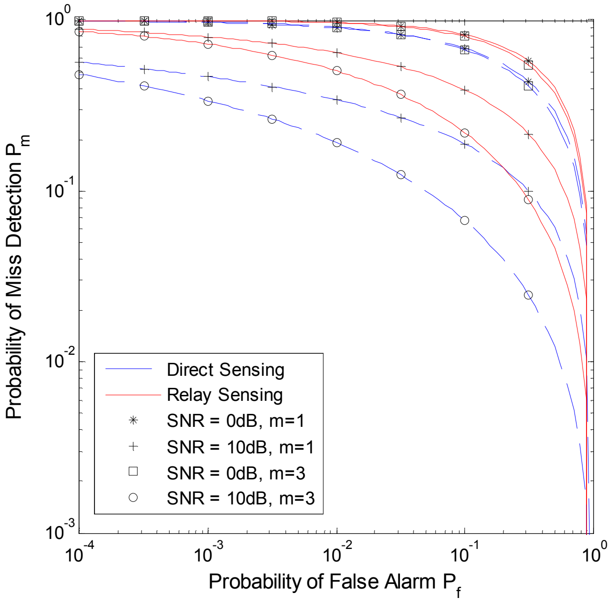


Fig. 2. Comparison of complementary ROC performance of direct and CAF relay based sensing with *u*=1.

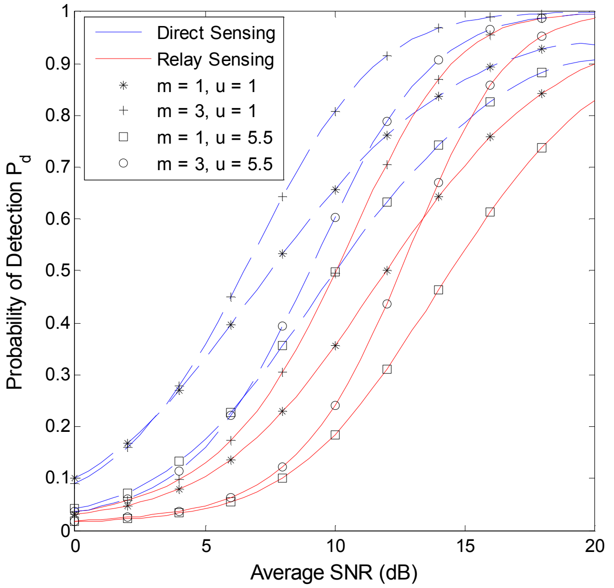


Fig. 3. Comparison of detection probability with average SNR (P_f =0.01).

4 Detection over multi-relay system

In the case of N cognitive relay nodes, the CH coordinates the PU signal detection. We assume here that the CNs have the knowledge of the particular bandwidth to be detected. Also, the communication between the CNs and CH is through orthogonal channels either through TDMA or CDMA techniques. The CNs simultaneously amplify the signal of the predetermined bandwidth and forward it to the CH to make the sensing decision. The received signal at the CH could be combined either before or after detection resulting into either of maximum ratio combining (MRC) or square law combining (SLC) respectively. Practical implementation of MRC is difficult to achieve as the CH receiver requires the complete channel state information (CSI) of each diversity branch in order to achieve coherent detection. This is not required in the case of SLC as the signals are combined after individual non-coherent detection leading to twice the degree of freedom of the MRC and ultimately reduce the detection probability compared to MRC.

4.1 Maximum ratio combining (MRC)

The output SNR, γ_{MRC} of the MRC combiner is the sum of all the SNRs on all branches and it is given by

$$\gamma_{MRC} = \gamma_{p,d} + \sum_{i=1}^{N} \frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}+1} \simeq \gamma_{p,d} + \sum_{i=1}^{N} \frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}} \qquad (16)$$

where N is the number of CNs. The decision variable Y_{MRC} is i.i.d χ^2_{2u} for H_0 and $\chi^2_{2u}(\epsilon_j)$ for H_1 and is defined by

$$Y_{MRC} \sim \begin{cases} \chi^2_{2u} & : H_0 \\ \chi^2_{2u}(\epsilon_{MRC}) & : H_1 \end{cases} \qquad (17)$$

where ϵ_{MRC} is the centrality parameter given by $2\gamma_{MRC}$. Therefore, the P_d at the MRC output for AWGN channels can be evaluated as

$$P_{d,MRC} = Q_u(\sqrt{2\gamma_{MRC}}, \sqrt{\lambda}) = 1 - \sum_{k=0}^{\infty} \frac{\gamma^k e^{-\gamma}}{k!} \frac{G(u+k, \frac{\lambda}{2})}{\Gamma(u+k)} \qquad (18)$$

The MGF of the combined output could be derived from (16) and it is expressed as

$$\phi_{MRC} = \phi_{p,d} \prod_{i=i}^{N} \phi_i \qquad (19)$$

The derivative of (19) can obtained from Appendix A and substituting this into (11), we obtain

$$\overline{P_{d\ MRC}} = 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \sum_{n_0=0}^{k} \sum_{n_1=0}^{n_0} \sum_{n_2=0}^{n_2} \cdots \sum_{n_{N-1}=0}^{n_{N-2}} \binom{k}{n_0}\binom{n_0}{n_1}\binom{n_1}{n_2}\cdots$$
$$X\binom{n_{N-2}}{n_{N-1}} \phi_0^{(k-n_0)} \phi_1^{(n_0-n_1)}(s) \phi_2^{(n1-n2)}(s) \cdots \phi_N^{(n_{N-1})}(s) \bigg|_{s=1} \qquad (20)$$

where $\phi_0(s) = \phi_{p,d}(s)$ and its k^{th} order derivative are listed in [13, Table 1] for different fading channels. $\phi_{1\ldots N}(s)$ is the MGF of the relay part and the k^{th} order derivative of any of them can be either of (14) or (15) as the case maybe. For example, the average P_d for i.i.d Nakagami-m can be expressed as

$$\overline{P_{d\ MRC-Nak}} = 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \sum_{n_0=0}^{k} \sum_{n_1=0}^{n_0} \sum_{n_2=0}^{n_2} \cdots \sum_{n_{N-1}=0}^{n_{N-2}} \binom{k}{n_0}\binom{n_0}{n_1}\binom{n_1}{n_2}\ldots\binom{n_N-2}{n_N-1}$$
$$\times \frac{(-\Omega)^{(k-n_0)} m^m \Gamma(m+(k-n_0))}{(m+\Omega)^{m+(k-n_0)} \Gamma((m)} \left(\frac{-\Omega_1}{4m}\right)^{(n_0-n_1)} \frac{(m)_{(n_0-n_1)} (2m)_{(n_0-n_1)}}{(m+\frac{1}{2})_{(n_0-n_1)}}$$
$$\times {}_2F_1\left(m+(n_0-n_1), 2m+(n_0-n_1); m+\frac{1}{2}+(n_0-n_1); \frac{-\Omega_2}{4m}\right)\ldots \qquad (21)$$
$$\times \left(\frac{-\Omega_i}{4m}\right)^{(n_N-1)} \frac{(m)_{(n_N-1)} (2m)_{(n_N-1)}}{(m+\frac{1}{2})_{(n_N-1)}} \ldots$$
$$\times {}_2F_1\left(m+(n_N-1), 2m+(n_N-1); m+\frac{1}{2}+(n_N-1); \frac{-\Omega_N}{4m}\right)$$

and the false alarm probability $P_{f,MRC}$, still remainings as in (9).

4.2 Square-law combining (SLC)

Since the output decision is combined in this scheme, the decision variable Y_{SLC} is the sum of $N+1$ i.i.d χ^2_{2u} for H_0 and $\chi^2_{2u}(\epsilon_j)$ for H_1 and is defined by [8]

$$Y_{SLC}=\sum_{i=0}^{N}Y_i\sim\begin{cases}\chi^2_{2(N+1)u} & :H_0\\ \chi^2_{2(N+1)u}(\epsilon_{SLC}) & :H_1\end{cases}\qquad(22)$$

where the non-centrality $\epsilon_{SLC}=\sum_{i=0}^{N}\epsilon_i=\sum_{i=0}^{N}2\gamma_i=2\gamma_{SLC}$, hence the false alarm and the detection probabilities can be expressed as

$$P_{f,SLC}=\frac{\Gamma((N+1)u,\frac{\lambda}{2})}{\Gamma((N+1)u)}\qquad(23)$$

$$P_{d,SLC}=Q_{(N+1)u}(\sqrt{2\gamma_{SLC}},\sqrt{\lambda})=1-\sum_{k=0}^{\infty}\frac{\gamma^k e^{-\gamma}}{k!}\frac{G((N+1)u+k,\frac{\lambda}{2})}{\Gamma((N+1)u+k)}\qquad(24)$$

Since $\gamma_{SLC}=\gamma_{MRC}=\gamma_{p,d}+\sum_{i=1}^{N}\frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}+1}\simeq\gamma_{p,d}+\sum_{i=1}^{N}\frac{\gamma_{p,i}\gamma_{i,d}}{\gamma_{p,i}+\gamma_{i,d}}$, the MGF of SNR at the output of SLC combiner is the same as (19) and therefore the average P_d is given by

$$\overline{P_d}_{SLC}=1-\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\frac{G((N+1)u+k,\frac{\lambda}{2})}{\Gamma((N+1)u+k)}\sum_{n_0=0}^{k}\sum_{n_1=0}^{n_0}\sum_{n_2=0}^{n_2}\cdots\sum_{n_{N-1}=0}^{n_{N-2}}\binom{k}{n_0}\binom{n_0}{n_1}\binom{n_1}{n_2}\cdots\qquad(25)$$

$$X\binom{n_{N-2}}{n_{N-1}}\phi_0^{(k-n_0)}\phi_1^{(n_0-n_1)}(s)\phi_2^{(n1-n2)}(s)\cdots\phi_N^{(n_{N-1})}(s)$$

In the special case of Nakagami-m i.i.d channels we obtain similar to (21)

$$\overline{P_d}_{SLC}=1-\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\frac{G((N+1)u+k,\frac{\lambda}{2})}{\Gamma((N+1)u+k)}\left[\phi_0(s)\phi_i^N(s)\right]^{(k)}\Big|_{s=1}\ .(26)$$

5 Numerical results

Here we analyse the performance of energy detector using the complementary ROC curves for each of MRC and SLC diversity techniques and later compare their detection probabilities at different SNR. Unless other wise stated the following are the values of parameter used in our simulations: for the complementary ROC, u=1.5, m=2.5 and mean SNR = 5dB and for detection probability against the mean SNR, m=2, u=1 and P_f =0.01. Fig. 4 shows the performance of MRC diversity of CR with and without direct sensing. It is evident from the figure that CAF relay-based sensing (combined relay and direct sensing) is better than just relay-based sensing. Also, it is observed that as the number of cooperating cognitive relay increases, the detection accuracy is increased. The figure also, shows the capability of our proposed method in handling non-integer u and m. Although not shown, similar performance is observed for the case of SLC diversity. Fig. 5 compares the performance of SLC to MRC and it shows that MRC outperforms SLC with and without direct sensing as expected due to higher degree of freedom of SLC diversity system. However, marginal contribution of additional SLC branch to the sensing is approximately the same as that of the MRC system. Therefore, we recommend SLC as a viable option to MRC as MRC implementation might be very practically impossible to achieve.

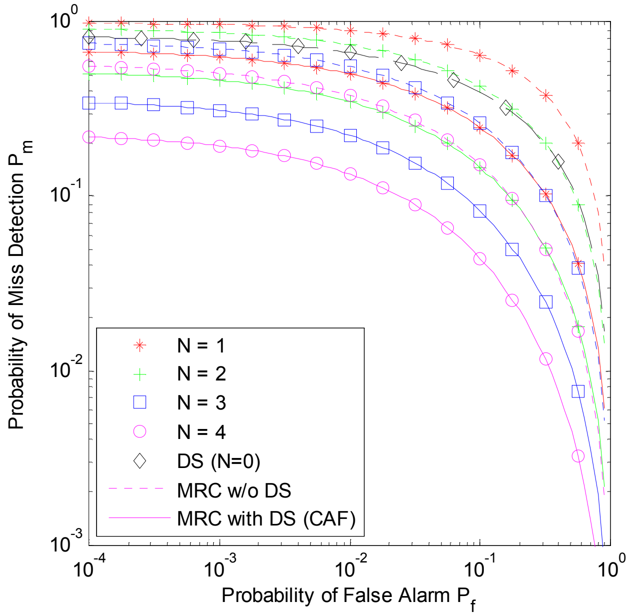


Fig. 4. Complementary ROC curves for MRC diversity, N= {0, 1, 2, 3, 4} with and without direct sensing in Nakagami-m Channel.

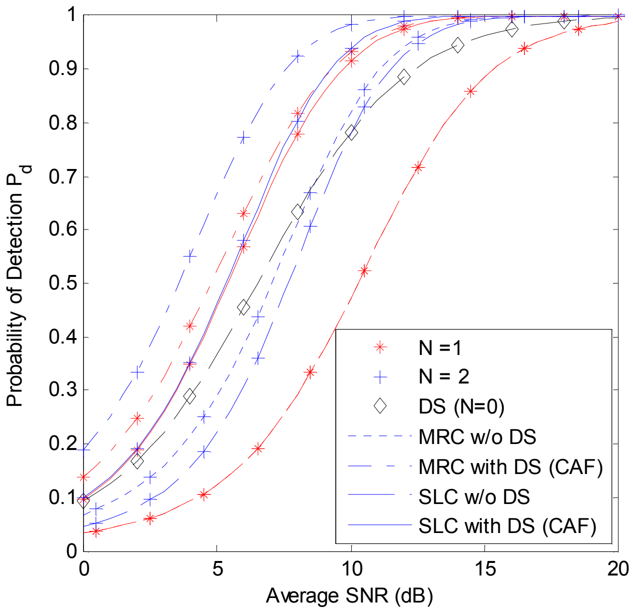


Fig. 5. Performance of SLC and MRC diversity systems, N={0, 1, 2} with and without direct sensing.

6 Conclusions

In this paper, we provide analytical framework for performance analysis of energy detector in cognitive CAF relay system with numerical results on Rayleigh and Nakagami-m channel. The mathematical expressions are based on canonical series representation of Marcum Q-function in conjunction with derivatives of MGF of SNR of fading channel. In our numerical analysis we have used “exact” MGF expression for the harmonic mean SNR unlike [14] and [17] that used MGF of the SNR bound and limited to Rayleigh fading channel. It is worth to mention here that our method is able to consider fractional m and half integer u and

Lu and we could revealed after careful consideration of related publications, that this approach has never been applied in the performance analysis of the CAF based energy detector in literature until now.

We have also shown that MRC diversity performs better that the SLC diversity as expected but there is no different in marginal contribution of additional diversity branch for both diversity schemes. This work fundamentally address the performance expected of CAF cognitive relay sensing with or without directing sensing. Our result could be easily used in deciding the number of diversity branches and the energy threshold value required to achieve a specified false alarm rate for different scenario of energy detector receiver in CAF cognitive relay system.

Acknowledgment

This work is supported in part by funding from the US Army Research Office (W911NF-04-2-0054), Clarkson Aerospace, and National Science Foundation (0931679 & 1040207).

Appendix A

In this appendix, a general method is outlined for treating the analysis of fading channels with dissimilar statistics. Let us consider an MGF of the combiner's output SNR that is written in a product form:

$$\phi_\gamma(s) = \prod_{i=1}^{L} \phi_i(s) \tag{A.1}$$

Where $\phi_\gamma(s)$ denotes the MGF of the SNR of the ith diversity branch. Our aim is to derive $\phi_\gamma^{(k)}(s)$ in closed form, given that the derivatives $\phi_i^{(k)}(s)$ are known. By applying Leibnitz's differentiation rule [10, eq. (0.42)]

$$\frac{d^k}{ds}(uv) = \sum_{n=0}^{k} \binom{k}{n} \frac{d^n}{ds}(v) \frac{d^{k-n}}{ds}(u) \tag{A.2}$$

recursively in eq. (A.1) leads to

$$\begin{aligned}\phi_\gamma^{(k)}(s) &= \sum_{n_1=0}^{k} \binom{k}{n_1} \phi_1^{(k-n_1)}(s) \sum_{n_2=0}^{n_1} \binom{n_1}{n_2} \phi_2^{(n_1-n_2)}(s) \ldots \sum_{n_{L-1}=0}^{n_{L-2}} \binom{n_{L-2}}{n_{L-1}} \phi_{L-1}^{(n_{L-2}-n_{L-1})} \phi_L^{(n_{L-1})}(s) \\ &= \sum_{n_1=0}^{k} \sum_{n_2=0}^{n_1} \ldots \sum_{n_{L-1}=0}^{n_{L-2}} \binom{k}{n_1}\binom{n_1}{n_2} \ldots \binom{n_{L-2}}{n_{L-1}} \phi_1^{(k-n_1)}(s) \phi_2^{(n_1-n_2)}(s) \ldots \phi_{L-1}^{(n_{L-2}-n_{L-1})} \phi_L^{(n_{L-1})}(s)\end{aligned} \tag{A.3}$$

which for the important practical cases of $L = 2$ and $L = 3$, Eq. (A-3) reduces to

$$\frac{d^k}{ds}\left[\prod_{i=1}^{2} \phi_i(s)\right] = \sum_{n_1=0}^{k} \binom{k}{n_1} \phi_1^{(k-n_1)}(s) \phi_2^{(n_1)}(s) \tag{A.4}$$

$$\frac{d^k}{ds}\left[\prod_{i=1}^{3} \phi_i(s)\right] = \sum_{n_1=0}^{k} \sum_{n_2=0}^{n_1} \binom{k}{n_1}\binom{n_1}{n_2} \phi_1^{(k-n_1)}(s) \phi_2^{(n_1-n_2)}(s) \phi_3^{(n_2)}(s) \tag{A.5}$$

From the above, it is apparent that only the k^{th} order derivative of the MGF of the SNR for a single channel reception is required to derive closed-form expressions.

References

[1] H. Urkowitz, "Energy detection of Unknown deterministic Signals," in *Proc IEEE,* vol. 55, no. 4, pp. 523-531, Apr 1967.

[2] V.I. Kostylev, "Energy detection of a signal with random amplitude," *IEEE Int. Conf. ICC* 2002, vol. 3, pp. 1606 – 1610, Apr-May 2002.

[3] Fadel F. Digham, Mohamed-Slim Alouni and Marvin K. Simon, "On the Energy Detection of Unknown Signals Over Fading Channels," *IEEE Int.Conf. ICC'03*, vol 5, pp. 3575 - 3579, May 2003.

[4] Fadel F. Digham, Mohamed-Slim Alouni and Marvin K. Simon, "On the Energy Detection of Unknown Signals Over Fading Channels," IEEE Trans. Commun., vol 55, no.1,pp.21-24, Jan. 2007.

[5] S.P. Herath, N. Rajatheva, C. Tellambura,"On the energy detection of unknown deterministic signal over Nakagami channels with selection combining," *IEEE CCECE'09,*2009, pp. 745–749, May 2009.

[6] S.P. Herath, N. Rajatheva, C. Tellambura, "Unified Approach for Energy Detection of Unknown Deterministic Signal in Cognitive Radio Over Fading Channels" in *Proc. IEEE ICC'09*, 2009, pp. 1-5, June 2005

[7] S.P. Herath, N. Rajatheva, "Analysis of Equal Gain Combining in Energy Detection for Cognitive Radio over Nakagami Channels," *IEEE GLOBCOM'08,* 2008, pp. 1-5, Dec. 2008.

[8] S.P. Herath, N. Rajatheva, "Analysis of Diversity Combining in Energy Detection for Cognitive Radio over Nakagami Channels," *IEEE GLOBCOM'08,* 2008, pp. 1-5, Dec. 2008.

[9] M. K. Simon and M-S Alouini, *Digital Communication over Fading Channels*, New York: Wiley, 2 edition, 2005.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products,* 5th ed., San Diego, CA: Academic, 1994.

[11] A. Annamalai, C. Tellambura, "An MGF-derivative based unified analysis of incoherent diversity reception of M-ary orthogonal signals over fading channels," in *Proc* 54th *IEEE VTC'01,* 2001, pp. 2404 - 2408, Oct. 2001.

[12] K. B. Letaief, W. Zhang, "Cooperative Communications for Cognitive Radio |Networks," *Proceedings of the IEEE,* Vol. 97, No. 5, pp. 878 893, May 2009.

[13] A. Annamalai, O. Olabiyi, S. Alam, O. Odejide, and D. Vaman, "Unified Analysis of Energy Detection of Unknown Signals over Generalized Fading Channels," *to appear in IEEE IWCMC 2011* Conference, Turkey, July, 2011.

[14] S. Atapattu, C. Tellambura, H. Jiang, "Energy Detection Based Cooperative Spectrum Sensing in Cognitive Radio Networks," *IEEE Transactions on Wireless Communications,*, vol.PP, no.99, pp.1-10, 2011.

[15] A. Annamalai, B. Modi, R. Palat and J. Matyjas, "Tight Bounds on the Ergodic Capacity of Cooperative Analog Relaying with Adaptive Source Transmission Techniques," *Proc. IEEE PIMRC'10*, pp. 18-23.

[16] S. Atapattu, C. Tellambura, and H. Jiang, "Performance of Energy Detection: A Complementary AUC Approach," *Proc. IEEE GLOBECOMC'10*, pp. 1-5.

[17] S. Atapattu, C. Tellambura, and H. Jiang, "Relay Based Cooperative Spectrum Sensing in Cognitive Radio Networks," *Proc. IEEE GLOBECOMC'09*, pp. 1-5.

[18] A. Annamalai, O. Olabiyi, S. Alam, "Accurate Approximations of Error Rates for Cooperative Non-Regenerative Relay Systems over Generalized Fading Channels," *to appear in IEEE IWCMC 2011* Conference, Turkey, July, 2011.

[19] Weifeng Su, K. S. Ahmed and K. J. Ray Liu, "Cooperative Communication Protocols in Wireless Networks: Performance Analysis and Optimum Power Allocation," *Wireless Personal Communication*, vol. 44, 2008, pp. 181-217.

[20] M.O. Hasna, M. S. Alouini, "Harmonic Mean and End-to-End Performance of Transmission Systems With Relays," *IEEE Transactions on Communications*, vol. 52, pp. 130–135, Jan. 2004.

[21] M.O. Hasna, M. S. Alouini, "End-to-End Performance of Transmission Systems with Relays over Rayleigh-Fading Channels," *IEEE Transactions on Communications*, vol. 2, pp. 1126–1131, Nov. 2003.

[22] Hypergeometric2F1 (2010) Wolfram Research homepage. [Online]. Available:http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric2F1/20/02/05/

[23] Y. Liu, D. Yuan, M. Jiang, W. Fan, G. Jin, F. Li, "Analysis of Square-Law Combining for Cognitive Radios over Nakagami Channels," *Proc. IEEE GLOBECOMC'09*, pp. 1-5.

Formal Analysis of Mobility Management for Ad Hoc Networks

Shakeel Ahmed[1], A. K. Ramani[1], Nazir Ahmad Zafar[2]
[1]School of Computer Science and Information Technology, Devi Ahilya University, Indore, M.P, INDIA
[2]Department of Computer Science, King Faisal University, Hofuf, SAUDI ARABIA

Abstract - *Mobility is one of the most challenging issues in Mobile Ad hoc Networks (MANET) where nodes are self organized, there is no infrastructure or centralized control and the nodes move freely. Different protocols have been designed for routing the packets from source to destination, earlier to design of any protocols for MANET one of the most important issues to consider is how these protocols can cope up with the unpredictable motion and the unreliable behavior of mobile nodes. The nodes are free to move from one place to another place this can be greatly achieved by mobility management where it deals with storage, maintenance, and retrieval of the mobile host location information. In this paper, a procedure is described to propose a model for mobility management of the AODV network nodes by using graph theory and formal techniques. The searching technique is based on considering the dynamic graphs in accordance with the nodes in MANET where the nodes are not stable. The Z notation is used to transform the graph model into formal specifications. Finally, the specification is analyzed and validated using Z Eves tool.*

Keywords: Ad hoc networks, Formal methods, Mobility management, Z notation, Validation

1 Introduction

Ad Hoc network is a collection of wireless nodes, which form a temporary network without relying on the existing network infrastructure or centralized administration [1]. Mobile Ad hoc Networks (MANETs) are self-organized wireless network of mobile nodes without any fixed infrastructure and is capable of communicating with each other without the assistance of base stations [2]. Nodes roam through the network, causing its topology to change rapidly and unpredictably with the passage of time. New nodes can join the network, while at the same time other nodes leave it or just fail to connect for the short term because they move to a region that is not in the covered range of the network [3]. MANETs are applicable for both military and civilian applications, in which there are no dedicated routers each individual node acts as a router and transmits packets from source to destination [4]. If the source node does not have the destination node within the transmission range the intermediate nodes forward the packet to the destination.

Routing protocol for Ad hoc wireless networks should have the special characteristics [5]. For example it must be fully distributed, adaptive to frequent changes in topology, involving a minimum number of nodes for route computation and maintenance, having minimum time for connection set-up, is localized, loop-free and independent of stale routes [5].

Many of the MANET routing protocols proposed are classified based on the mobility strategy they follow to discover route to the destination where the nodes are mobile. These protocols are based on various factors including the issues about the power consumption, low bandwidth, and high error rates, in decision of selecting the best nodes for routing the packets [6]. Mobility of nodes in MANETs makes a challenging problem that is represented in designing of an efficient and reliable routing strategy to select the path from the source to destination [7]. Routing protocols are to be designed in order to use the limited resources within the transmission range of the nodes.

Most of the proposed protocols are focused on simulation and few implementations are proposed in which environments had no more than a dozen of nodes. Graph theory has much of its applications in the area of parallel and distributed algorithms and is an effective tool for modeling and visualizing the communication networks. Graph theory does not have much computer tool support for verifying and validating the systems. Formal techniques are best approaches for specification and proving the computerized models. In this research, formal methods in terms of Z notation [8] are used by linking with graph theory for describing the mobility management and updating the routing table. Z notation is used because of abstraction and encapsulation of objects for further enhancement of the description of the system. Rest of the paper is organized as follows: Section 2 provides an outline of the related work. Section 3 presents an introduction to formal methods. In section 4, formal specification of mobility management and routing table is described. Finally, conclusion and future work are discussed in section 5.

2 Related Work

There are three main categories of Ad hoc routing protocols: Proactive (table-driven), Reactive (on demand) and Hybrid [9]. Proactive protocols build their routing tables continuously by broadcasting periodic routing updates through the network; reactive protocols build their routing tables on demand and have no prior knowledge of the routes they will take to get to a particular destination. Hybrid protocols create reactive routing zones interconnected by proactive routing links and usually adapt their routing strategy to the amount of mobility in the network [10]. Mobility impacts conditions where routing protocols must operate.

Designing communication and networking protocols for MANET is a challenging task because of the dynamic nature of MANET. Protocols are to be designed where the nodes have to establish and maintain the route from source to destination with the multi hop transmission of data from source to destination. A number of researches have been done in this area, and many multi-hop routing protocols have been developed. The Optimized Link State Routing (OLSR) protocol [11] [12], Dynamic Source Routing protocol (DSR) [13], Ad Hoc on Demand Distance Vector protocol [14], Temporally Ordered Routing Protocol (TORA) [15], and others protocols that establish and maintain routes on a best-effort basis.

Mobility-based method was proposed for improving the performance of the Ad hoc On-demand Distance Vector routing (AODV). Mobility metric was defined and used in both route discovery and route maintenance [15]. In route discovery, the standard AODV hop-count metric is dropped and replaced with a combination of two mobility parameters: average and mean of the calculated mobility along the path between any source node and destination.

In AODV Hello packets were used to enhance mobility awareness [16]. When receiving a Hello packet with the Global Positioning System (GPS) coordinates of the source node, a lightweight mobility aware agent on each node of the network compares these coordinates with previous ones and then can determine information about the mobility of the originator node. Now, when a node receives a RREQ packet and has to send a RREP (it is either the destination, or it has an active route to the desired destination), it will use the mobility awareness to choose the best neighbor which is not moving frequently.

3 Formal Methods

Formal methods are extensively used in a variety of areas including software engineering, modeling and simulation, verifying network protocols, designing and development of parallel and distributed systems, model checking, theorem proving and for checking hardware systems. Hence Formal methods are best choice for modeling of mobile Ad hoc networks due to its distributive nature and because of much component of software as compared to its counterpart hardware. An important aspect of a wireless networks is that nodes use multi-hop communication on an unreliable medium, further the network is subject to dynamic changes and environmental interferences therefore algorithms and protocols should be used for analysis, writing formal specification and producing refinements [18].

A formal specification is a description that is abstract, precise and in a sense is complete. The abstraction allows a human reader to understand the big picture; the precision forces ambiguities to be questioned and removed; and the completeness means all aspects of behavior are described [18]. Secondly, the formality of the description allows us to carry out rigorous analysis. By looking at a single description one can determine useful properties such as consistency or deadlock-freedom [19]. By writing different descriptions from different viewpoints one can determine important properties such as satisfaction of high level requirements or correctness of a proposed design

Z notation [20, 21] is a model-based approach which is a strongly typed, mathematical specification language, not an executable notation and it cannot be interpreted or compiled into a running program. There are few tools for checking Z texts for syntax and type errors in much the same way that a compiler checks code in an executable programming language. In Z notation, schemas are used which are small pieces for decomposing a specification into manageable components. The schema is the feature that distinguishes Z from other formal notations. In Z schemas are used to describe both static and dynamic aspects of a system [22]. Z specification enables to produce a model that is unambiguous, verifiable and traceable. Z is more mature and has an ISO standard [23].

In MANETS nodes are free to move causing changes in the network topology and highly dynamic. This dynamic nature increases the complexity of the algorithms designed for Ad hoc networks and the verification of AODV algorithms is a difficult error-prone task that requires much effort. Formal methods have a lot to offer where mobility of the nodes can be modeled from a complex system to mathematical entities resulting in a rigorous model by using these techniques it is possible to model and verify the mobility of nodes in a more thorough and detailed fashion than the empirical testing and simulation techniques.

4 Formal Analysis

In this section, formal analysis of the network management and routing table for Ad hoc on-demand distance vector routing protocol using Z notation is presented. Initially, formal definitions of basic data types will be described then moving objects and network being complex structures needed for Ad hoc network will be defined. Finally, AODV network and routing table management procedures will be described.

4.1 Formal Model of AODV Network

An interconnected collection of objects that help and allow users to share information and resources is termed as communication network. A mobile Ad hoc network is a collection of self-configuring objects inter-connected by wireless links and devices which are free to move in any direction in the domain. This network might be a part of another larger network of communicating objects. In this paper, the communication network is defined by a graph relation where the moving objects are considered as nodes and communication links are assumed as edges. The graph relation is not static or fixed on the other hand it is dynamic, i.e., its any two nodes may be connected at one time while might be disconnected at another time. In formal definition, the identifier of a moving object is denoted by *Node* as given below. Four types of nodes, i.e., source, destination, internal and nil are assumed here which will be needed to analyze and search the route for data transmission. The power of battery is also considered improving quality of service and assumed as dead, low or normal denoted by *Dead*, *Low* and *Normal* respectively.

$[Node]; Power ::= Dead \mid Low \mid Normal$

$Type ::= Source \mid Destination \mid Internal \mid Nil$

The above types are assumed as sets types in Z notation, where we do not impose any restriction upon number of elements in a set and as a consequent a high order of abstraction is supposed. Moreover, we do not insist upon any effectual procedure to decide about an arbitrary element if it is a member of the given set in Z notation. Consequently, the *Node* is a set of nodes over which we cannot define an operation of cardinality to know the number of elements. Likewise, the complement and subset operations are not well-defined over sets in Z notation. The sets *Power* and *Type* defined above are assumed as free types in which at one time only one value is assumed.

The moving object of the network is defined as a schema given below which is denoted by *Object* and has four components namely, identification (*id*), type (*type*), battery (*battery*) and set of the neighbors (*neighbors*). The type of object is considered because it might be a source, destination or an internal node. Further, the node is given a Nil value if it is not part of any route stored in the routing table.

Object

$id: Node;$

$type: Type$

$battery: Power$

$neighbours: \mathbb{F}\ Node$

The possibility of communication between two objects of the graph is defined by an edge which is described by the schema *Connectivity* given below. It consists of three components: *connection*, *status* and *weight*. The first one is used to define the link based on two nodes of the graph. The second is used to represent its status that is the nodes are connected or disconnected which is represented by *Active* and *NotActive*. And the last one *weight* is used to represent the time needed a node to communicate with the other. Because an object cannot communicate to itself therefore it is checked, first element of the connection cannot be same as its second element.

$Status ::= Active \mid NotActive$

Connectivity

$connection: Object \times Object$

$status: Status$

$weight: \mathbb{N}$

$connection . 1 . id \neq connection . 2 . id$

Description of the communication at object level is extended to define the entire communication network and is initially assumed as a complete graph. This is because we have supposed that there is an edge if communication between two objects is possible. Since any two objects in a network can communicate and hence it is a complete graph that is there is an edge between any two given objects. However, a link (edge) is either active or dead which was considered in the definition of connectivity. The formal specification of the network is described by the schema *Network* given below consisting of two components which are *objects* and *connections*. The variable, *objects*, is a collection of nodes of the graph defined as a finite power set of *object*. And the second variable, *connections*, is a finite power set of *Connectivity* used to represent the edge set. It is proved in predicate part that for any two objects there must be an edge because any two nodes can communicate if the link is active. Similarly, for any edge there must be two nodes in the network which is a natural constraint to define an edge of a graph.

Network

$objects: \mathbb{F}\ Object;\ \ connections: \mathbb{F}\ Connectivity$

$\forall o1, o2: Object \mid o1 \in objects \land o2 \in objects$

$\bullet\ \exists con: Connectivity \mid con \in connections \bullet con.$

$connection= \mid\ (o1, o2)$

$\forall con: Connectivity \mid con \in connections$

$\quad \bullet\ \exists o1, o2: Object \mid o1 \in objects \land o2 \in objects$

$\qquad \bullet\ con . connection = (o1, o2)$

AdhocNetwork

$adhoc: Network$

$\forall con: Connectivity \mid con \in adhoc . connections$

$\quad \bullet\ \exists o1, o2: Object \mid o1 \in adhoc . objects \land o2 \in adhoc .$

$\quad objects \bullet\ con . connection = (o1, o2)$

$\forall o1, o2: Object \mid o1 \in adhoc . objects \land o2 \in adhoc .$

$\quad objects \bullet\ \exists con: Connectivity \mid con \in adhoc . connections$

$\qquad \bullet\ con . connection = (o1, o2) \Rightarrow con . status = Active$

$\forall o1, o2: Object \mid o1 \in adhoc . objects \land o2 \in adhoc\ \ .$

$\quad objects \bullet\ \exists con: Connectivity \mid con \in adhoc . connections$

$\qquad \bullet\ con . connection = (o1, o2) \Rightarrow o1 . id \neq o2 . id$

$\forall o1, o2: Object \mid o1 \in adhoc . objects \land o2 \in adhoc .$

$\quad objects \bullet\ \exists con1: Connectivity \mid con1 \in adhoc . connections$

$\qquad \bullet\ con1 . connection = (o1, o2)$

$\qquad\ \Rightarrow (\exists con2: Connectivity \mid con2 \in adhoc . connections$

$\qquad\qquad \bullet\ con2 . connection = (o2, o1))$

Invariants: (i) For any two objects in the graph relation there must be an edge connecting it which is active or passive (ii) For any link, there must be two objects which can communication to each other. (iii) The identifiers of any of the two objects must be different. (iv) It is supposed that if an object A can communicate to object B then vice versa is also possible that is the graph is a symmetric relation.

As mentioned above, if a node is connected with another node at one time it might be disconnected at another time. Hence we have supposed that communication is possible only if the nodes are connected and the link between the nodes is active. The formal definition of the mobile Ad hoc network is described above based on the definition of network.

4.2 Formal Analysis of Mobility Management

When a node changes its location, activated or newly introduced in the network, the AODV topology will be changed. Following three possible operations are defined based on the change in the state of the network topology:

1. mobility of an object
2. activation of an object
3. new introduction of an object

When an object moves from one location to another it is possible that it will be disconnected from some of the objects and connected to few others causing change in its neighbours. In this way the network topology will be changed. To update the network, a schema *NodeMobility* is described below. The schema consists of four components that are *ΔAdhocNetwork*, *object*, *added* and *removed*. The first one component is used to describe the network topology which represents to collection of all the objects connected at a time. The delta notation is used to represent the change in the network state. The second variable *object* is the moving entity which changes its location from one point to another causing change in its neighbors. The third variable *added* is used to characterize the newly connected nodes of the network with the moving object. And the last one variable *removed* is used to show the disconnected objects from the moving object. All of the four components discussed above are put in the first part of the schema and change in the network state is described in the second part.

In the next an operation is defined when an object is activated from its dead state. This is the case when the object is already part of the network but was deactivated due to any of the reasons. After its activation, it may be connected to the network by introducing a set of neighbors. As a result the network topology will be changed which is updated by the schema *NodeActivation* given below. The schema consists of three components: *ΔAdhocNetwork*, *object* and *added* which are already explained. The *removed* variable is not considered here because for a newly activated node there does not exit any neighbor before its activation.

```
NodeMobility
  ΔAdhocNetwork
  object: Object
  added: 𝔽 Node
  removed: 𝔽 Node
  ─────────────────
  ∃o: Object ∣ o ∈ adhoc . objects
    • object = o ∧ o . neighbours = object . neighbours ∪
      added \ removed ∧ o . type = object . type
      ∧ o . battery = object . battery
  ∀n: Node ∣ n ∈ removed
    • ∃o: Object ∣ o ∈ adhoc . objects • o . id = n
        ∧ (∀con: Connectivity ∣ con ∈ adhoc . connections
            • con . connection ≠ (o, object))
  ∀n: Node ∣ n ∈ added
    • ∃o: Object ∣ o ∈ adhoc . objects • o . id = n
        ∧ (∃con: Connectivity ∣ con ∈ adhoc . connections
            • con . connection = (o, object))
```

Invariants: (i) The moving object must exists in the AODV network topology and after change in the position of the moving object the neighbors are updated by taking union of the newly connected and removing the disconnected nodes. (ii) The edges of the nodes which are disconnected from the moving object are removed from the graph relation. (iii) The edges of the nodes which are connected with the moving object are added in the graph relation.

```
NodeActivation
  ΔAdhocNetwork
  object: Object
  added: 𝔽 Node
  ─────────────────
  ∃o: Object ∣ o ∈ adhoc . objects
    • object = o
      ∧ o . neighbours = added ∧ o . type = object . type
      ∧ o . battery = Normal
  ∀n: Node ∣ n ∈ added
    • ∃o: Object ∣ o ∈ adhoc . objects
        • o . id = n
        ∧ (∃con: Connectivity ∣ con ∈ adhoc . connections
            • con . connection = (o, object))
```

Invariants: (i) The activated object must be in the collection of existing objects of the network. After activation, its neighbors are the only which are newly introduced. (ii) The edges of the nodes which are connected with activated object in the AODV network are added in the graph relation.

If an object does not exist already in the network and is newly introduced it will be connected to its neighbors. It means the node is new to the network which will be first included in the collection of the objects in the network and then linked with the graph (network) relation. The network (graph relation) is changed which is updated by the schema *NodeIntroduction* given below consisting of three components same as in case of node activation operation.

NodeIntroduction

$$\begin{array}{l}
\Delta \mathit{AdhocNetwork}\\
\mathit{newobject}: \mathit{Object}\\
\mathit{added}: \mathbb{F}\ \mathit{Node}\\
\hline
\forall o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects}\\
\quad \bullet\ \mathit{newobject} \neq o \land \mathit{newobject} . \mathit{neighbours} = \mathit{added}\\
\quad\quad \land\ \mathit{newobject} . \mathit{type} = \mathit{Nil}\\
\quad\quad \land\ \mathit{newobject} . \mathit{battery} = \mathit{Normal}\\
\forall n: \mathit{Node} \mid n \in \mathit{added}\\
\quad \bullet\ \exists o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects}\\
\quad\quad \bullet\ o . \mathit{id} = n\\
\quad\quad\quad \land\ (\exists \mathit{con}: \mathit{Connectivity} \mid \mathit{con} \in \mathit{adhoc} . \mathit{connections}\\
\quad\quad\quad\quad \bullet\ \mathit{con} . \mathit{connection} = (o, \mathit{newobject}))
\end{array}$$

Invariants: (i) The newly added object is not in the collection of objects of the existing network. After activation, its neighbors are only which are introduced after its introduction. (ii) The edges of the neighboring nodes are established with newly introduced object and are added in the graph relation.

4.3 Routing Table Management

In this section, routing table is defined then possible operations to manage, after addition and removal of a route, are described. The history of routes is stored in the routing table which is defined by the schema *Routings* consisting of two variables, i.e., graph relation and routes stored. The variable *route is* a collection of routes whereas each route is a sequence of nodes in the network as described below.

Routings

$$\begin{array}{l}
\mathit{AdhocNetwork}\\
\mathit{routes}: \mathbb{F}\ (\mathrm{seq}\ \mathit{Node})\\
\hline
\forall \mathit{route}: \mathrm{seq}\ \mathit{Node} \mid \mathit{route} \in \mathit{routes}\\
\quad \bullet\ \mathrm{ran}\ \mathit{route} \subseteq \{\ o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects} \bullet\ o . \mathit{id}\ \}\\
\forall \mathit{route}: \mathrm{seq}\ \mathit{Node} \mid \mathit{route} \in \mathit{routes} \bullet\ \#\ \mathit{route} \geqslant 1\\
\quad\quad \land\ (\exists o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects}\\
\quad\quad\quad \bullet\ (o . \mathit{id} = \mathit{route}\ 1 \land o . \mathit{type} = \mathit{Source}))\\
\quad\quad \land\ (\exists o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects}\\
\quad\quad\quad \bullet\ (o . \mathit{id} = \mathit{route}\ (\#\ \mathit{route}) \land o . \mathit{type} = \mathit{Destination}))\\
\quad\quad \land\ (\forall i: \mathbb{N} \mid i \in 2 \ldots \#\ \mathit{route} - 1\\
\quad\quad\quad \bullet\ (\exists o: \mathit{Object} \mid o \in \mathit{adhoc} . \mathit{objects}\\
\quad\quad\quad\quad \bullet\ (o . \mathit{id} = \mathit{route}\ (\#\ \mathit{route}) \Rightarrow o . \mathit{type} = \mathit{Internal})))\\
\forall \mathit{route}: \mathrm{seq}\ \mathit{Node} \mid \mathit{route} \in \mathit{routes} \land \#\ \mathit{route} > 1\\
\quad \bullet\ \forall i: \mathbb{N} \mid i \in 1 \ldots \#\ \mathit{route} - 1\\
\quad\quad \bullet\ \exists \mathit{con}: \mathit{Connectivity} \mid \mathit{con} \in \mathit{adhoc} . \mathit{connections}\\
\quad\quad\quad \bullet\ (\mathit{route}\ i, \mathit{route}\ (i + 1)) = (\mathit{con} . \mathit{connection} . 1 .\\
\mathit{id}, \mathit{con} . \mathit{connection} . 2 . \mathit{id})
\end{array}$$

Invariants: (i) In this property, it is stated that a route must be a path whose all nodes are objects of the network topology. (ii) In this property, it is verified that for every route the first node is a source and the last one is a destination. All others must be internal nodes. (iii) The connectivity of nodes in the route is checked and verified.

A route is searched only if it does not exist in the routing table. After route is established and data is transmitted, the route is stored in the routing table for its future use. After discovery of a new route, the routing table is updated by the schema *AddingRute* given below. The schema takes routing table and a new route as input and updates the routing table as an output.

AddingRute

$$\begin{array}{l}
\Delta \mathit{Routings}\\
\mathit{route}: \mathrm{seq}\ \mathit{Node}\\
\hline
\forall \mathit{node}: \mathit{Node} \mid \mathit{node} \in \mathrm{ran}\ \mathit{route}\\
\bullet \exists \mathit{object}{:}\mathit{Object} | \mathit{object}{\in}\mathit{adhoc.objects} \bullet\ \mathit{object} . \mathit{id} = \mathit{node}\\
\exists \mathit{object1}, \mathit{object2}: \mathit{Object}\\
\quad \mathit{object1} \in \mathit{adhoc} . \mathit{objects} \land \mathit{object2} \in \mathit{adhoc} . \mathit{objects}\\
\quad \bullet\ \#\ \mathit{route} \geqslant 1 \Rightarrow \mathit{route}\ 1 = \mathit{object1} . \mathit{id}\\
\quad\quad \land\ \mathit{route}\ (\#\ \mathit{route}) = \mathit{object2} . \mathit{id} \land \mathit{object1} . \mathit{type} = \mathit{Source}\\
\quad\quad \land\ \mathit{object2} . \mathit{type} = \mathit{Destination}\\
\mathit{routes}' = \mathit{routes} \cup \{\mathit{route}\}
\end{array}$$

Invariants: (i) Each node of the route must be an element of the collection of objects of the AODV network. (ii) There exist two objects in the network (graph relation), one is source and the other is the destination of the route. (iii) The new state of routing table is union of its previous routes and the newly established route in the network.

A route is deleted from the routing table if it is not used for a long time or due to any other reason. To delete a route from the routing table a schema *DeleteRute* is presented below. The schema takes same components, as in case of adding a route to the routing table, as input and updates the routing table after deleting the given route.

DeleteRute

$$\begin{array}{l}
\Delta \mathit{Routings}\\
\mathit{route}: \mathrm{seq}\ \mathit{Node}\\
\hline
\forall \mathit{node}: \mathit{Node} \mid \mathit{node} \in \mathrm{ran}\ \mathit{route}\\
\quad \bullet\ \exists \mathit{object}: \mathit{Object} \mid \mathit{object} \in \mathit{adhoc} . \mathit{objects} \bullet\ \mathit{object} . \mathit{id} = \mathit{node}\\
\exists \mathit{object1}, \mathit{object2}: \mathit{Object}\\
\quad \mid \mathit{object1} \in \mathit{adhoc} . \mathit{objects} \land \mathit{object2} \in \mathit{adhoc} . \mathit{objects}\\
\quad \bullet\ \#\ \mathit{route} \geqslant 1\\
\quad\quad \Rightarrow \mathit{route}\ 1 = \mathit{object1} . \mathit{id}\\
\quad\quad\quad \land\ \mathit{route}\ (\#\ \mathit{route}) = \mathit{object2} . \mathit{id}\\
\quad\quad\quad \land\ \mathit{object1} . \mathit{type} = \mathit{Source}\\
\quad\quad\quad \land\ \mathit{object2} . \mathit{type} = \mathit{Destination}\\
\mathit{routes}' = \mathit{routes} \setminus \{\mathit{route}\}
\end{array}$$

Invariants: (i) All nodes in the input route must be in the collection of objects of the AODV network topology. (ii) There exist two objects in the network, one is called source and the other is named as destination of the route to be deleted. (iii) The new state of the routing table is updated by taking set difference of previous routes and the route to be deleted.

5 Conclusions

In this paper, a formal procedure of managing mobile Ad hoc network and routing table is presented by integrating graph theory and Z notation. The objects of the network are represented as nodes and communication between objects is assumed as edge set of the graph. Because objects are free to move from one place to another consequently the communication links might be active at one time and dead at another time. Hence the network is not fixed and it needs frequent management to update the active nodes and live communication links.

We have described the formal model to manage the network due to change of location, activation and new addition of a node in the network. Further, the routing table is maintained when a new route is added or removed. Graph theory is used in this research because it has several applications in modeling of communication networks. But it does not have much computer tool support for verifying and validating the computer models. On the other hand, formal methods are approaches based on mathematical techniques and have a rigorous computer tools support used for analysis, specification and proving of the computerized models. That is why integration of graph theory and formal methods in terms of Z notation used in this research.

It was observed that inconsistencies and ambiguities were removed by application of formal methods for the specification of the above procedures. We believe that this integrated approach is an effective tool for further analysis and optimization of the route request and reply procedures of the AODV routing protocol.

6 References

[1] S. A. Al-Omari and P. Sumari, "*An Overview of Mobile Ad Hoc Networks for the Existing Protocols and Applications*", The International Journal on Applications of Graph Theory in Wireless Ad hoc Networks and Sensor Networks, Vol. 1, no. 1, 2010.

[2] M. A. Ali, A. El-Sayed and I. Z. Morsi "*A Survey of Multicast Routing Protocols for Ad-Hoc Wireless Networks*", Minufiya Journal of Electronic Engineering Research, Vol. 17, no. 2, pp. 185-198, 2007.

[3] V. Daza, J. Herranz, P. Morillo and Carla. "Cryptographic Techniques for Mobile Ad hoc Networks", Comput. Networks, Vol. 51, no. 18, pp.4938–4950, 2007.

[4] S. Zaki, M. Ngadi and S. Razak, "A Review of Delay Aware Routing Protocols in MANET", Computer Science Letters, Vol.1, 2009.

[5] C. S. R. Murthy and B. S. Manoj, "Adhoc Wireless Networks Architecture and Protocols", Prentice Hall, 2004.

[6] D. Djenouri, A. Derhab, and N. Badache, "Ad hoc Networks Routing Protocols and Mobility", Int. Arab J. Inf. Technol. Vol. 3, no. 2, pp.126–133, 2006.

[7] M. Abolhasan, T.Wysocki, E.Dutkiewicz, E., "A Review Of Routing Protocols For Mobile Ad Hoc Networks", Elsevier, Ad Hoc Networks, Vol. 2, no. 1, pp. 1-22, 2004.

[8] J. M. Spivey, "The Z Notation: A Reference Manual", Prentice Hall, 1989.

[9] X. Niu, Z.Tao, G.Wu, C.Huang, Li Cui, "Hybrid Cluster Routing: An Efficient Routing Protocol for Mobile Ad Hoc Networks", IEEE International Conference on Communications, 2006.

[10] L.Wang and S. Olariu, "A Two-Zone Hybrid Routing Protocol For Mobile Ad Hoc Networks", IEEE Transactions on Parallel and Distributed Systems, Vol. 15, no. 12, 2004.

[11] T. Clausen and P. Jacquet. "Optimized Link State Routing protocol (OLSR)", RFC 3626 Experimental, October 2003.

[12] T. H. Clausen, G. Hansen, L.Christensen, and G. Behrmann. "The Optimized Link State Routing Protocol, Evaluation through Experiments and Simulation". In Proceedings of the IEEE Symposium on Wireless Personal Mobile Communications, September 2001.

[13] D. B. Johnson, D.A. Maltz, and Y.C. Hu. "The Dynamic Source Routing protocol for mobile ad hoc networks", 2004.

[14] C. Perkins, E. Belding-Royer, and S. Das. "Ad hoc On-demand Distance Vector (AODV) routing", RFC 3561, 2003.

[15] V. Park and M. Corson. "Temporally-Ordered Routing Algorithm (TORA)", Ver. 1 Functional Specification. Internet-Draft, IETF. draft-ietf-manettora-spec-04.txt, 2001.

[16] Y. Khamayseh, O.M. Darwish and S. A. Wedian, "MAAODV: Mobility Aware Routing Protocols for Mobile Ad hoc Networks", in proceedings of IEEE/CS, 2009.

[17] S. Black, Paul P. Boca, Jonathan P. Bowen, J. Gorman and M. Hinchey, "Formal Versus Agile: Survival of the Fittest", Computer, Vol. 42, no.9, pp. 37–45, 2009.

[18] A. Hall. “Realising the Benefits of Formal Methods”. Journal of Universal Computer Science, Vol. 13, no. 5, pp.669–678, 2007.

[19] S. Chiyangwa and M. Kwiatkowska, “Modeling Ad hoc On-demand Distance Vector (AODV) Protocol with Time Automata”, Proceedings of Third Workshop on Automated Verification of Critical Systems, 2003.

[20] J.M. Spivey, “The Z Notation, A Reference Manual”, 2nd edition. Prentice Hall International, 1992.

[21] J. Woodcock, J. Davis; “Using Z Specification, Refinement and Proof” , Prentice Hall, 1996.

[22] Teyseyre, A. “A 3D Visualization Approach to Validate Requirements”, Proc. Congreso Argentino de Ciencias dela Computacion, Argentina, October 2002.

[23] ISO: Information technology – “Z formal specification notation, syntax, type system and semantics”, 2002 ISO/IEC 13568:2002, International.

Shakeel Ahmed received his B.Sc. (Computer. Science in 1997) from Kakatiya University and M.C.A. (Master of Computer Applications in 2000) from M. K. University, India. Currently he is a PhD student at Devi Ahilya University, Indore, India. His current research interests include Mobile Ad hoc networks, Software Engineering, Modeling of Systems using Formal Approaches, Integration of Approaches etc.

A. K. Ramani received his Master of Engineering and PhD, from Devi Ahilya University Indore, India. He is currently professor and Head of the School of Computer Science and Information Technology, at Devi Ahilya University, Indore, India. His research interest areas are Computer Architecture, Information Systems, System Analysis and Design, Database Systems, Information Technology Project Management, Communication and Computer Networks, Information Architecture, High Performance Computing, Modeling and Simulation. He has over 80 papers and supervised more than a dozen of PhD students.

Nazir A. Zafar received his M. Sc. (Math. in 1991), M. Phil (Math. in 1993), and M. Sc. (Nucl. Eng. in 1994) from Quaid-i-Azam University, Islamabad, Pakistan. He earned his PhD degree in Computer Science from Kyushu University, Japan, in 2004. He has served at various universities and scientific organizations in Pakistan. Currently, he is working at King Faisal University as Associate Professor in the College of Computer Sciences and Information Technology. He is the founder of Center for Research in Computer Science at University of Central Punjab (UCP) Pakistan. He is the Chair of Formal Methods Research Group at UCP. His current research interests include Modeling of Systems using Formal Approaches, Integration of Approaches etc. He is an active member of Pakistan Mathematical Society. Dr. Zafar has lectured at national level promoting use and applications of formal methods at academic and industry and is author of more than 80 research articles.

SESSION

PROTOCOLS

Chair(s)

TBA

AISAC: A novel Artificial Immune System-based Area Coverage Protocol for WSNs

Arash Nikdel[1], **S. Mahdi Jameii**[2], **and Mohamad Sabour**[3]
[1]Department of Computer Engineering, Islamic Azad University, Ramhormoz Branch, Ramhormoz, Iran
[2]Department of Computer Engineering, Islamic Azad University, Shahr-e-Qods Branch, Tehran, Iran
[3]Department of Computer Engineering, Islamic Azad University, Khouzestan Science & Research Branch, Ahwaz, Iran

Abstract-The coverage problem deals with the ability of the network to cover a certain area or some certain events. In this paper, we focus on the problem of area coverage and propose a novel artificial immune system-based area coverage protocol (AISAC) for Wireless Sensor Networks (WSNs). In this protocol, proper sensing radius can be determined using artificial immune system. We have simulated our protocol and compared its functionality to some other protocols. Simulation results show high efficiency of the proposed protocol.

Keywords- Wireless Sensor Networks; Artificial Immune System; Area Coverage; Energy Consumption; Network Lifetime.

1. Introduction

The wireless sensor network (WSN) has emerged as a promising tool for monitoring the physical world. This kind of networks consists of sensors that can sense, process and communicate [1]. Wireless sensor networks are developing quickly and have been widely used in both military and civilian applications such as target tracking, surveillance, and security management [2]. Due to their portability and deployment, nodes are usually powered by batteries with finite capacity. Although the energy of sensor networks is scarce, it is always inconvenient or even impossible to replenish the power. Thus, one design challenge in sensor networks is to save limited energy resources to prolong the lifetime of the WSN [3].

Another challenge in the area of sensor networks is the coverage problem. This challenge deals with the ability of the network to cover a certain area or some certain events. Various coverage formulations have been proposed in literature among which following three are most discussed: *Area coverage*, *Point coverage* and *Barrier coverage*. Covering (monitoring) the whole area of the network is the main objective of area coverage problem [4].

In this paper, we focus on the problem of *area coverage*. We propose a novel artificial immune system-based area coverage protocol (AISAC) for WSNs. In this protocol, each node adjusts its sensing radius using artificial immune system algorithm and considers the sensing radius of its neighbors and network status. The sensing radius will be between minimum sensing radius and maximum sensing radius.

The remaining of this paper is organized as follow: related works are explained in section 2. Section 3 is problem definition. Artificial immune system will be discussed in Section 4. Proposed protocol is explained in section 5. Simulation results are shown in section 6. Section 7 is the conclusion.

2. Related Works

So far, many protocols have been introduced for area coverage control in sensor networks. The coverage concept is a measure of the QoS[1] of the sensing function. The coverage problem is an NP-complete problem [1]. In this paper, we focus on the area coverage problem with random sensor deployment.

The energy consumption of the network can be reduced by allowing the idle sensors to go into the sleep mode. For example, in [5] a node scheduling scheme is proposed to reduce the energy consumption by turning off some redundant nodes in the sensor network, but this centralized solution requires a large number of nodes to operate in the active mode. In [6], sensor scheduling problems of p-percent coverage is studied and two scheduling algorithms to prolong the network lifetime are proposed: CPCA[2] and DPCP[3]. The CPCA is a centralized algorithm which selects the least number of nodes to monitor p-percent of the monitored area. Also the DPCP is a distributed algorithm which can determine a set of nodes in a distributed manner to cover p-percent of the monitored area. Both mentioned algorithms can guarantee network connectivity. The solution used in [7] is to use a transition radius R that is at least twice the sensing range r (i.e. $R \geq 2r$), such that area coverage implies connectivity of active sensors. The authors in [8] address the problem of network coverage and connectivity and propose an efficient solution to maintain coverage, while preserving the connectivity of the network. This solution aims to cover the area of interest, while minimizing the count of the active sensor nodes. In [9], several schemes are designed for sensing coverage subject to different

[1] Quality of Service

[2] Centralized P-Percent Coverage Algorithm

[3] Distributed P-Percent Coverage Protocol

requirements and constraints respectively. In [10], a new distributed and localized coverage control protocol is proposed. This protocol called LDCC[4]. The LDCC protocol does not requires any information about the node location coordinates for selecting the active nodes. Instead, it exploits hop count information, which is easily obtained in a WSN, to select active sensor nodes. In [11], a simple distributed algorithm is developed that allows mobile nodes to autonomously navigate through the field and improve the area coverage. An important element of the proposed algorithm is the ability of each mobile node to autonomously decide its path based on local information. In [12], a set of nodes are made active to maintain coverage while others are put into sleeping modes to conserve energy. This algorithm called PEAS. In PEAS, by adjusting the probing range of sensor nodes, it can achieve different coverage redundancy, but it can't preserve the original sensing coverage completely after turning off some nodes.

In our proposed area coverage control protocol, each node will consider the sensing radius of its neighbors.

3. Problem Definition

Regarding that it is impossible to cover all network points, we use the cellular network for doing area coverage. In this method, the coverage is done according to nodes location and sensing radius such a way that the network area with dimensions x×y is divided in square cells with dimensions c×c. the sensing area of each node is showed by R_S.

The cell is covered if it is completely within the sensing area of a sensor node. We consider a calculated sensing radius (R_{CS}) for each sensor node defined on the basis of real sensing radius and cell size as (1):

$$\begin{cases} R_{CS} = R_S - (2C)^{1/2} \\ R_{C\,\min} = R_{\min} - (2C)^{1/2} \\ R_{C\,\max} = R_{\max} - (2C)^{1/2} \end{cases} \qquad (1)$$

where $(2c)^{1/2}$ is the length of cell diameter. The reason for such definition is that if one of the cell vertices has the least overlap with a node calculated sensing radius, it will be covered completely based on the real sensing radius of node. As a result, one cell is covered if it is within the calculated sensing radius of an active sensor node as shown in Fig. 1. In this figure the dark cells are covered by N node based on R_{CS}. The number of covered cells is divided by all area cells to obtain the network area coverage.

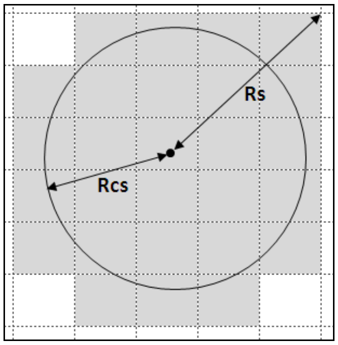


Figure 1: The Covered cells by active node N based on R_{CS}

In this section, we assume that each node has adjustable sensing radius that can be between minimum and maximum area. R_{min} is sensing area with minimum power, R_{max} is sensing area with maximum power and R_S is selective sensing area of node. The value of R_S should be between the R_{min} and R_{max} ($R_{min}<R_S<R_{max}$). Value of sensing area R_{min} and R_{max} will be calculated based on R_t . Value of sensing area R_t identified proportionate with network density [13].

When one cell has some overlap with $R_{C\ max}$ $node_i$, we assume it is one of cells in *sensing set* of $node_i$. Because we sure that this node covered by R_{max} .

Cells of *sensing set* are in four different groups. Sets of A_{min} , A_S and A_{max} are calculated according to the following equations:

$$\begin{cases} c_i \in \textit{sensing set if } c_i \textit{ covered by } R_{max} \\ c_i \in A_{min} \textit{ if } c_i \textit{ covered by } R_{min} \\ c_i \in A_s \textit{ if } c_i \textit{ covered by } R_S \textit{ \& not covered by } R_{min} \\ c_i \in A_{max} \textit{ if } c_i \textit{ covered by } R_{max} \textit{ \& not covered by } R_S \end{cases} \qquad (2)$$

In this equation, c_i is cell number.

$$\begin{cases} \textit{sensing set = all cells have overlap with } R_{C\max} \\ A_{\min} = \textit{all cells have overlap with } R_{C\min} \\ A_S = \textit{all cells have overlap with } R_{CS} \text{ - } A_{\min} \\ A_{\max} = \textit{all cells have overlap with } R_{C\max} \text{ - } (A_S \cup A_{\min}) \end{cases} \qquad (3)$$

$$A_{\max} \cup A_S \cup A_{\min} = \textit{sensing set} \qquad (4)$$

$$A_{\max} \cap A_S \cap A_{\min} = \{\} \qquad (5)$$

A_C consists of cells that are sensed by selective sensing area or are covered by selective sensing area of other nodes. A_{min} is proper subset of A_C, because each node has minimum sensing area A_{min} . Sensing area and set of cells are covered by node n in Fig. 1. The main problem in this paper is choosing minimum sensing area R_S between R_{min} and R_{max} for each node without decreasing the area coverage.

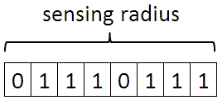
[4] Layered Diffusion-based Coverage Control

4. Artificial Immune System

AIS are distributed adaptive systems for problem solving using models and principles derived from the Human Immune System [14]. The capabilities of the AIS is mainly the inner working and cooperation between the mature T-Cells and B-Cells that are responsible for the secretion of antibodies as an immune response to antigens. The different theories regarding the functioning and organizational behavior of the Natural Immune System (*NIS*) are discussed in literature. These theories inspired the modeling of the NIS into an Artificial Immune System (*AIS*) for application in non-biological environments [14].Many different AIS algorithm models have been built, including Classical View Models, Clonal Selection Theory Models, Network Theory Models, Danger Theory Models [15].

5. Proposed Protocol

Suppose that the network has been clustered and each node is a member of one cluster with a single-hop or multi-hop distance to cluster head. The proposed area coverage protocol consists of two phase: *startup phase* and *sensing radius selection phase*. The sensing radius coding is binary and required *B* bit. *B* is calculated based on (6):



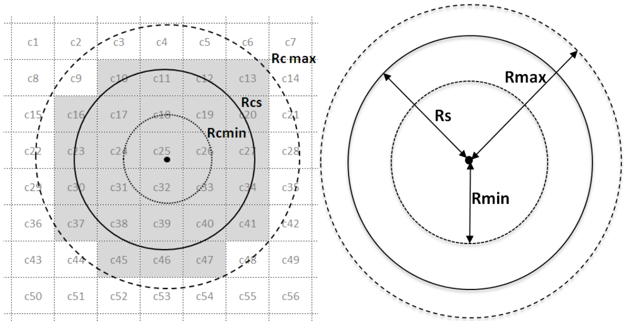
$$B = [log_2\,(R_{\max} - R_{\min})]\ .\ (1/\mu) \qquad (6))$$

The sensing radius of each node is calculated based on (7):

$$R_S = R_{\min}+(R_{\max} - R_{\min})\ (\Sigma_{\ b=1\ to\ B}\, 2^{b-1}\, a_{b-1}\, /\Sigma_{\ b=1\ to\ B}\, 2^{b-1}).\ \mu \quad (7))$$

Where μ is the interval variations rate of sensing radius.

The sensing area and set of cells are depicted in Fig. 2.



Amin = { c17,c18,c19,c24,c25,c26,c31,c32,c33}

As = {c10,c11,c12,c13,c16,c20,c23,c27,c30,c34,c37,c38,c39,c40,c41,c45,c46,c47}

Amax = {c2,c3,c4,c5,c6,c8,c9,c14,c15,c21,c22,c28,c29,c35,c36,c42c44,c48,c52,c53,c54}

FIGURE 2.SENSINNG AREA AND SET OF CELLS

5-1 Startup phase

At first, the sensing rate of each node is set between R_{min} and R_{max} randomly. In this phase, each node with R_S sensing radius, sends its information consist of *ID* and location. In this way, the nodes take their neighbor information. According to them, each node calculates *sensing set*, A_{min}, A_{S} and A_{max} of their neighbors.

5-2 Transition radius selection phase

In this phase, each node adjusts its sensing radius regarding its sets, the sensing radius of neighbor nodes. This phase consists of two stages: *updating sets* and *producing new sensing radius.*

5-2-1 Updating Sets

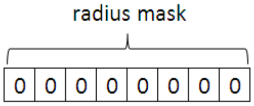
In this stage, for each cell, the member of A_{min} set are placed in its A_{C} set and then, the distance between nodes is added to $A_{\text{max}}/A_{\text{S}}$ sets and the members of A_{C} set is calculated. If the calculated distance for one node is less than A_{min}, it will be removed from A_{S} or A_{max} sets and added to A_{C} set. Moreover, it is considered that whether the A_{max} nodes can be reachable through A_{S} nodes or not. If that is possible, that node is removed from A_{max} set and is added to A_{S} set. See more details in (8):

$$\begin{array}{l}
A_{\text{C}} = A_{\text{min}} \\
\textit{then:} \qquad\qquad\qquad\qquad\qquad\qquad\qquad (8) \\
\forall c_{\text{i}} \in A_{\text{S}}(N)\ Or\ A_{\text{max}}(N)\ \textit{that} \\
\qquad \exists n_{\text{j}} : c_{\text{i}} \in A_{\text{min}}(n_{\text{j}}) \Rightarrow \\
\left\{\begin{array}{l} A_{\text{min}}\ (N) = A_{\text{min}}\ (N) + c_{\text{i}} \\ A_{\text{S}}(N) = A_{\text{S}}(N) - c_{\text{i}}\ \ Or\ \ A_{\text{max}}(N) = A_{\text{max}}(N) - c_{\text{i}} \end{array}\right. \\
\textit{then:} \\
\forall n_i \in A_S\ (N)\ \ \exists n_j \in A_{min}(n_i)\ \ AND \\
\qquad (n_j\ \in A_{max}(N)) \Rightarrow \\
\left\{\begin{array}{l} A_S\ (N) = A_S\ (N) + n_j \\ A_{max}(N) = A_{max}(N) - n_j \end{array}\right.
\end{array}$$

Regarding to A_S and A_{max} condition, the node performs a sensing radius mask ($Mask_{\text{sensing}}$) and determines a sensing radius mask operation ($Operation_{\text{mask\_sensing}}$) with *OR/AND*. The method of determining sensing radius mask and sensing radius mask operator is calculated according to the four conditions:

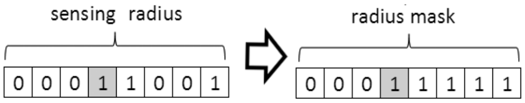
- *Both A_S* and A_{max} sets are empty

The sensing radius of node is equal to A_{min} . So, radius mask is as below: (Mask operator is *AND)*



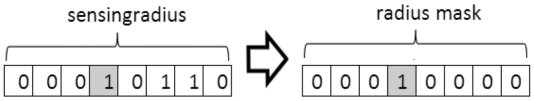
- A_S set is not empty and A_{max} set is empty

The node can select its sensing radius between both R_{min} and R_S. The sensing radius of this mask is as below. (Mask operator is *AND)*



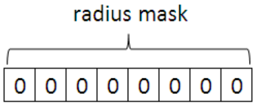
- A_S set is empty and A_{max} set is not empty

The node can select its sensing radius between R_S and R_{max}. Sensing radius of this mask is as below: (the Mask operator is *OR*)



- Both A_S and A_{max} sets are not empty

The node can select its sensing radius between A_{max} and A_{min}. This sensing radius of Mask is as below: (the Mask operator is *OR*)



Thus, regarding to A_S and A_{max} condition, the node determines $Mask_{\text{sensing}}$ and $Operation_{\text{mask\_sensing}}$ as (9):

$$
\begin{aligned}
&\text{If } A_S=\Phi \text{ And } A_{\max}=\Phi \Rightarrow \\
&\qquad \begin{cases} Mask_{\text{transition}} = 0 \\ Operation_{\text{mask\_transition}} = \text{'AND'} \end{cases} \qquad (9)\\
&\text{If } A_S<>\Phi \text{ And } A_{\max}=\Phi \Rightarrow \\
&\qquad \begin{cases} Mask_{\text{sensing}} = [Log_2 R_S]*2-1 \\ Operation_{\text{mask\_sensing}} = \text{'AND'} \end{cases}\\
&\text{If } A_S=\Phi \text{ And } A_{\max}<>\Phi \Rightarrow \\
&\qquad \begin{cases} Mask_{\text{sensing}} = [Log_2 R_S] \\ Operation_{\text{mask\_sensing}} = \text{'OR'} \end{cases}\\
&\text{If } A_S<>\Phi \text{ And } A_{\max}<>\Phi \Rightarrow \\
&\qquad \begin{cases} Mask_{\text{sensing}} = 0 \\ Operation_{\text{mask\_sensing}} = \text{'OR'} \end{cases}
\end{aligned}
$$

Regarding what mentioned above, if a node sensing radius selection is performed surely, it informs its neighbors of this selection and its A_C sets.

5-2-2 Producing New Sensing Radius

After performing the "startup phase", new sensing radius process is started. The nodes select their own sensing radius within cycles. The *affinity* rate of each node is depended on selected sensing radius of that node and its neighbors. At each cycle, any node selects its new sensing radius. Each node, based on its selected sensing radius, sends its A_C set with A_{max} and A_S to its neighbors as "*status-pack*" package. After that, the node receives its neighbor's sensing radius and then, based on them, determines the *affinity* of its selected radius.

If the node receives all neighbors "*status-pack*" packages, it updates its sets based on received A_C. whenever the node can't received some of neighbor's "*status-pack*" package due collision, it is supposed that the selected sensing radius of neighbor nodes is R_{min} (the smallest sensing radius).

Whenever one cell of A_{C} becomes a member of A_{S} or A_{max}, that cell is removed from A_{S} or A_{max} and adds to the A_{C} set. See more details in (10):

$$
\begin{aligned}
&\forall c_i \in A_C(N)\\
&\quad \exists n_j: \ (c_i \in A_S(n_j) \text{ Or } c_i \in A_{\max}(n_j)) \Rightarrow \qquad (10)\\
&\begin{cases} A_C(N) = A_C(N) + c_i \\ A_S(N) = A_S(N) - c_i \text{ Or } A_{\max}(N) = A_{\max}(N) - c_i \end{cases}
\end{aligned}
$$

Where Ax(y) is Ax set of node y.

After updating the sets, the node determines the *affinity* of its selected sensing radius regarding to neighbor's selected sensing radius. For this purpose, the node considers a temporary TA_C set. As can be seen in (11), this set, at first, is equal to A_C.

$$TA_C(N) = A_C \qquad (11)$$

Then, according to (12), the node adds the A_{S} set of its neighbors to the same neighbor A_{C} set:

$$\forall n_i : A_C(n_i) = A_C(n_i) + A_S(n_i) \qquad (12)$$

After that, regarding the (13), the node updates TA_C set:

$$
\begin{aligned}
&\forall n_i \ \exists c_j \in TA_C(n_i) \text{ And}\\
&\quad (c_j \in A_S(N) \text{ Or } n_j \in A_{\max}(N)) \Rightarrow \qquad (13)\\
&\quad TA_C(N) = TA_C(N) + c_j
\end{aligned}
$$

After updating TA_{C} set, the process of determining sensing radius *affinity* of node is as below:

If $A_S \subset TA_C$ and $A_{\max} \subset TA_C$

At this situation, more closely the sensing radius rate to R_{min}, more fit the sensing radius. So:

$$affinity = \lambda_1 + \psi_1 * (my\text{-}cell\text{-}A \ / \ max\text{-}cell\text{-}A)(1/(R_S - R_{\min} + \varepsilon)) \qquad (14)$$

Where ε shows very small positive number, λ_1 shows the minimum acceptable rate for *affinity* of node and ψ_1 is selected as the *affinity* rate.

- If $A_S \not\subset TA_C$ or $A_{\max} \not\subset TA_C$

The node adds A_S set to TA_C set and updates TA_C set again by (13). Then, If $A_{\max} \subset TA_C$ (evidently $A_S \subset TA_C$), so the sensing radius will be fit and can be smaller. The details can be seen in (15):

$$\text{affinity}=\lambda_2+\psi_2*(\textit{my-cell-A} \,/\, \textit{max-cell-A})(1/(R_S-R_{\min}+\varepsilon)) \quad (15)$$

If $A_{max} \not\subset TA_C$, more closer the node sensing radius to $R_{\max}$, more affinity of it. So, affinity can be defined as (16):

$$\text{affinity}=\lambda_2+\psi_2*(\textit{my-cell-A} \,/\, \textit{max-cell-A})(1/(R_{\max}-R_S+\varepsilon)) \quad (16)$$

Where ε shows very small positive number, λ_2 shows the minimum acceptable rate for affinity node, and ψ_2 is selected as the affinity rate doesn't exceed a given limit.

In (14), (15) and (16), *my-nodes-A* shows all member of two A_S, $A_{\max}$ sets of cell. The rate of *max-nodes-A* is calculated by (17). In this relation, $A_x y$ shows the A_x set of node *y*.

$$\begin{aligned}
&\textit{for each node } N\text{:}\\
&\quad \textit{max-node-A} = \max(a,b)\\
&\begin{cases} a = \max\left(|A_S n_j| + |A_{\max} n_j| \ \forall n_j \textit{ that is neighbor of } N\right)\\ b = \textit{my-node-A} = |A_S N| + |A_{\max} N| \end{cases} \qquad (17)
\end{aligned}$$

After calculating the *affinity* rate, each node after receiving the mentioned package acts as follows:

- If the node *affinity rate* is less than *affinity threshold rate*, it will release its selected radius without any change. (i.e. this node selected as a memory cell).
- If the node *affinity rate* is less than *affinity threshold rate*, its transition radius with *mutation rate* τ, is mutated. The ratio of *mutation rate* τ to *affinity rate* is inverses, as a result, the node with more *affinity rate* will have less mutation and with less *affinity rate*, they have more mutation. The bits mutate and are selected randomly and the selected bit will be inverted (zero change to 1 and vice versa). This process is shown in Fig. 3.

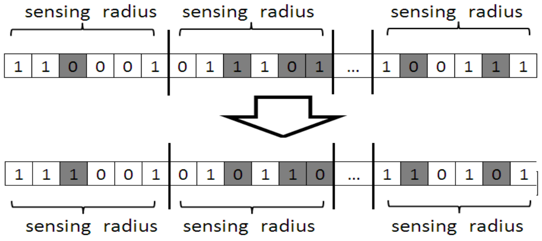


Figure 3. Process of mutating the sensing radiuses

Fig. 4 shows a big binary number with four different mutation ranges. Regarding the binary rate come before and after mutation, we observe that the performance of mutation operator makes the small number bigger and big ones smaller very likely. Regarding the reverse ratio of *mutation rate* to *affinity rate*, the less node's *affinity* have the more node's *mutation* and if a number is big, it will becomes smaller and vice versa. The nodes with more *affinity* have less *mutation* and also less change.

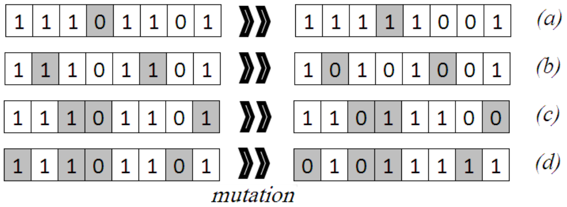

1 1 1 0 1 1 0 1 ⟫ 1 1 1 1 1 0 0 1 (a)
1 1 1 0 1 1 0 1 ⟫ 1 0 1 0 1 0 0 1 (b)
1 1 1 0 1 1 0 1 ⟫ 1 1 0 1 1 1 0 0 (c)
1 1 1 0 1 1 0 1 ⟫ 0 1 0 1 1 1 1 1 (d)
mutation

Figure 4. A sensing radius with four different mutation rates

So, the next cycle begins and again nodes determine the sensing radius by using mentioned algorithm and this process continues.

The cycles of sensing radius selection continue until one of these conditions is achieved:

- The *affinity rate* of selected sensing radius exceeds threshold rate *TP*. At this condition, node releases its selected sensing radius without any change.
- The numbers of sensing radius selection cycles are as threshold rate *TS* or all neighbor nodes select their sensing radius: regarding the condition of sets, the node determines its sensing radius in the way that if the $A_{\max}$ set is not empty, $R_{\max}$ sensing radius is selected. If A_S set is not empty, R_S sensing radius is selected, otherwise $R_{\min}$ sensing radius will be selected.
- A_S and $A_{\max}$ sets are empty: this condition results from updating the cell sets according to sent A_C sets from neighbors. In this situation, the node selects $R_{\min}$ sensing radius.

In all condition mentioned above, whenever a node determines its fix selection, sends this selection with its updated A_C set to the neighbors. The neighbors, after receiving this message and after updating their sets, remove that node from their neighbor list.

6. Simulation Result

According to different energy consumption models, the power consumed by a working node to deal with a sensing task in a round is proportional to r_S^2 or r_S^4 , where r_S is the sensing radius of the working node [16]. In this paper, we take the sensing energy consumption as $u.r_S^2$, where *u* is a factor.

The coverage energy consumption of the sensor set, which is related to the sum of the sensor's sensing radius squared, is defined as (18):

$$E_{\text{total}} = u. \sum_{\text{i=1 to N}} r_{\text{i}}^{2} \qquad (18)$$

So, the energy consumption per area is calculated based on (19):

$$E_{\text{area}} = E_{\text{total}} / A_{\text{area}} = u. \sum_{\text{i=1 to N}} (f_i \times r_{\text{i}}^{2}) / A_{\text{area}} \qquad (19)$$

Where N is the number of all nodes and A_{area} is the monitoring rate of the sensor set.

In (19), the value of function f_i is calculated according to (20):

if the node$_i$ is in sleep mode

$f_i = 0$

Else (20)

$f_i = 1$

In the simulation, based on (21), we define the coverage rate of the sensor set, R_{area}, as the proportion of the monitoring area A_{area} to the total area A_{S}:

$$R_{\text{area}} = A_{\text{area}} / A_{\text{S}} \qquad (21)$$

To evaluate the proposed protocol, it is compared with OGDC protocol [7].

In the simulations, we assume an area with a size of 150×150 which is divided in cells with size of 1×1. We deploy the sensor nodes randomly in the area. The number of nodes, N, in different configurations are considered as 110, 120, 130, 140, 150, 160, 170, 180, 190 and 200 respectively.

In the proposed protocol, AISAC, the nodes sensing radius is considered as $R_{\text{S}} \in [8,23]$ and interval variations rate is equal to μ=0.25. As a result, the number of bits being required to preserve the selected sensing radius of each node equals B=6.

Regarding the ability of proposed protocol to adjust the sensing radius of nodes, it is compared with OGDC protocol with three sensing radiuses with size of 8, 10 and 12 m.

As shown in fig. 5, due to its ability to adjust sensing radius, the proposed protocol is able to provide higher coverage rate with less energy consumption in comparison with OGDC. Due to the proposed protocol accuracy in adjusting the nodes sensing radius, it is able to provide the full coverage in less densities. Moreover, as the result of increasing nodes density, the proposed protocol decreases both nodes sensing radius and energy consumption, but OGDC protocol is not able to decrease energy consumption and to provide the full coverage in low density because of using fixed sensing radius.

Another noticeable point in the proposed protocol is the number of active nodes. As shown in fig.6, the number of active nodes is equal to the number of all nodes while energy consumption in the proposed protocol is less in different configurations. As a result, the proposed protocol maintains more balance in using nodes energy so that it prolongs the network lifetime.

On the other hand, as shown in fig. 7, the energy consumption per area in different configurations of our protocol is less than the OGDC protocol.

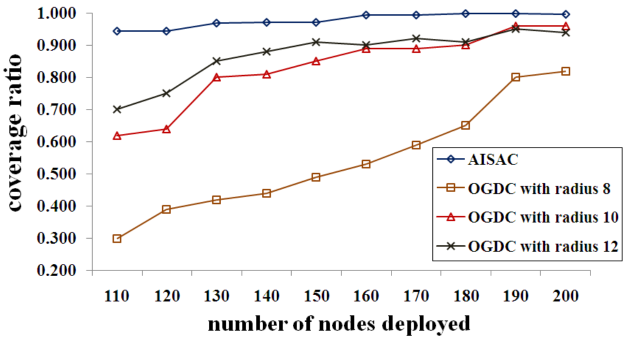


Figure 5. The coverage rate of the sensor set in different configurations

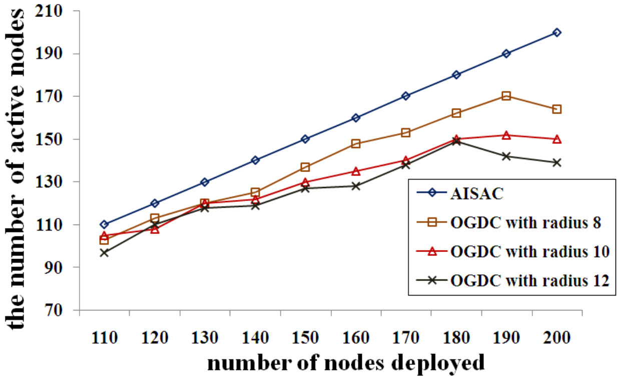


Figure 6. The number of active nodes in different configurations

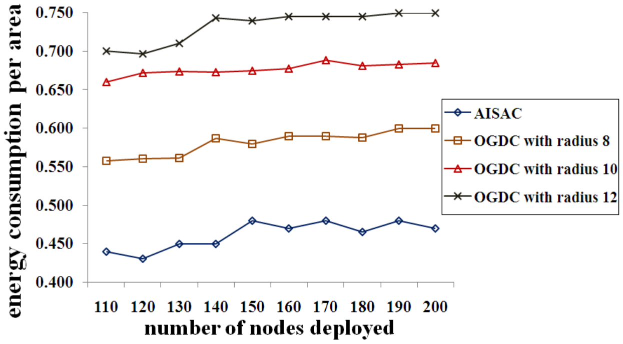


Figure 7. The energy consumption per area in different configurations

7. Conclusion

In this paper, we proposed a topology control protocol based on artificial immune system. In this protocol, nodes can select proper transition radius. Simulation results showed that the proposed protocol has some advantages compared to the previous protocols. First advantage is minimum average of transition area and adjusting the radio radius dynamically, unlike previous protocols that should select radio radius among predefined values. Second advantage is that in our protocol the average number of neighbors is less compared to existing protocols. So, the energy consumption in our protocol is less than others and the network lifetime will be prolonged. In addition, we showed that the network connectivity in our protocol is in the acceptable level.

References

[1] A. Jie Jia, C. Jian, C. Guiran, C. Zhenhua , "**Energy efficient coverage control in wireless sensor networks based on multi-objective genetic algorithm**" ,Elsevier , Computers and Mathematics with Applications 57 (2009) 1756_1766

[2] P. Santi, "**Topology Control in Wireless Ad Hoc and Sensor Networks**", Wiley, 2005.

[3] A. Jie Jia, C. Jian, B. Guiran, W. Yingyou, S. Jingping , "**Multi-objective optimization for coverage control in wireless sensor network with adjustable sensing radius**" , Elsevier, Computers and Mathematics with Applications 57 (2009) 1767_1775

[4] M. Ilyas, I. Mahgoub, Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press, London, Washington, DC, 2005.

[5] D. Tian, N.D. Georganas, A coverage-preserving node scheduling scheme for large wireless sensor networks, in: Proc. of the First ACM Intel workshop on Wireless Sensor Networks and Applications, New York, USA, 2002, pp. 32_41.

[6] Y. Li, C. Ai, Z. Cai. R. Beyah**, *"Sensor scheduling for p-percent coverage in wireless sensor networks",*** Springer Science+Business Media, LLC 2009.

[7] H. Zhang, J.C. Hou, Maintaining sensing coverage and connectivity in large sensor networks, Ad-hoc and Sensor Wireless Networks 1 (2005) 89_124.

[8] S. Misra, M. Kumar, M. S. Obaidat**, *"Connectivity preserving localized coverage algorithm for area monitoring using wireless sensor networks",*** 2010 Elsevier B.V. All rights reserved.

[9] X. CHEN, Q. ZHAO, X. GUAN**, *"Enegy-efficient sensing coverage and communication for wireless sensor network",*** 2007 Springer Science + Business Media, LLC.

[10] B. Wang, C. Fu, H. B. Lim**, *"Layered Diffusion-based Coverage Control in Wireless Sensor Networks",*** 2008 Elsevier, Computer Networks 53 (2009) 1114–1124.

[11] T.P. Lambrou, C. G. Panayiotou, S. Felici, B. Beferull**, *"Exploiting Mobility for Efficient Coverage in Sparse Wireless Sensor Networks",*** Springer Science+Business Media, LLC. 2009.

[12] F. Ye, G. Zhong, J. Cheng, S. Lu, L. Zhang, PEAS: a robust energy conserving protocol for long-lived sensor networks, in: Proc. of the 23rd International Conference on Distributed Computing Systems, ICDCS, Providence, USA, 2003, pp. 28_37.

[13] D. Stauffer and A. Aharony, ***"Introduction to Percolation Theory",*** London: Taylor & Francis, 1994.

[14] A. M. Bidgoli, A. K. Javanmardi, A. M. Rahmani,"Application of AIS algorithm for optimization of TORA protocol in ad hoc network", in: IEEE 2010

[15] A. P. Engelbrecht , "Computational Intelligence An Introduction" ,second edition, wiley 2007

[16]M. Lu, J. Wu, M. Cardei, M. Li, Energy-efficient connected coverage of discrete targets in wireless sensor networks, in: Proc. of the International Conference on Computer Networks and Mobile Computing, ICCNMC, Zhangjiajie, China, 2005, pp. 43_52.

A new model for the ns-3 simulator of a novel routing protocol applied to underwater WSN

S. Climent, J.V. Capella, A. Bonastre, and R. Ors
ITACA, Universitat Politècnica de València
46022 València, Spain
+34963877000 Ext. 85703

Abstract— *Providing reliability, scalability and energy efficiency in Underwater Wireless Sensor Networks (UWSN) is very challenging due to its special features. UWSNs usually employ acoustic channels for communications, which compared with radio-frequency channels, allow much lower bandwidths and have several orders of magnitude longer propagation delays.*

Some routing protocols have been proposed to address these problems, but these approaches usually present unrealistic assumptions or do not consider important questions such as energy consumption or reliability.

In this paper we propose the application of EDETA (Energy-efficient aDaptive hiErarchical and robusT Architecture) to the subaquatic acoustic medium. This routing protocol allows a multiple-sink architecture without introducing extra cost. Furthermore, it adds fault tolerant mechanisms to the network structure.

An implementation of a subset of EDETA in the ns-3 simulator has been carried out. Extensive simulations have been executed. The results show very high packet delivery ratios with reduced energy consumption while offering scalability and reliability.

Keywords: Simulation, Protocol design and analysis, Low-power design, Routing protocols, Sensor networks.

1. Introduction

During last decade there have been large improvements in terrestrial sensor networks. In the last few years, researchers are increasing their efforts in underwater sensor networks. This networks enable a broad range of applications like environmental monitoring, seismic monitoring or distributed tactical surveillance [1].

In this networks the transmission is done by means of acoustic waves, since electromagnetic waves are heavily attenuated underwater. But, this kind of transmission also has drawbacks. Signal attenuation changes with distance and frequency [2]. Signal propagation changes with distance and it can become cylindrical or spherical [3].

Shallow waters present high multipath interference and methods to mitigate ISI have been developed. Also there are interferences from ambient noise generated by shipping activity or surface waves [4].

Another drawback is related to the signal propagation, which is 1500 m/s, five orders of magnitude lower than its radiofrequency counterpart. On radio-frequency networks this delay is negligible, but on underwater acoustic networks, it has to be considered. Although lots of results have been achieved at the physical and medium access layer to address these differences, little work has been done at the routing layer [5], [6].

In this paper we present an adaptation of a recently proposed routing protocol [7], [8] for wireless sensor networks to the subaquatic acoustic medium. In this line, we have implemented a model of the protocol using the ns-3 simulator and its underwater model [9].

The ns-3 is an open source software project. Its objectives are building and maintaining a discrete-event network simulator. Our experience with it has been quite pleasing compared to our previous experience with other simulators. It has been faster to learn and easier to debug than its predecessor, the ns-2. The inbuilt memory management, the use of just one programming language and the absence of split objects, has helped the implementation of the protocol.

The remaining of this paper is organized as follows. In Section 2 some of the routing protocols for underwater acoustic networks are described. In Section 3 a brief introduction to the protocol we want to propose for UWSN is performed. In Section 4 we discuss how was done the implementation of the protocol in the ns-3 simulator, providing implementation details. Furthermore, we explain some changes to the underwater model that were necessary in order to support all the required features of the routing protocol. In Section 5 results obtained from simulations are shown. And finally, in Section 6 conclusions and future work in this line are drawn.

2. Related work

Underwater acoustic transmission has been heavily studied during last decade. Recently, significant advances in MAC and routing protocols for underwater sensor networks have been achieved. Good surveys referring the recent advances and challenges in underwater sensor networks can be found at [4], [6], [5].

The propagation delay is one of the most studied factors in MACs for UWSN. Researchers have been trying to adapt existing protocols and have been proposing new ones in order to address the differences between terrestrial and underwater acoustic networks.

To this end, some authors propose in [10] a modified version of ALOHA in order to adapt it to the new transmission medium. Their results show throughput increase due to a reduction on the number of collisions.

The original FAMA (Floor Acquisition Multiple Access) protocol [11] assures no packet collisions provided that the RTS and CTS frames are long enough. Given the long propagation delay of the underwater acoustic medium, theses packet lengths are very high, hence Molins et. al. propose in [12] the Slotted FAMA MAC protocol. This protocol provides some energy savings since nodes have not to transmit long RTS/CTS frames. Although, the slot length needs to be equal to the maximum propagation delay plus the transmission time of a CTS packet, which can lead to a low channel utilization.

T-Lohi (Tone-Lohi), an hybrid between RTS/CTS and CSMA, is proposed in [13]. This protocol adapts by itself the contention time to the number of contending nodes. The nodes send a short packet (called tone) prior to the actual data packet to count the number of terminals contending for the channel. If a node does not receive any other tones, it starts the transmission. However, if it receives more tones, it adapts its backoff time depending on the number of tones received. The channel utilization of this protocol is within 30% of the theoretical maximum.

Pompili et. al. advocate for a CDMA-based MAC [14]. This MAC switches at each sender node from an ALOHA scheme, to transmit the header, to a CDMA-based scheme, to transmit the payload. This payload and the header are sent back-to-back in one transmission. A node first sends a short header with the spreading code and immediately it sends the payload using this spreading code. Results show that this MAC scheme can achieve high network throughput in deep water communication and it can dynamically adapt to compensate the multipath effect in shallow waters.

On the network layer, despite the huge amount of routing protocols proposed for terrestrial sensor networks, in underwater acoustic sensor networks there is a lot of work to be done.

In [15] the author studies the behavior of DSDV (Destination-Sequenced Distance Vector), AODV (Ad hoc On Demand Distance Vector) and DSR (Dynamic Source Routing) in the underwater acoustic medium. An adaptation of the ns-2 wireless model is done in order to simulate the underwater conditions. The results show that AODV has better performance than DSDV and DSR.

One of the main characteristics of the underwater acoustic channel is the propagation delay. Since it is five orders of magnitude higher than its radio frequency counterpart, protocols must be adapted [3].

A centralized routing protocol is proposed in [16] to mitigate the high packet delay using a sliding window approach. The main drawback of this protocol is the increasing need of different transmission channels with the number of nodes.

Several authors have been trying to adapt routing protocols to the underwater channel. For instance, in [17] a modified AODV algorithm called AODV-BI is proposed. Results show that AODV-BI (AODV-Bi-directional) has less latency and lost packets than AODV.

Zorzi and Casari study in [18] the effects of the differences between the terrestrial and underwater transmission mediums and the relationship between energy consumption and different radio modes. Furthermore, they design a set of new routing protocols considering these studied factors. Although they make the assumption that each node has information about its position, it is not specified how the location is performed.

There are some approaches that use geographic routing protocols. In the VBF (Vector-Based Forwarding) algorithm [19] a node forwards a packet if the node is close enough to the estimated routing vector. The sender piggybacks into the data packet its position and the receiver position. With this information an intermediate node will forward the packet if it is close enough to the routing path.

QELAR (Q-learning-based Routing) is presented at [20]. It is an adaptive routing protocol based on a machine learning approach. When a node needs to transmit data, it piggybacks some state information. Each time a node receives a packet, even if it is not its destination, it reads the added state information and updates its state and routing function. Authors compare its performance versus the VBF algorithm and conclude that QELAR can achieve the same performance in terms of delivery rate and routing efficiency as VBF, but with higher energy efficiency.

In [21] a centralized routing protocol based on geographic information is proposed. The master node computes the topology and the routing paths. The main drawback lies in the complexity of the algorithm that computes the topology and the routing paths, since it is NP-complete.

Minimum Cost Clustering Protocol (MCCP) is a distributed clustering protocol proposed in [22]. The authors propose a cluster-centric cost-based optimization problem for the cluster formation. Although cluster-heads have the ability to send the data in a multi-hop manner to reach the sink, all nodes are supposed to be able to reach the sink. The cluster-head selection algorithm does not assure that the cluster-heads far away from the sink are going to be able to relay their data to other cluster-heads.

Although these approaches try to minimize the energy consumption selecting energy efficient routes or minimizing the overhead of control packets, no low power consumption modes of the nodes are used. The proposed protocol takes advantage of this and sets the nodes to a low-power consumption mode when they do not need to recollect data. Moreover, as explained in [7], [8] the protocol adds fault tolerant mechanisms and time-constrained properties.

3. Energy-efficient aDaptive hiErarchical and robusT Architecture

EDETA [7], [8] is a two levels protocol, the first formed by clusters and the second formed by a dynamic tree. It elects its clusters randomly and recalculates the achieved network structure after certain number of rounds.

A tree structure of Cluster-Heads (CH) is built in order to send data to the sink, which occupies the root of the tree.

The protocol supports the existence of more than one sink in order to provide more scalability and increase its fault tolerance.

The CH election is based on a calculated threshold (T) given by equation (1):

$$T(n) = \frac{c}{|N| - 2c} \times \alpha,\, n \in N \tag{1}$$

Where c is the optimum number of clusters in the network, N is the set of the nodes in the network and α is a parameter that depends on the moment in which the equation is computed.

The nodes will compute the above equation in two situations. The first case corresponds to the beginning of the network configuration, with $\alpha = 1$. The other case appears during the network configuration, if there is any CH without tree connectivity and thus, needs some leaf nodes to become CH. In this case α has to be greater than one to increase the node's probability to become CH.

When a node tries to become CH, if the number randomly generated is lower than the calculated threshold, a node can become a CH only if its remaining energy is greater than (2):

$$E(n) = E_T \times \frac{2T_{Config}}{2T_{Config} + MAX_{Rounds}T_{SuperFrame}},\, n \in N \tag{2}$$

Where E_T is the mean of the remaining energy of the CHs that are around the node. T_{Config}, $T_{SuperFrame}$ and MAX_{Rounds} are protocol parameters explained later at Section 3.1.

EDETA is a time-constrained protocol. As will be seen on the next section, the operation of EDETA is divided into phases. The duration of these phases is limited. This way, EDETA can be applied to applications in which time is an important variable.

3.1 Operation

EDETA is divided into two phases called, the initialization phase and the normal operation phase. There are two variables that limit the duration of these phases. The T_{Config} variable limits the initialization phase and the $T_{SuperFrame}$ variable limits one round of the normal operation phase. The normal operation phase has a limited number of rounds defined by the parameter MAX_{Rounds}. So, the normal operation phase lasts $MAX_{Rounds} \times T_{SuperFrame}$.

3.1.1 Initialization phase

In this phase the network structure is built. CHs are self-elected and the rest of the nodes, also called leaf nodes, choose a cluster to join and ask for admission. After that, clusters organize themselves in a tree structure to deliver the collected data to the sink. More specifically, this phase is divided into three sub-phases.

On the first part, with duration of half T_{Config}, each node decides on its own if it is going to be a CH, based on the above explained procedure.

When a node becomes a CH it sends HEAD messages to announce its role. At the same time, a CH receives HEAD messages from the others and decides which CH is the best option to send its data to the sink.

This decision is based on the signal strength of the HEAD messages. However, a CH will only try to join another CH which has established a path to the sink, so it can reach the sink directly or through other CHs. This one will be its parent CH.

Meanwhile, leaf nodes, also receive HEAD messages. They store them to decide which CH they will join on the second part of this phase. The selection of the CH for these nodes uses the same criteria as the selection of parent CH.

If a CH doesn't receive any HEAD message, it sends a NEED_CH message. When a normal node receives it, it reruns the procedure to decide whether it is going to become a CH or not, but with an increased value of α. As α is increased the probability of a node to become a CH increases too. In this case, the probability must be increased because of the need to achieve full connectivity as soon as possible. This mechanism, along with the random distribution of CHs allows the protocol to adapt rapidly the population of CHs to the needs of the network.

At the end of this sub-phase, the tree structure is build and leaf nodes have the necessary information to decided which cluster will join in.

On the second sub-phase, with duration of half T_{Config}, normal nodes try to join their selected clusters. CHs send, in the response message, the time schedule in which each node has to send its data. After that, the leaf nodes enter into the low-power state.

A CH only allows a limited number of leaf nodes to join in. This number is given by MAX_{Soft} and MAX_{Hard}. A CH accepts all the join request petitions until it reaches its MAX_{Soft}. After that, it will only accept join petitions that have activated a last-resort bit. When a CH reaches the MAX_{Hard} threshold, it will no longer allow new joins.

Finally, on the third sub-phase, with duration of one T_{Config}, each leaf CH in the tree sends to its father the amount of time it needs to have all the data recollected from its leaf nodes. Each father CH collects this information from all its sons and decides the time schedule in which each son can send its data. After that, the father repeats the process with its own father, sending the amount of time needed to collect all the data from its nodes and its sons. Then, the grandfather decides the time schedule in which its sons have to send it the data. This process continues until the entire tree schedule is done.

3.1.2 Normal operation phase

At this moment, the network structure is done and every leaf node must send its data to the CH at its scheduled time. During the remaining time the nodes enter in a low-power state. When a CH has received all the data from all its nodes, it aggregates it and sends it to his father at the established time.

As has been said, the father of a cluster informs at which time its sons have to send their data. Sons will send their data when they receive a POLL message. This allows the father to decide exactly when the data will be sent. This makes applicable some fault tolerant mechanisms, as discussed in [7], [8], without inquiring collision of messages. Moreover, this polling mechanism allows timing synchronization between all the CHs in the tree.

This phase is repeated during some amount of rounds, with duration $T_{SuperFrame}$, defined by parameter MAX_{Rounds}. After that, the network structure is considered obsolete and every node restarts from the beginning at the initialization phase.

3.2 EDETA-enhanced

EDETA-e is a subset of EDETA that allows the engineer to assume control over the network formation and the delays.

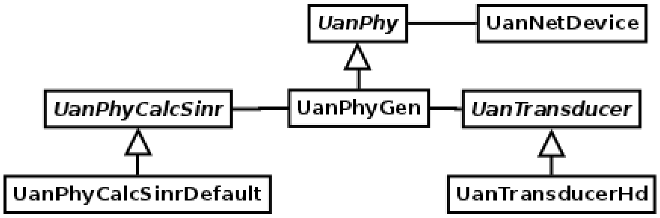


Fig. 1: Simplified class diagram of UAN modified methods

With EDETA-e, the engineer, and not the protocol, decides which node becomes a CH. This way, the engineer decides where to place CH nodes and which type of power supply they will have.

Since CHs are fixed at design time, EDETA-e only considers one initialization phase. After that, all nodes will always remain in the normal operation phase.

4. Protocol Implementation

As stated before, the ns-3 simulator and its Underwater Acoustic Network (UAN) model were used to conduct the experimentation.

EDETA requires some functionality that the last UAN available code was missing at the time when the implementation was done. It needs to be aware of the received packets signal level and needs support to transmit in different radio channels. Figure 1 depicts a simplified class diagram of the modified classes of the UAN model.

The received signal strength is passed to the upper layers attaching a tag to the received packet at the StartRxPacket method of the UAN phy layer.

To implement the different radio channels functionality, absence of inter-channel interference was assumed and an attribute was added to the phy layer class to keep track of the current channel. The packet's channel was also attached to the packet with another tag. It was also necessary to modify the CalcSinrDb, SendPacket and Receive methods form the UanPhyCalcSinrDefault, UanPhyGen and UanTransducerHd classes to correctly handle SNR calculations, send and receive packets.

A Data Link Layer was implemented in order to make the EDETA implementation independent of the MAC layer. This layer extracts the received signal strength of the packet and sends it to the EDETA model along with the rest of the EDETA packet. Sender and destination addresses from the UAN packet model are also extracted and sent to the upper layer.

5. Simulation results

A set of EDETA-e simulations was performed to carry out the adequate tuning of parameters and study the adaptability of the protocol to this new transmission medium.

The protocol was tested in three different square scenarios. The first one with 100 × 100 meters, the second one with 150 × 150 meters and the last one with 200 × 200 meters. In all of them 100 and 200 nodes were randomly deployed, which gave us 6 different scenarios to test. Each scenario has been simulated several times in order to achieve a confidence interval of ±1% with a confidence level of 95%.

All the simulations were seeded using the number 1299602291 and each repetition was done advancing the run number [23].

Leaf nodes start with 150 Joules of energy. The transmission power was set to 0.203 wats, the reception and idle power to 0.024 wats and the sleep power to 3×10^{-6} watts. The transmission range of the nodes was limited to 100 meters adjusting the model transmission power. These values were extracted by A. Sanchez et al. from their low-cost, low-power underwater acoustic modem [24]. The transmission speed was set to 500 bps. This speed could be increased, but we wanted to take into account the speed reduction produced by the use of CDMA spreading codes.

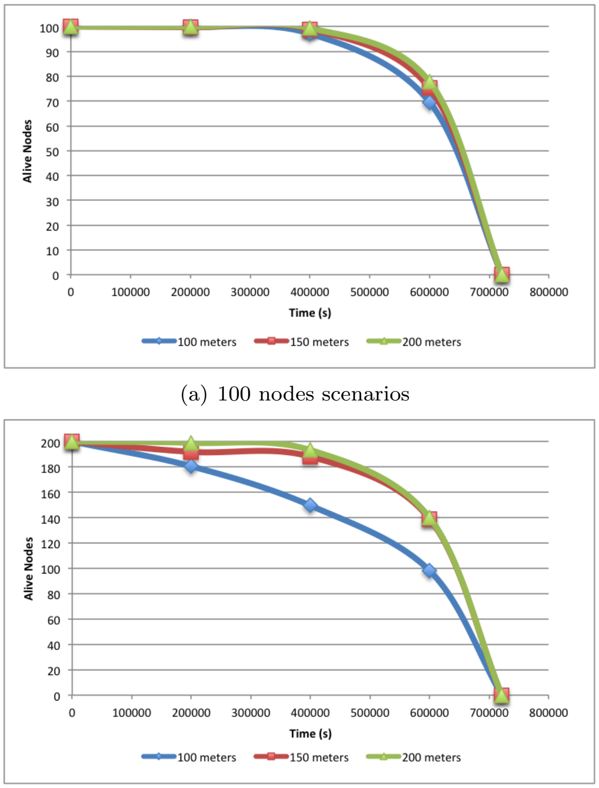
More specific parameters of the EDETA protocol are the $T_{SuperFrame}$ which was set to 250 seconds and the T_{Config} parameter which was set to 9000 seconds.

Each node woke up at its defined time interval and sent one byte with its collected data to its CH. After that, each CH had to aggregate these data and sent it to its parent in the tree structure until the data reached the sink.

Table 1 shows the difference in network lifetime for each scenario. A confidence interval in which the specified average values will reside is given with a confidence level of 95%. The average value of the network lifetime for each scenario is fairly stable, varying at most between ±1.01%. A decrease of the network lifetime can be observed with the increase of the scenario's area. We believe that this decrease is strongly related to the existing node density.

This can be better observed at Figure 2, where a plot of alive nodes over time for the six different scenarios is given. Figure 2(a) depicts these values for the 100 nodes scenarios showing no significant differences between them.

Figure 2(b) shows the evolution of the alive nodes over time for the 200 nodes scenarios and it provides more interesting data to analyse. Although the simulation end time for the three different scenarios is almost the same, the evolution of the alive nodes over time is different. It can be seen that, at the 100 × 100 meters scenario, there are some nodes that deplete their energy almost at the beginning of the simulation and as we increase the scenario area the nodes start to die latter. This behaviour is given by the node density of the scenario. When there is high density, some nodes will remain more time awake at the configuration phase. This leads them to spend more energy, hence they will die sooner than expected.



(a) 100 nodes scenarios

(b) 200 nodes scenarios

Fig. 2: Alive nodes vs. Time with (a) 100 nodes and (b) 200 nodes

The optimum node density will vary for EDETA and EDETA-e with the transmission range. Further studies have to be conducted in order to give more insights in the effect of this parameter. Furthermore, the increase of the transmission range will lead to higher propagation delays and will likely increase the collision probability during the configuration phase.

Figure 3 depicts the accumulated spent energy by all nodes over time for the first 40,000 seconds. Figure 3(a) shows this spent energy for the 100 nodes scenarios. As expected from Figure 2(a) there are no big differences between scenarios. It can be observed that there is one big slope from the beginning of the simulation to 4,000 seconds. Here the energy consumed increases really fast. This part corresponds to the first sub-phase of the initialization phase where normal nodes are trying to join a CH, so there is huge activity in the network. After this first sub-phase is done and the nodes have joined their cluster, only the CHs will remain awake building the tree schedule to send their data to the sink. So, there is really low energy consumption until the end of

Table 1: Network lifetime

Meters	100 × 100		150 × 150		200 × 200	
Nodes	100	200	100	200	100	200
Average	702400 sec.	704000 sec.	693250 sec.	689450 sec.	700550 sec.	693900 sec.
Confidence interval	±0.78%	±0.36%	±0.45%	±0.53%	±1.01%	±0.28%

the initialization phase. After the initialization phase, the normal operation phase begins. In these scenarios it corresponds to the period beginning at 18,000 seconds until the end of the simulation.

At Figure 3(b) the same analysis was done for the 200 nodes scenarios. It depicts differences in the accumulated energy spent by the nodes depending on the area size of the scenario, corroborating the results given by Figure 2(b). As can be seen, the increase in energy consumption happens at the initialization phase and after that, during the normal operation phase, the slope is similar for all the scenarios. This shows that node density is a determining factor at the initialization phase but, it has not a big influence on the normal operation phase.

This increase in the energy spent given by the node density comes with an increase of packet collisions at the initialization phase. New MAC protocols may avoid some of these collisions and thus, increase EDETA network lifetime.

When the network simulation is finished, no more collisions appear, so 100% data packet delivery rate to the sink is achieved.

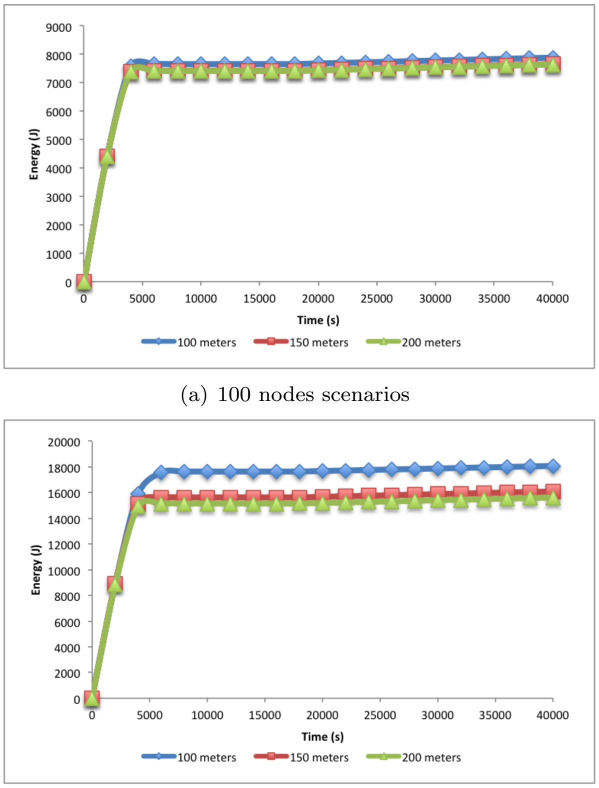
Finally, at Figure 4 we study the effect of the $T_{SuperFrame}$ parameter in the network lifetime. We performed several simulations increasing its value from 250 seconds up to 800 seconds. As expected, network lifetime grows linearly with this parameter. We calculated a linear regression and obtained the expression given by (3) with $R^2 = 0.999$,

$$y = 2523.5T_{SuperFrame} + 88841 \qquad (3)$$

Where R^2 is the correlation coefficient and y the estimated network lifetime. This predictable energy consumption allows us to accurately anticipate the network lifetime in absence of failures. For simplicity, just results for 100 nodes and 100 × 100 meters scenario are shown, since results for the other five scenario behave in the same way.

6. Conclusions

It has been presented an application of a novel routing protocol for underwater wireless sensor networks. EDETA is a power-aware routing protocol which tries to minimize the energy consumption organizing nodes in



(a) 100 nodes scenarios

(b) 200 nodes scenarios

Fig. 3: Energy spent vs. Time with (a) 100 nodes and (b) 200 nodes

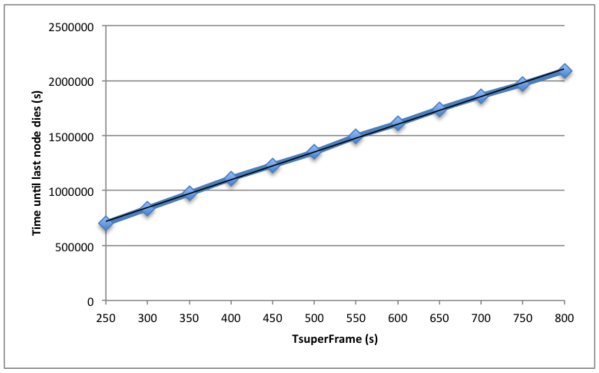


Fig. 4: Effect of the $T_{SuperFrame}$ on the network lifetime

clusters and using low-power modes at the times in which nodes have no need to be awake. In addition, as explained at [7], [8] the protocol adds fault tolerant mechanisms and has time-constrained properties.

The behaviour of EDETA-e has been studied in the subaquatic medium by means of simulation in ns-3. The results show high reliability with no data packet loss and an optimal energy management during the normal operation phase, allowing the nodes to remain in a low-power state when they have no data to deliver to the sink.

The performance evaluation of EDETA-e protocol shows a stable and efficient behaviour of clusters and tree structures. Moreover, paths can be dynamically adapted to topology changes and node failures, offering the maximum energy saving without exact location information.

As a consequence of these studies and results, it can be concluded that EDETA-e is a very suitable protocol for subaquatic sensor networks, presenting, in addition, new features in this field.

Future work will include a full implementation of the EDETA protocol on a UWSN and a better optimization of the initialization phase. Moreover, further study of the influence of the node density and its relation with the transmission rage has to be done. In addition, new MAC protocols, aside the ALOHA protocol used in this work, should be studied in order to try to reduce packet collisions the initialization phase.

Acknowledgements

The authors gratefully acknowledge financial support from the CICYT (research projects DPI2007-66796-C03-01 and PET2008-0011).

References

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, *State-of-the-art in protocol research for underwater acoustic sensor networks.* New York, New York, USA: ACM Press, Sept. 2006.

[2] M. Stojanovic, "On the relationship between capacity and distance in an underwater acoustic communication channel," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 4, p. 34, Oct. 2007.

[3] R. Urick, *Principles of underwater sound.* McGraw-Hill, 1983.

[4] L. Lanbo, Z. Shengli, and C. Jun-Hong, "Prospects and problems of wireless communication for underwater sensor networks," *Wireless Communications and Mobile Computing*, vol. 8, no. 8, pp. 977–994, 2008.

[5] D. Pompili and I. Akyildiz, "Overview of networking protocols for underwater wireless communications," *Communications Magazine, IEEE*, 2009.

[6] Y. Xiao, *Underwater Acoustic Sensor Networks.* CRC Press, 2010.

[7] J. V. Capella, A. Bonastre, R. Ors, and S. Climent, "A New Energy-Efficient, Scalable and Robust Architecture for Wireless Sensor Networks," in *2009 3rd International Conference on New Technologies, Mobility and Security.* IEEE, Dec. 2009, pp. 1–6.

[8] J. V. Capella, "Wireless Sensor Networks: A new efficient and robust architecture based on dynamic hierarchy of clusters," Ph.D. dissertation, Universitat Politècnica de València, 2010.

[9] "Web page of the Network Simulator 3 project," 2011. [Online]. Available: http://www.nsnam.org

[10] N. Chirdchoo, W. Soh, and K. Chua, "Aloha-based MAC protocols with collision avoidance for underwater acoustic networks," *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, 2007.

[11] C. L. Fullmer and J. Garcia-Luna-Aceves, "Floor acquisition multiple access (FAMA) for packet-radio networks," *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, 1995.

[12] M. Molins and M. Stojanovic, "Slotted FAMA: a MAC protocol for underwater acoustic networks," *OCEANS 2006-Asia Pacific*, 2007.

[13] A. Syed, W. Ye, and J. Heidemann, "T-Lohi: A new class of MAC protocols for underwater acoustic sensor networks," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE.* IEEE, 2008, pp. 231–235.

[14] D. Pompili, T. Melodia, and I. F. Akyildiz, "A CDMA-based medium access control for underwater acoustic sensor networks," *IEEE Transactions on Wireless Communications*, 2009.

[15] O. Aldawibio, "A review of current routing protocols for ad hoc underwater acoustic networks," *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the*, pp. 431–434, 2008.

[16] G. G. Xie and J. Gibson, "A network layer protocol for UANs to address propagation delay induced performance limitations," in *MTS/IEEE Oceans 2001. An Ocean Odyssey. Conference Proceedings (IEEE Cat. No.01CH37295).* Marine Technol. Soc, 2001, pp. 2087–2094.

[17] K. Foo, P. Atkins, T. Collins, C. Morley, and J. Davies, "A routing and channel-access approach for an ad hoc underwater acoustic network," *OCEANS'04. MTTS/IEEE TECHNO-OCEAN'04*, vol. 2, pp. 798–795, 2005.

[18] M. Zorzi and P. Casari, "Energy-efficient routing schemes for underwater acoustic networks," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 9, pp. 1754–1766, 2008.

[19] P. Xie, J. Cui, and L. Lao, "VBF: vector-based forwarding protocol for underwater sensor networks," *Networking 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, pp. 1216–1221, 2006.

[20] H. Tiansi and F. Yunsi, "QELAR: A Machine-Learning-Based Adaptive Routing Protocol for Energy-Efficient and Lifetime-Extended Underwater Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 6, pp. 796–809, June 2010.

[21] D. Pompili, T. Melodia, and I. Akyildiz, "A resilient routing algorithm for long-term applications in underwater sensor networks," *Proc. of Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2006.

[22] P. Wang, C. Li, and J. Zheng, *Distributed Minimum-Cost Clustering Protocol for UnderWater Sensor Networks (UWSNs).* IEEE, June 2007.

[23] *ns-3 simulator reference manual*, 2011. [Online]. Available: http://www.nsnam.org

[24] A. Sanchez, S. Blanc, P. Yuste, and J. J. Serrano, "A low cost and high efficient acoustic modem for underwater sensor networks," in *OCEANS'11 IEEE SANTANDER*, 2011.

Variable Ranges Security Protocol for Wireless Sensor Networks

F. Bagci, A. Khalifeh, G. Jung, and C. Sturm
Department of Computer Science and Engineering, German University in Cairo, Egypt

Abstract— *Wireless sensor network applications pose novel challenges to networking and protocol design. Generally, traditional wireless devices directly link to a base station or beacon within their transmission range to facilitate ensuring integrity and security of communication, only rarely they make use of any multi-hop messaging. Wireless sensor networks, in contrast, are often intrinsically based on a peer to peer infrastructure, where messages inevitably traverse long distances through the network, passing multiple nodes to reach a certain destination.*

With the sensitivity of the communicated data (e.g., in safety or privacy critical areas such as hospitals or power plants) the required level of security increases, making effective security mechanisms a prime concern of the protocol design. This higher level of security, in the context of the limited resources and the high distributivity of ad-hoc sensor networks, gives raise to complex issues in communication protocol design.

In this paper we present a novel security architecture for wireless sensor networks. The focus of our protocol is twofold: First, a cluster-based organization of the network with few more powerful cluster nodes that implement required security features for a large number of simple sensor nodes provides for confidentiality and integrity of sensible data. Second, a dynamic transmission range adaption protocol substantially prolongs the lifetime of the nodes through energy efficient communication without significantly decreasing the node connectivity.

1. Introduction

The study of sensor networks, while being a research field in and of itself, forms a basis for various areas where the collection of environmental data through sensors is essential (e.g., security, traffic control, ubiquitous computing, etc.). The combination of sensing/sensoring, computational aspects, and communication solutions provides for a broad range of applications such as as smart hospitals, intelligent battlefields, earthquake response systems, and learning environments [1] [2]. Generally, the term *sensor network* has come to describe a dynamically self-organizing collaborative network of widely distributed, tiny, low-cost, sensoring nodes (“smart dust”) that are capable to cover an area and automatically communicate the collected data to a beacon or base station over multi-hop paths.

Sensor nodes are usually tiny, self contained, battery powered devices. Under normal circumstances it is impossible to replace or recharge these batteries, therefore the lifetime of a wireless sensor network is intrinsically restricted by the initially available power in each individual node, making power consumption considerations an essential part of any new protocol design. Similarly, very small memory, low processing power, and a limited communication bandwidth, all in comparison to traditional wireless devices, further restrict the options. Also, high failure rates, occasional shutdowns, and sporadic communication interference force continuous dynamic changes upon the topology. Finally, the sheer number of individual sensing devices of a sensor network, ranging from hundreds to thousands, make it infeasible to rely on previous solutions of ad-hoc networking protocols such as. For example flooding-based standard routing schemes for ad hoc networks simply do not scale adequately [3].

As in most communication systems security becomes more and more important in wireless sensor networks. The aim of a security architecture is to ensure confidentiality, integrity and availability. Confidentiality is given if any kind of unauthorized access to data is prevented. Integrity means that data cannot be modified or deleted without being detected. For a sensor network to serve its purpose, the data must be available when it is needed. A security architecture can never cover all types of threats simultaneously. The application determines which attacks are probable in a certain scenario. A comprehensive security architecture can increase protection, but on the other hand this would lead to undesirably high hardware costs. It also may increase energy requirements of sensor nodes significantly due to several successive protocols.

In [4] we proposed a security architecture for wireless sensor networks called *SecSens* which fulfills security requirements on multiple levels. SecSens focuses mainly on three security aspects: key management, secure routing, and verification of sensor data. The sensor network in SecSens consists of clusters, each containing a number of simple sensor nodes and one powerful node that acts as a cluster-head. Sensor nodes connect directly to the cluster-head, i.e. routing in clusters is not necessary. A node can be a member of several clusters at the same time. All cluster-heads form together an inter-cluster network used for sending messages to base stations. We assume that sensor nodes do not change their position once they are attached to a location. SecSens works with multiple base stations in order to avoid single-point-of-failures (see Figure 1). In the first version of SecSens clusters were built in the initial phase and remained unmodified for the whole lifetime of the network. Furthermore, all nodes used the same sending configuration, i.e. transmission power and range were set to maximum on each node. This paper describes an extension of this approach with dynamic features. The new variable ranges (VR) security protocol optimizes the communication range of each cluster-head. This optimization results in more efficient usage of energy throughout the overall sensor network.

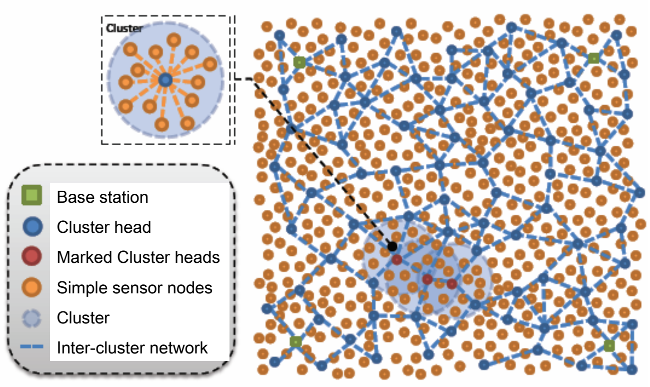


Fig. 1: Basic sensor network architecture

The next section describes related security and energy efficient communication approaches for sensor networks. Section 3 introduces the variable ranges security protocol. We evaluated the new architecture in a wireless sensor network simulator. Section 4 presents the evaluation results. The paper ends with the conclusion.

2. Related Work

The proposed architecture in this paper combines security relevant features with energy efficient communication. There are several security architectures in the area of wireless sensor networks that disregard energy concerns. And energy efficient protocols are often not protected against a broad range of attacks. We want to present here some security architecture and as well as communication protocols for wireless sensor networks with low energy consumption.

In [5] a security protocol for sensor networks called *SPINS* was presented for hierarchical sensor networks with one or more trustworthy base stations. SPINS consists of two parts: a secure network encryption protocol (SNEP) and authenticated broadcasts (μTESLA). SNEP provides the security properties confidentiality, authenticity, integrity, and timeliness. Each sensor node receives on a secure channel an individual, symmetric master key, which is only known by the base station and the node. Using this master key the sensor node is able to generate all keys. The disadvantage of SNEP is that secure communication can be built only between a base station and nodes, and it is not possible to protect the communication in or between clusters. The second part of SPINS is μTESLA that provides sending of authenticated broadcasts. For symmetric encryption, sender and receiver must share the same secret. Consequently, a compromised receiver is able to act as a designated sender by transferring forged messages to all receivers. μTESLA uses delayed disclosure of symmetric keys for generating an asymmetry between a sender and a receiver. This approach requires weak time synchronization of the sender and the receiver in order to achieve time shifted key disclosure. Storage cost increases because each node has to buffer packets which can be only verified after receiving the key in the future time-slots. Also, this causes new possibilities for DoS-attacks. An attacker can force a buffer overflow by sending planned broadcasts. Furthermore, μTESLA leads to scalability problems, which are described in [6].

[7] suggests an adjusted key distribution for different security requirements. For this reason, four different kinds of keys are used. The individual key is similar to the master key of SPINS. The second kind of key is a pair-wise shared key, which is generated in the initial phase for each known neighbor. Furthermore, the nodes have a cluster key for secure communication between cluster members. The last key is the group key that is used for secure broadcasting. This approach provides more flexibility but contains a security risk during the initial key distribution phase.

The Sensor-MAC (S-MAC) protocol [8] is an energy efficient communication protocol for wireless sensor networks. S-MAC is a slot-based protocol where each sensor node has alternating sleep and awake phases. The network is divided into clusters and all members of a cluster are awake or asleep at the same time. All cluster nodes exchange schedules in an initial phase. Within a cluster only one schedule is used, i.e. the schedule of the first node that sent a schedule. If a node receives multiple schedules it follows all of them. Such nodes have a higher energy consumption. Within an awake phase all cluster nodes contend with each other for medium access. The contention mechanism of S-MAC is the same as that in IEEE 802.11, i.e., using RTS (Request To Send) and CTS (Clear To Send) packets. S-MAC needs a strict timer synchronization in order to achieve correct functionality. Periodic synchronization among neighboring nodes is performed to correct their clock drift. An extension of S-MAC by *adaptive listening* is described in [9]. If a node A notices an ongoing communication of node B whom it wants to send a message, it sleeps the time until B is ready.

A modification of S-MAC called Timeout-MAC or T-MAC is introduced in [10]. In S-MAC all nodes need to be awake in the contention phase even if they have nothing to send or receive. T-MAC uses a specific timer T_A to shorten the awake phase if the node does not need to communicate. Obviously the T_A is smaller than the contention phase, thus the energy consumption is reduced. But if the timer T_A is chosen too small, the node sleeps early missing possible message requests of other nodes. This *early sleeping problem* could even lead to unfairness. In further extensions of T-MAC this problem is solved by *future request to send (FRTS)* messages, but this increases again the energy consumption. Nevertheless T-MAC gains better energy results compared to S-MAC.

The *Wireless Token Ring Protocol (WTRP)* [12] was developed for mobile ad-hoc networks. All nodes build a single ring in the initial phase. The aim of WTRP is to maximize the throughput and minimize the latency without restraining the mobility. Energy efficiency is not considered in WTRP because the nodes are mobile devices with strong energy resources like Laptops or PDAs. A mapping of WTRP on wireless sensor networks is *E^2WTRP* that is described in [13]. E^2WTRP aims to enhance the energy balance by dynamic adaptation of the token holding time. An active node can send more messages if the token holding time is increased. The frequency of token hand-over is decreased at the same time that reflects in lower energy consumption. *ESTR* [14] is an energy saving token ring protocol for wireless sensor networks that introduces sleep periods for nodes which does not need to send or receive messages. This leads to a very good energy balance.

3. Variable Ranges Security Protocol

The energy consumption of sensor nodes that send with maximum transmission power lies significantly higher than with reduced power. Decreasing transmission power results in exponentially decreased energy consumption. Therefore, the Variable Ranges (VR) protocol saves energy by adjusting and optimizing the signal strength to particular circumstances of the sensor network in order to extend the lifetime of the sensor nodes. Additionally, the initial state of reduced transmission power of the nodes ensures that complexity of network is low. With high signal strength nodes are confronted with frequent interferences and redundant paths within the network. Low signal strength means that number of neighboring nodes is less, i.e. probability for message collisions decreases.

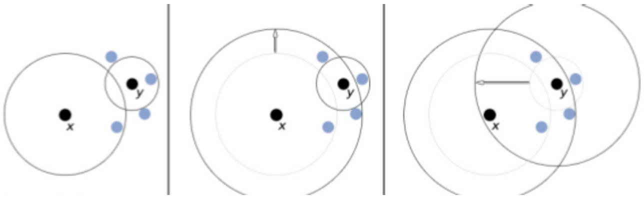


Fig. 2: Range adjustment in VR protocol

Regarding the security architecture, the number of neighbors is also an important parameter. Each cluster-head has to manage several keys with each other neighboring sensor node and cluster-head for securing the communication and ensuring authentication. More neighbors means higher management effort and more storage space, as well as more encryption and decryption processing. All of this result again in increased energy consumption. Therefore, it is essential that the security architecture works hand in hand with the underlying communication protocol.

In the initial phase of the VR protocol, the nodes search for neighboring nodes starting with a minimum signal strength. For this reason, each node sends a *DiscoverNodes* message containing its own range parameter and ID. Then the node waits a certain time to get a response. The waiting time is also dynamic, i.e. the time is low, if the range is low and increases, if the range increases. The reason for this is that a node with a low range will reach less neighbors. Therefore, there is no need to wait a long time for a response. The nodes increase stepwise their transceiver power and send new discovery messages until a pre-configured number of nodes is found. A node which receives a *DiscoverNodes* message of an unknown node, extracts the range information of its seeking new neighbor. In the next step, the node compares the received range value with its own range. If the own range is lower, the node increases its range and sends a *DiscoverNodesReply* message back. Since the signal strength of the nodes would increase in this way until all nodes would settle at the range of the largest distance between two nodes, the VR protocol performs only a temporary range adjustment. This means that nodes discard the adaptation after a certain time and return again to their previous values.

Figure 2 illustrates this adjustment scheme. Assuming a maximum number of neighbors is set to three in this figure, it would be unfavorable for node *y* to use the range parameter of node *x*, since it can reach three neighbors with a much lower signal strength. For this reason, *y* will only increase range temporarily to answer node *x* and return to its previous range, in order to proceed with its own search. Cluster-heads find in this way a minimum number of other cluster-heads and as well as sensor nodes.

Actually, the aim of each cluster-head is to be reachable through the inter-cluster network by at least one base station. After the initial phase, each base station sends a broadcast through the new built network. This sink message is also important to generate new keys for further authentication. In [4] this key generation is described in detail. Therefore, reachability of base stations is essential to establish the security architecture in the sensor network. If a cluster-head does not receive a broadcast message after a certain time, it starts a new search phase to find new cluster-heads. The cluster-head uses this time a different message *DiscoverNewNodes*. It first uses the current range, since there could be cluster-heads which are in range, but not discovered. Cluster-heads who receive such a message, adjust their signal strength to answer, but keep their new range value this time. If the node does not get any answers, it increases its range and repeats the procedure until it finds new connections. Figure 3 shows the *TryToConnect* phase. You can see on the right side of the figure, that after the initial search phase, several local cluster networks are established, but not all of them can reach a base station marked here as green squares at the four corners. On the left picture you can recognize that the connectivity is enhanced after the *TryToConnect* phase, but nevertheless there are still local clusters remaining unreachable by any base station. The reason is that nodes are deployed randomly. There is a small probability that some nodes cannot reach a base station, even if transmission power is set to the maximum or due to message collisions in the initial phase.

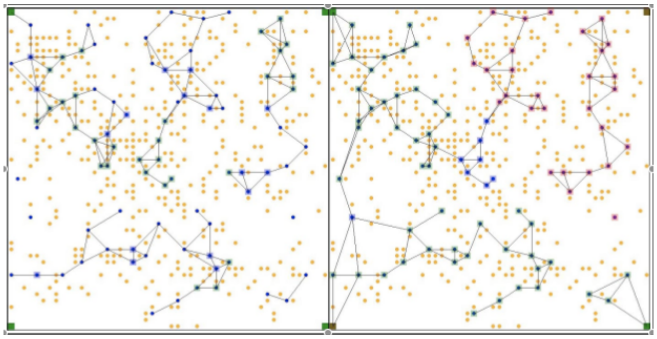
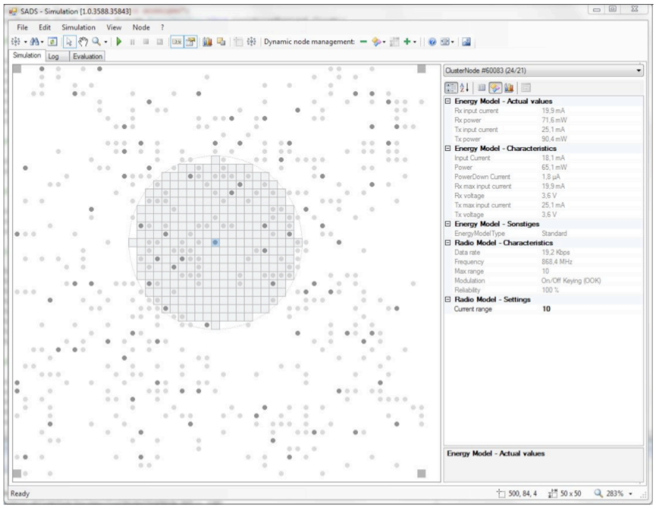


Fig. 3: Initial phases of VR protocol: a) neighbor search b) TryToConnect

Cluster-heads check periodically their neighborhood for node losses or new arriving nodes. During lifetime the VR protocol ensures that nodes can dynamically adapt to changes in their environment. This network adaptation goes hand in hand with security adjustments. Basically, the security architecture contains four components, which interact with each other: authenticated broadcasts, key management, secure routing, and en-route filtering. Authenticated broadcasts ensure that the stated sender is identified as the true sender. Each message contains a *Message Authentication Code (MAC)* generated by using a shared key. In case of only one recipient knowing the shared key, the sender can be easily authenticated. But if there are multiple recipients with the same shared key, potentially each receiver could be the sender, i.e. authentication is not possible. Therefore, our security architecture uses an extension of the MAC approach with key chains and delayed disclosure of keys (see [4]). This ensures the authentication of several base stations and cluster-heads. After each adaptation of VR protocol the cluster-heads distribute group keys to own members used for encryption of group messages and exchange pair-wise shared keys with each member node for one-to-one messages. In this way, every type of message exchange is secured in the network.

Additionally, routing information is updated by cluster-heads after each VR adjustment phase. Simple sensors do not need routing capability, because they exclusively communicate with the cluster-head. Routing is used only within the inter-cluster network established by cluster-heads. Our security architecture uses probabilistic multi-path routing based on the level values to forward messages from cluster-heads on the way to the corresponding base station. Cluster-heads build up a trust matrix, where each transmission to its neighbors is recorded. Based on this trust information, cluster-heads calculate a probability value and write it into the packet header. This value is used to decide in which direction the packet has to be send. Each cluster- head modifies the probability value and sends the message over the most trustworthy route. Furthermore, our architecture provides passive participation, i.e. sensor nodes listen to packet transmissions of their neighbors. If cluster-head *u* detects a packet addressed to its neighbor *v*, and recognizes that *v* is not forwarding the message, *u* takes responsibility with a certain (low) probability. Also, if *u* assumes that *v* forwards the message to a non-existent node, *u* takes care of transferring.



SADS - Simulation [1.0.3588.35843]

ClusterNode #60083 (24/21)

Energy Model - Actual values	
Rx input current	19,9 mA
Rx power	71,6 mW
Tx input current	25,1 mA
Tx power	90,4 mW
Energy Model - Characteristics	
Input Current	18,1 mA
Power	65,1 mW
PowerDown Current	1,8 µA
Rx max input current	19,9 mA
Rx voltage	3,6 V
Tx max input current	25,1 mA
Tx voltage	3,6 V
Energy Model - Sonstiges	
EnergyModelType	Standard
Radio Model - Characteristics	
Data rate	19,2 Kbps
Frequency	868,4 MHz
Max range	10
Modulation	On/Off Keying (OOK)
Reliability	100 %
Radio Model - Settings	
Current range	**10**

Energy Model - Actual values

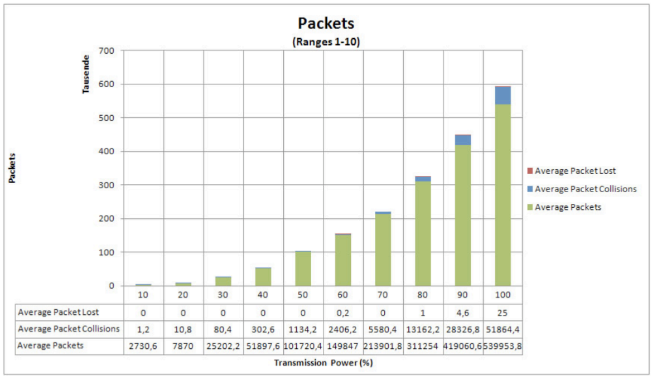
Ready 500, 84, 4 50 x 50 283 %

Fig. 4: The Simulator GUI

Attacks like *report fabrication* or *false data injection* threaten the network by manipulating and infiltrating sensor data. The security architecture prevents such threats using en-route filtering. Cluster-heads generate data reports containing sensor information of cluster members for sending them to base stations. These reports are verified during transfer through the inter-cluster network. Sensor nodes belonging to same cluster report events (sensor data) collectively by generating MACs based on their en-route keys, whereas keys must be chosen from different partitions of a global key pool. These multiple MACs collectively act as the proof that a report is legitimate. Finally, the cluster-head forwards the report to the base station over the inter-cluster network. A cluster-head receiving a report checks, if it has one of the keys, that were used to generate the MACs in the report. With a certain probability, it will be able to verify the correctness of MACs. A report with an insufficient number of MACs will not be forwarded. A compromised node would have access to keys from one partition and could generate MACs of one category only. Since keys and indices of distinct partitions must be present in a legitimate report, the compromised node has to forge the other key indices and corresponding MACs. This can be easily detected by cluster-heads possessing these keys. If cluster-head has none of the keys and the number of MACs is correct, it forwards the report to the next cluster-head. Even if a forged report receives a base station, it can be detected, because base stations know all global keys.

4. Evaluation

To evaluate the efficiency of our variable ranges security architecture we implemented a simulation tool where it is possible to establish different sizes of sensor networks. Figure 4 shows the GUI of the simulator. The simulation is divided into three phases: node distribution, initialization of network, and report sending. In the first phase, a predefined number of nodes is distributed randomly over a given area. Sensor nodes and as well as cluster-heads are deployed setting for each a maximum transceiver range.



	10	20	30	40	50	60	70	80	90	100
Average Packet Lost	0	0	0	0	0	0,2	0	1	4,6	25
Average Packet Collisions	1,2	10,8	80,4	302,6	1134,2	2406,2	5580,4	13162,2	28326,8	51864,4
Average Packets	2730,6	7870	25202,2	51897,6	101720,4	149847	213901,8	311254	419060,6	539953,8

Fig. 5: Traffic load in relation to signal strength

Basic parameters for the network are total number of nodes, initial node range, initial node energy, and network density. Type, range, and position of nodes can be changed easily using the simulator GUI. Furthermore, new nodes can be added or existing nodes can be deleted before the next phase of the simulation is started. Figure 4 shows a screenshot after the first phase. Dark circles are cluster-heads whereas light dots are simple sensor nodes. The squares at the corners represent again four base stations. In this case one cluster-head is selected and you can see the communication range of the current node.

The second phase initializes the network based on a communication protocol. We implemented three protocols that can be selected by the user in the beginning of this phase. These are the SMAC protocol, the energy saving token ring protocol (ESTR), and the variable ranges (VR) protocol. The user can change a set of parameters depending on the selected communication protocol, e.g. cluster size, timer settings, update periods. In this phase security and routing information is exchanged too. At the end of initialization, the network is established and nodes can start to exchange secured messages. This is simulated in the last phase by randomly generated reports that are sent to base stations. The user can halt the simulation at any time, in order to change parameters for nodes. For example, one can turn off a node to simulate a node loss. It is also possible to simulate a compromised node that sends false reports into the network.

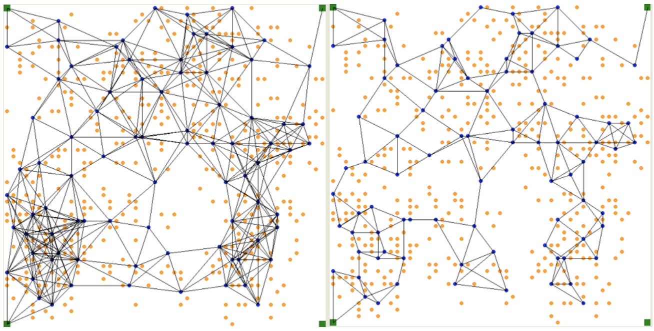


Fig. 6: Network complexity without (left) and with VR protocol (right)

Nodes consume energy for processing data, like encryption and decryption, and sending or receiving messages. For some communication protocols nodes can switch to a sleep mode, where energy consumption is minimal. Our simulator bases on an energy model that uses specifications of real sensor boards: ESB 430/1 and MSB-430 of Freie University Berlin [15].

In a first evaluation we measured the number of message sent in the initialization phase using the VR protocol. We performed several simulations where we modified the maximum range of nodes in order to get average number of sent packets, collisions, and lost packets. Figure 5 shows the results of a network with 500 nodes. Traffic load increase with higher range of single nodes, because nodes can reach more neighbors to exchange messages with.

As mentioned in the previous section, the complexity of sensor network is much lower using VR protocol. On the left side of Figure 6 the network was established with maximum node range. It is clearly seen that in dense areas of the network, the number of different connections is rather high. In a second simulation, we used the VR protocol to establish the network. Each node increased range starting from minimum until at least three cluster-head neighbors and three sensor neighbors were found, denoted by the annotation *VR 3-3*. As seen on the right part of Figure 6 using the VR protocol lowers the complexity of the network.

An important value for the network is connectivity, i.e. the percentage of nodes that can be reached by at least one base station. A %100 connectivity means that every sensor node can send its data to at least one base station. In Figure 7 we measured the connectivity of a network with 500 nodes in relation to maximum signal strength. If the signal strength is set to lower than %50, only a small part of network is connected. Actually, only clusters near to a base station get connected. But the connectivity increases considerably with higher signal strength and reaches nearly full connectivity after %70 transmission power.

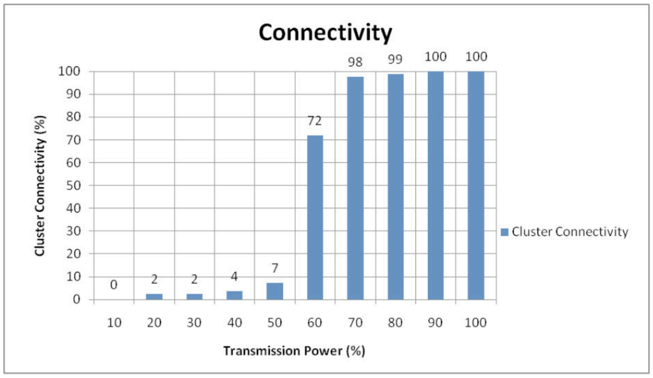


Fig. 7: Network connectivity in relation to signal strength

The optimal usage of signal strength in VR protocol shows its advantages also in energy consumption. Figure 8 illustrates the energy consumption for sending reports from a sensor node to the base station. Level represents here the distance between sending node and nearest base station, e.g. level 6 means that messages traverse six intermediate cluster-heads until they reach the base station. We performed the energy measurement in four different networks with the same size. For the first three networks the maximum range parameter of each node was set to a fixed value, i.e. range 6 stands for %60 maximum signal strength. The last network used the VR protocol with at least three cluster-head and three sensor node neighbors and a further optimization step to increase the connectivity (*VR 3-3 ExtCon*). One can clearly see that the VR protocol has the best energy balance leading to a longer lifetime of the network.

5. Conclusion

This paper presented the variable ranges security protocol for wireless sensor networks. By dynamically adapting the range of each node, the network can be established with low complexity, but still with high connectivity. Since nodes do not sent messages with full transmission power, the energy consumption decreases considerably. This results in an extended lifetime of the overall sensor network. The security architecture consists of four components, which interact with each other: authenticated broadcasts, key management, secure routing, and en-route filtering. The VR protocol positively influences security functions, since nodes have fewer communication partners to deal with.

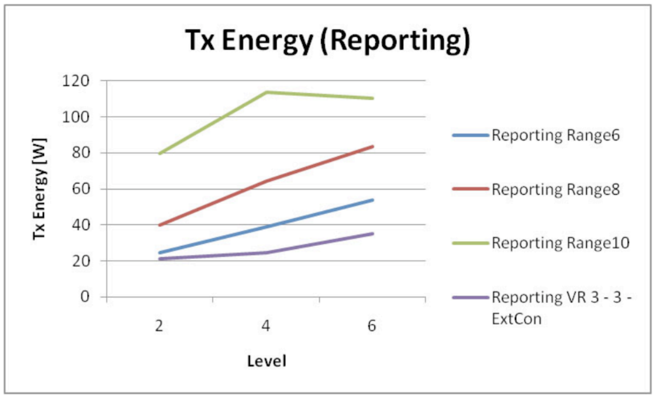


Fig. 8: Energy consumption for reporting

References

[1] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," in *ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02)*, Atlanta, GA, USA, September 2002.

[2] G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Communications of the ACM*, vol. 43, no. 5, pp. 51–58, 2000.

[3] P. Downey and R. Cardell-Oliver, "Evaluating the Impact of Limited Resource on the Performance of Flooding in Wireless Sensor Networks," in *Proceedings of the 2004 international Conference on Dependable Systems and Networks*, Washington, DC, USA, June 2004.

[4] F. Bagci, T. Ungerer, and N. Bagherzadeh, "Multi-level Security in Wireless Sensor Networks," *International Journal On Advances in Software*, vol. 4, no. 6, 2010.

[5] Adrian Perrig, Robert Szewczyk, J. D. Tygar, Victor Wen, and David E. Culler, "SPINS: Security Protocols for Sensor Networks," *Wireless Networks*, vol. 8, no. 5, pp. 521–534, 2002.

[6] Donggang Liu and Peng Ning, "Multilevel μTESLA: Broadcast authentication for distributed sensor networks," *Trans. on Embedded Computing Sys.*, vol. 3, no. 4, pp. 800–836, 2004.

[7] Sencun Zhu, Sanjeev Setia, and Sushil Jajodia, "LEAP: efficient security mechanisms for large-scale distributed sensor networks," in *CCS '03: Proceedings of the 10th ACM conference on Computer and communications security*, New York, NY, USA, 2003, pp. 62–72, ACM Press.

[8] Wei Ye, John Heidemann, and Deborah Estrin, "An Energy-Efficient MAC Protocol for Wireless Sensor Networks," in *Proceedings of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, New York, USA, June 2002, vol. 3, pp. 1567–1576.

[9] W. Ye, J. Heidemann, and D. Estrin, "Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 3, pp. 493–506, June 2004.

[10] Tijs van Dam and Koen Langendoen, "An adaptive energy-efficient MAC protocol for wireless sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*, Los Angeles, California, USA, Nov. 2003, pp. 1567–1576.

[11] IEEE, *IEEE CS, Token Ring Access Method and Physical Layer Specifications. ANSI/IEEE Standard 802.5*, 1985.

[12] M. Ergen, D. Lee, R. Sengupta, and P. Varaiya, "WTRP - Wireless Token Ring Protocol," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1863–1881, Nov. 2004.

[13] Zhenhua Deng, Yan Lu, Chunjiang Wang, and Wenbo Wang, “E^{2}WTRP: An Energy-Efficient Wireless Token Ring Protocol,” in *Proceeding of the IEEE conference on Personal, Indoor, and Mobile Radio Communications*, Barcelona, Spanien, 2004, pp. 398–401.

[14] F. Bagci, T. Ungerer, and N. Bagherzadeh, “ESTR - Energy Saving Token Ring Protocol for Wireless Sensor Networks,” in *Proceedings of the International Conference on Wireless Networks (ICWN ’08)*, Las Vegas, NV, USA, July 2008.

[15] http://www.scatterweb.com, *ScatterWeb Homepage*, 2007.

Empirical Study of Mobility effect on IEEE 802.11 MAC protocol for Mobile Ad-Hoc Networks

Mojtaba Razfar and Jane Dong

mrazfar, jdong2@Calstatela.edu
Department of Electrical and computer Engineering
California State University Los Angeles

ABSTRACT

To design an efficient and effective MAC layer protocol for Mobile Ad-Hoc networks is a challenging task. IEEE 802.11 MAC protocol, which supports ad hoc network mode, provides a good reference for research work in this area. In the recent years, many researchers have investigated the performance of IEEE 802.11 on MANET both theoretically and empirically. However, the impact of the mobility on MAC layer design has not been evaluated thoroughly. In our research, we used OPNET simulator to analyze the performance of IEEE 802.11 MAC protocol under various mobility patterns for different network topologies. The findings revealed interesting correlation between the speed of movement/ the mobility pattern and key network performance parameters including delay and throughput. We also investigated the impact of mobility on fairness issues in media access. The empirical study presented in this paper will be useful to enhance the MAC design of MANET with median or high mobility nodes.

Keywords: MANET (Mobile Ad-Hoc Networks), Mobility, (Medium Access Control) MAC, IEEE 802.11, OPNET

1. INTRODUCTION

Mobile Ad Hoc Networks (MANET) are becoming more and more popular due to its ability to offer convenient, flexible and low cost network service for many non-traditional applications. Unlike the widely used Wi-Fi network which relies on the access point to attach to the existing networking infrastructure, MANET is infrastructure-less, where each node acts as a sender, receiver, and router. While the freedom to deploy a mobile ad hoc network at anytime anywhere is very attractive, to make such network function properly presents a lot of technical challenges. For MAC layer protocols, the well-known challenges are imposed by hidden and exposed terminal problems, fairness access issues, limited bandwidth, limited power supply, as well as limited transmission range and mobility.

IEEE802.11was primarily designed for WLAN, but it also supports ad hoc network mode. The MAC layer protocol in IEEE 802.11 laid the foundation for many proposed MAC layer protocols for MANET. Therefore, it is worthwhile to evaluate the performance of IEEE 802.11 on MANET to see how to improve the design. Many existing research [1-3] focused on the effectiveness on handling hidden and exposed terminal problem, and some addressed fairness issues. Mobility, although an important design factor of MANET, its impact on MAC layer performance has not been fully analyzed yet. Most of the current researches that investigated the mobility effect are focused on the network layer since it is a major concern in routing [4-5]. In [6], the authors briefly compared the performance IEEE 802.11 and other MAC protocols under network scenarios with mobility. However, to develop a full understanding of the mobility effect on MAC layer performance including delay and throughput, fairness, collision probability, a more comprehensive and in-depth study is necessary. The objective of our research is to conduct such study using OPNET [7] to show how mobility impacts the key MAC layer performance parameters.

In this paper, we will present our findings of the empirical study using OPNET simulation. Due to the nice property of OPNET, it is possible to set up different network scenarios with different mobility patterns, which allowed us to better study the impact of various factors including speed, transmission range, and moving trajectory. The network parameters that were taken into account in our study includes delay, throughput, collision count, overhead of control traffic, and backoff time. The documented results will be useful to enhance the MAC design of MANET with different mobility.

The paper is organized as follows. Section 2 provides a brief overview of IEEE 802.11 and highlights the important design issues. The empirical study exploring the mobility effect using the OPNET software is presented in section 3. Experimental results are described in this section as well. Finally, we will conclude our findings in Section 4.

2. OVERVIEW OF IEEE 802.11 MAC PROTOCOL

IEEE 802.11 MAC layer protocol is referred to as Distributed Coordination function "*DCF*" which was based on virtual carrier sensing and the physical carrier sensing [8]. IEEE 802.11 DCF uses RTS/CTS/DATA/ACK when the size of the data frame is large enough; it may just use carrier sense or it may use both methods referred to as CSMA/CA with RTS/CTS as a MAC protocol. The three major issues related to MAC layer protocol over MANET are the ability to handle hidden and exposed terminal problems, the ability to ensure fair access of multiple stations, and the ability to cope with mobility.

2.1. Hidden and Exposed Terminal Problems

These two problems have become a major issue in MANET. Hidden terminal problem [3] occurs when two stations are out of the range of each other and trying to send to the same receiver. As a result, the effect significantly decreases the throughput and makes the delay longer. Exposed terminal problem, on the other hand, is when a node is blocked from transmission to the other stations due to the transmission of the adjacent node. This will cause collision and bandwidth waste (less spatial reuse) and will bring up the starvation problem of the unlucky node. IEEE 802.11 DCF proposed the RTS/CTS handshaking method in order to alleviate the negative impact of these issues on the whole network. In the literature, several schemes have been proposed to solve the hidden and exposed terminal problems using different mechanism. In [9], the authors explored the IEEE 802.11 MAC protocol with and without using RTS/CTS handshaking method. The total WLAN retransmissions, data traffic sent/received, WLAN Delay factors of the whole network was investigated using both methods. They demonstrated that in the scenarios that the Hidden terminal problem exists, it will be a good idea to use this option as it decreases the delay of the network dramatically. They also mentioned that this handshaking method is not necessary to be used where the hidden nodes are not present due to the overhead that it adds to the network. The mobility factor was investigated when the hidden nodes exist. However, the speed of the nodes and the location of them have not been studied in this paper. MACA [10] on the other hand, did not use the carrier sensing option and instead, it used the RTS/CTS/DATA handshake to reserve and use the channel. Although this protocol was a simple design, the control channel collisions made the scheme not effective in the MAC layer. Moreover, in these papers, they never spoke of the effect of the Mobility of the performance of the whole network.

2.2. Fairness issue

Another important factor that should be considered in designing MAC protocols is to make sure that all nodes have fair access to the channel in order to transmit their data. So far most of the single channel MAC schemes rely on the back-off procedure. Upon collision, the mobile nodes will go through the back-off procedure and will try to retransmit after a certain amount of time. Because the backoff time is different for different nodes, some nodes may have more chance to transmit than the others and they are favored in data transmission. This will result in starving problem of the unlucky nodes with long contention window size. Therefore, designing good strategies for back-off procedures and providing fair chances among nodes to access the channel is one of the important aspects in MANET. MACAW [11] for wireless LANs is another single channel schemes which tried to improve the performance of MACA protocol. A five handshake RTS/CTS/DS/DATA/ACK has been used in this protocol which leads to alleviation of the hidden and exposed terminal problem and better fairness among nodes. By using a different back-off approach (MILD), this protocol allowed the nodes to access the channel in a fair manner which is more desirable in ad hoc networks. However, the effect of mobility on fairness issue using this protocol has not been investigated.

2.3. Mobility issue

For the infrastructure-based networks, the access point has the major influence on the delivery of the data to the destination. Within a Basic Service Set (BSS), the stations have to share information using the access point and therefore their position towards each other is not that important. Hence, the mobility of nodes does not have a major effect on the MAC layer protocol [9]. For infrastructure-less network as MANET, the mobile nodes are in direct contact with each other. Since they can be sender, receiver and router, mobility has a significant impact on the performance of their data delivery. One may wonder what influence may the mobility of the nodes cause on the performance of the network. Will mobile nodes be treated the same way as they move? Will the efficiency of the transmission stays the same as the mobility varies? How will the delay and overall throughput be affected via different mobility pattern? Can we enhance the network performance using the handshaking RTS/CTS method under high mobility? Most of the questions do not have a solid answer yet. In our research, we will explore the relationship between mobility and all these factors using OPENT simulation. The results presented in this paper will shed some light to answer some of the questions related to the impact of the mobility on MANET networks.

3. EMPIRICAL STUDY USING OPNET

3.1. OPNET Simulator

OPNET modeler is one of the powerful simulation software allowing the users to implement different network topologies using a friendly graphic user interface. As lots of research papers in networking field used NS-2 simulator [12], OPNET makes it easier to use as it provides ready-to-use components without the need of writing codes to create real time network simulations. It also provides the flexibility for advanced users to create their own network node and link by hard coding. For our research, OPNET is selected since its Wireless Modeler includes a rich library of detailed mobile protocols and application models that can be utilized to create MANET with various mobility patterns.

3.1.1. MANET and Mobility in OPNET

OPNET [1] uses the IEEE 802.11 MAC protocol with DCF for Mobile Ad-Hoc Networks. RTS/CTS handshaking option is also included in case a user decides to implement it. The software has different objects for MANET networks such as the MANET station, MANET work station, and Mobility configuration options in order to set up the movement of the nodes. In fact, Mobility is one of the most valuable options that are included in the simulator so that the users can easily define the way the stations move. The speed of the stations can also be easily defined for various applications. This option makes it simpler in real time simulations in comparison to the other simulators where the mobility is a difficult task to define and implement. Figure 1 illustrates a MANET scenario with pre-defined node mobility. The statistics that are related to this work are explained briefly as follows:

1) MANET delay: the end to end delay of MANET packets for the whole network (seconds).
2) Throughput: the total number MANET traffic which is received in bits per second by all the MANET receivers.
3) Media Access delay: The global statistic for the total of queuing and contention delays of the data, management, delayed block-ACK and Block-ACK frames transmitted by all WLAN MACs in the network (seconds).
4) Back off slots: the number of slots that a stations needs to back off before transmission while contenting for the medium, and the number of slots in the contention window after the successful transmission of the station.
5) Retransmission attempts: the total number of retransmissions by all the WLAN MACs in the whole network until the delivery of the packet or being discarded as a result of reaching the short or long retry limits. We used this factor to study the impact of mobility on the collision counts as well as the effectiveness of RTS/CTS handshake.

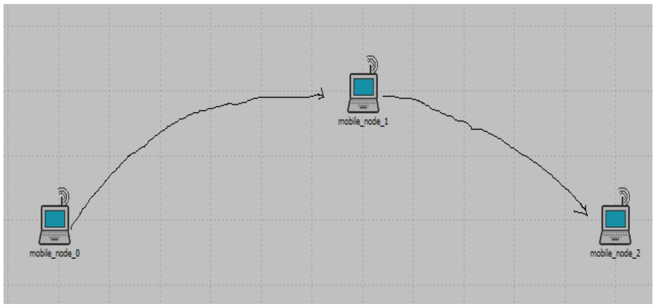


Fig.1. Mobility of the Mobile Ad-Hoc Networks

3.1.2. Simulation environment

In our study, two different network topologies were created and analyzed to evaluate the mobility effect on the node's behavior and the network performance. Different settings have been applied to the two topologies based on the needs of the network simulations.

In the first topology, the relationship between mobility, transmission range and the overall throughput and delay of the network is investigated. As Figure 2 illustrates, one subnet consists of eight nodes around each other. Another single node is approaching this sub network with a constant speed. To study the impact of mobility among the nodes inside the sub network, different scenarios were created to compare the delay and throughput where the nodes are either static or moving randomly. We also changed the speed of the nodes in different steps to see the influence of this factor on the network. To evaluate the effect of the transmission range, we also varied the transmission range in different scenarios according to the distance and the area that were used in the simulation. In addition, the mobility impact on this network with different traffic loads was also studied.

Table 1 shows the setting used for the first topology.

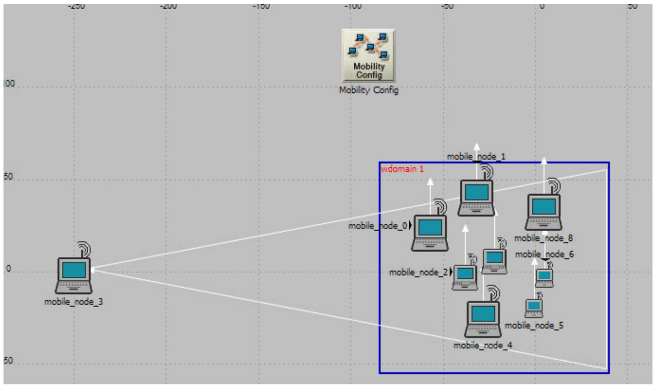


Fig.2. First topology where a node is approaching a static network

Attribute	Value
Transmission power (W)	Varies per scenario
Data Rate (bps)	11 Mbps
Physical Layer Method	Direct Sequence
Buffer Size (bits)	256000
Packet Size (bits)	Exponential (1024)
Traffic generated per node	Varies per scenario
Node's Speed	Varies per scenario
Nodes movement method	Defined/Vector trajectory
Simulation time (min)	60

Table.1. Topology 1 configurations

For the second topology, two similar subnets are created and each consists of 7 nodes. One of the nodes is static and it transmits to the other static node in the second subnet. The other nodes inside the subnet are either static or moving while trying to transmit data to the static station inside their subnet. The two subnets are moving towards each other with a constant speed. The internal nodes are located with different distances from the static receiver in order to study the effect of different movement trajectories on the fairness among nodes. Note that the internal nodes are in the transmission range of each other. That is, each subnet allows their nodes to transmit inside the region of the subnet. The transmission range for the static receiver is higher than the others due to the fact that the receiver will need to transmit to the other static receiver located at the second subnet. In this topology, we not only look into the delay and throughput factors, but also check the fairness among nodes and the effect of RTS/CTS. Table 2 shows the setting we used for the second topology.

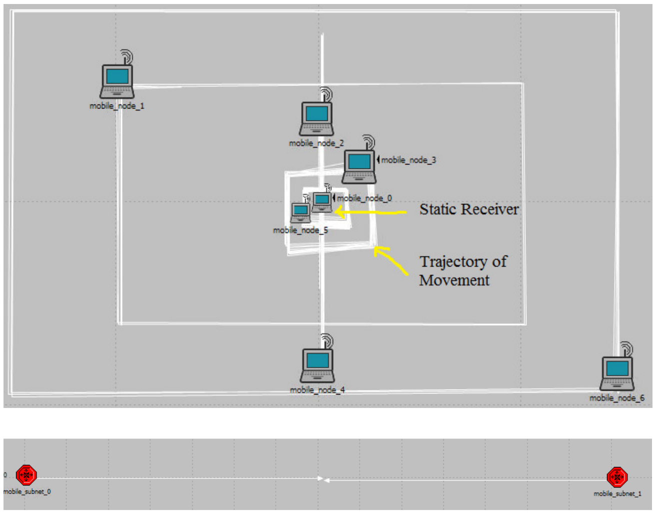


Fig.3. Second topology where the mobile nodes are moving around the static receiver inside the two subnets

Attribute	Value
Transmission power of the static receiver (W)	0.001
Mobile Nodes Transmission power (W)	0.0003
Data Rate (bps)	11 Mbps
Physical Layer Method	Direct Sequence
Buffer Size (bits)	256000
Packet Size (bits)	Varies per scenario
Traffic generated per node	Varies per scenario
Internal Node's Speed (m/s)	0.2
Subnet speed (m/s)	1
Nodes movement method	Defined trajectory
Simulation time (min)	30

Table.2. Topology 2 configurations

3.2. Experimental results

3.2.1. Impact of mobility on delay and throughput

A) The impact of mobility with lower transmission range

To evaluate the impact of mobility with lower transmission range, three scenarios were created under the first topology (figure 2). The transmission range and traffic load for these scenarios are the same, while the mobility inside the subnet is different:

1) Scenario 1: inside nodes are static
2) Scenario 2: inside nodes move with speed 0.2m/s
3) Scenario 3: inside nodes move with speed 1m/s

For all these scenarios, the single node in approaching the sub-network with a constant speed 0.2m/s. Table 3 shows the configuration of transmission power and traffic load of these scenarios. The Domain which covers an area of 120 × 120 square meter allows the nodes to move inside this region. A lower transmission range is defined so that the nodes can sense each other at a maximum of 80 meters distance. That is, the nodes will not be able to sense each other at some parts of the Domain. This will allow us to see the effect of lower transmission region on the networks using the mobility feature.

Attribute	Value
Transmission power (W)	3E-005
Packet Size (bits)	Exponential (1024)
External Node's Speed (m/s)	0.2
Traffic generated per node	Exponential (0.1)

Table 3: Common parameters

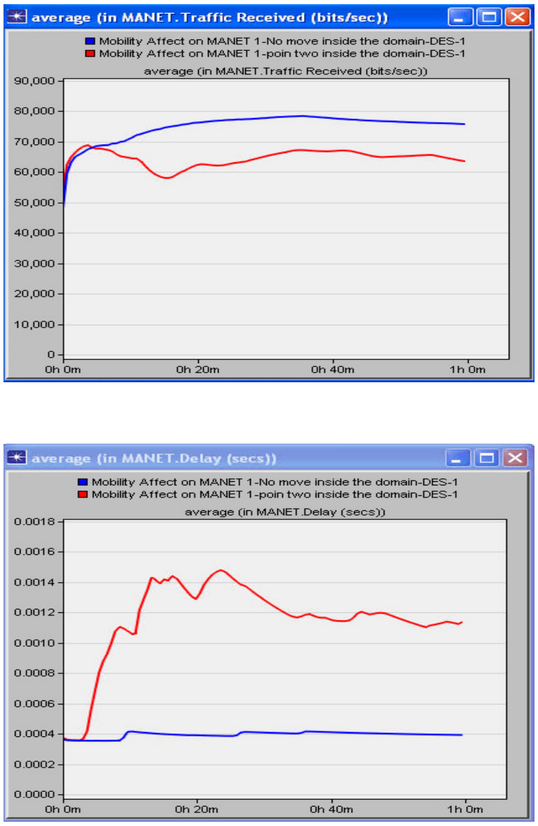


Fig.4. Comparison of average Delay and Throughput for static and moving inside nodes with speed 0.2m/s with lower transmission range

Figure 4 compares the results for scenario 1(static) and 2 (node moving with low speed 0.2/m). Results show that when the nodes are not moving inside the domain, the network has higher throughput and lower delay. This seems be to due to the fact that when the nodes move around, the transmission range decreases and the connection establishment among nodes becomes weaker. Hence, the overall throughput decreases coming up with higher delay.

Figure 5 compares the results for scenarios 2 (low movement speed) and 3 (high movement speed. Results demonstrate a higher delay and lower traffic being received as a result of an increase in the speed of the mobile nodes. Higher speed will make the node to move further from the receiver in a shorter amount of time and therefore, less chance to deliver their data to the destination. Higher delay is due to the fact the nodes are having more problem in delivering their data to the receiver and experiencing a higher back off time and retransmissions of the data. It also contributes to the internal collision or packet loss which prevents the delivery of the data.

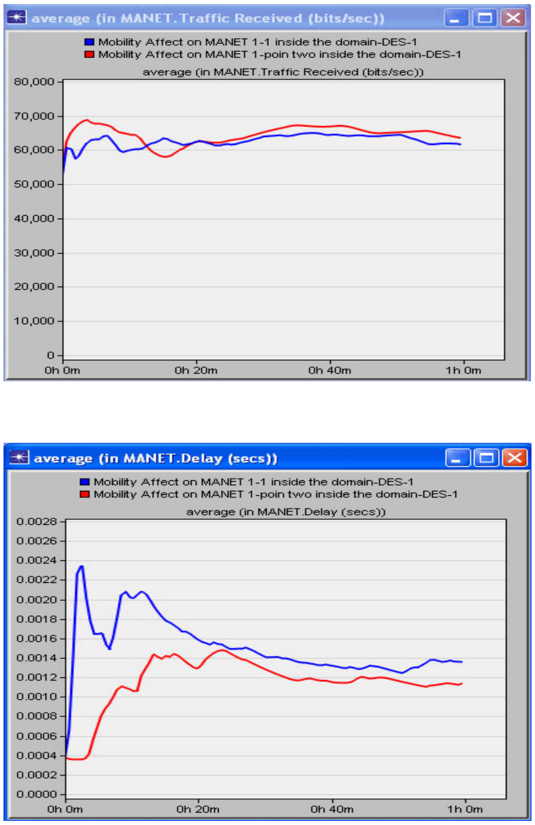


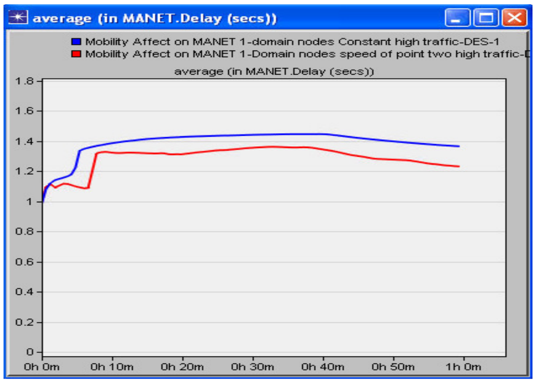
Fig.5. Comparison of average Delay and Throughput for the networks with low movement speed (0.2m/s) and high movement speed (1 m/s) with low transmission range.

B) The impact of mobility with higher transmission range

Attribute	**Value**
Transmission power (W)	0.0001
Packet Size (bits)	Exponential (1024)
Internal Node's Speed (m/s)	1
Traffic generated per node	Exponential (0.005)

Table 4: Common parameters

In this case, we increased the transmission range (0.0001 W) so that the range covers the whole area of the movement. We also increased the Traffic generated by each node to see how the mobile stations behave while generating more traffic. Figure 6 shows the comparison results for the two network scenario with static inside nodes and moving inside notes (speed 0.2m/s). It is interesting to see that in this case, the mobility will have a positive impact which leads to a little higher throughput and lower delay for the entire network.



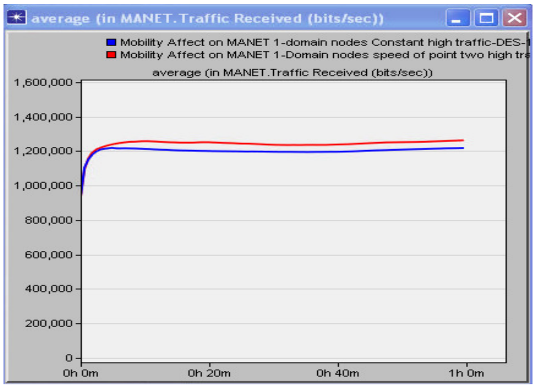


Fig.6. Comparison of average Delay and Throughput for static and moving inside nodes with speed 0.2m/s with higher transmission range

We also found out that increasing the speed of a network with high transmission range will slightly improve the performance of the network in case of throughput and delay. The reason may be due to the fact that all nodes are within the transmission range of each other no matter how they move. Therefore, the random movement pattern of the inside nodes may lead to a more even distribution of the nodes that helps with channel access.

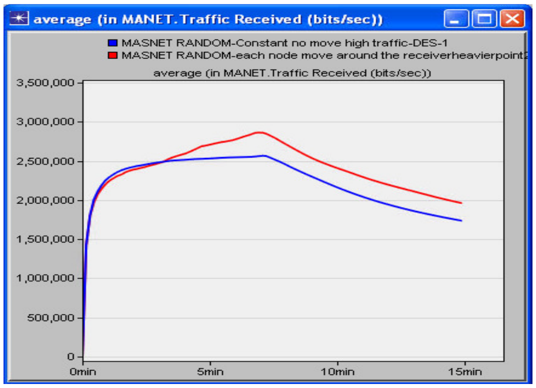
C) *Impact of group mobility on delay and throughput*

Starting from this subsection, we will describe our findings on the impact of *group mobility pattern*. The simulations were created using the second topology as discussed earlier (Figure 3). The two subnets are moving towards each other with a speed of 1 m/s. The internal nodes inside each subnet are moving with the speed of 0.2 m/s. The nodes have different distances to the fixed receiver. In the case where the nodes are moving, they move around the receiver based on their location and distance with regards to the receiver. Besides the delay and throughput factors, the fairness among nodes and the effect of RTS/CST method is investigated in the following section.

Attribute	**Value**
Transmission power (W)	0.0003
Packet Size (bits)	Exponential (8192)
Internal Node's Speed (m/s)	0.2
Traffic generated per node	Exponential (0.0008)

Table 5: Common parameters

In this scenario, we investigated the effect of mobility on a network with stations generating a relatively larger traffic inside the network. We also increased the packet size per station. Similar to the first topology, results demonstrate a better performance of the network with mobility in comparison to the static one when the transmission ranges covers the movement paths.



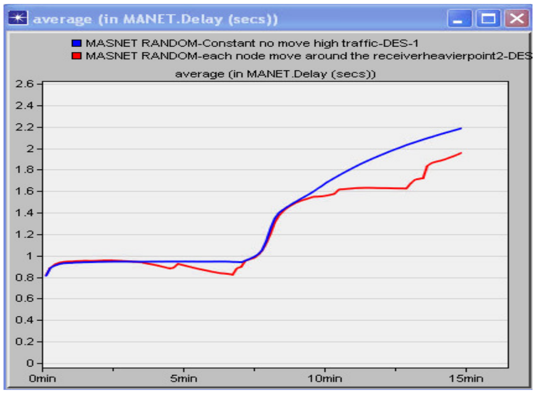


Fig.7. Comparison of average Delay and Throughput for static and moving inside nodes with speed 0.2m/s for group mobility

3.2.2. Impact of mobility on Fairness

A) *The fairness issue without RTS/CTS*

To evaluate the fairness of IEEE 802.11 MAC layer protocol for mobile network, we used the back off slot time which is the number of slots that a stations needs to back off before transmission while contenting for the medium, and the contention window size after the successful transmission of the station as the measurements. . Moreover, the retransmission attempt is a good factor to analyze the delivery efficiency of the packets per node when the stations are moving around the receiver. This factor reflects the impact of both the internal collision and the transmission errors including the loss of acknowledgment or an error occurred in the packet. The effect of the amount of traffic generated by each node on the fairness issues has also been investigated in this part. Node 3 and Node 6 are selected for our results. As mentioned before, the reason for this selection is the distance difference between the nodes and the receiver on their moving paths. Therefore, we can clearly see the effect of the different distances caused by mobility on the fairness among these nodes.

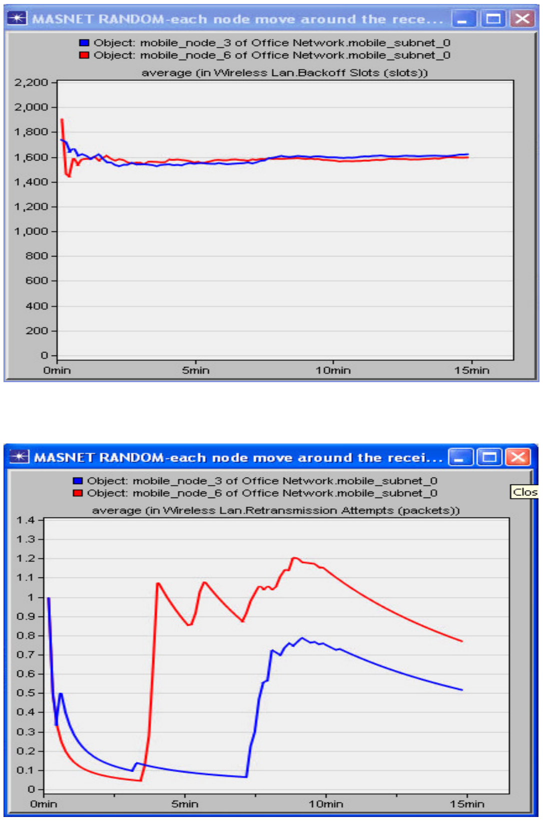


Fig.8. Average Back off slot time and retransmission attempts for the two selected nodes with low traffic load

Figures 8 and 9 present our simulation results for CSMA/CA without RTS/CTS. For the low load generated by each station, the two nodes have almost the same back off slot time (as shown in Figure 8) demonstrating that they have the same chance to access the channel. On the other hand, the retransmission attempts per node are much more less for the closer node to the receiver (Node 3). This might be due to the internal collision or the error inside the packets resulting in the failure of the delivery of the packets.

We repeated the same procedure but changing the amount of traffic generated by each station to a higher level (exponential (0.0008)). We also increase the packet size up to eight times (exponential (8192)).

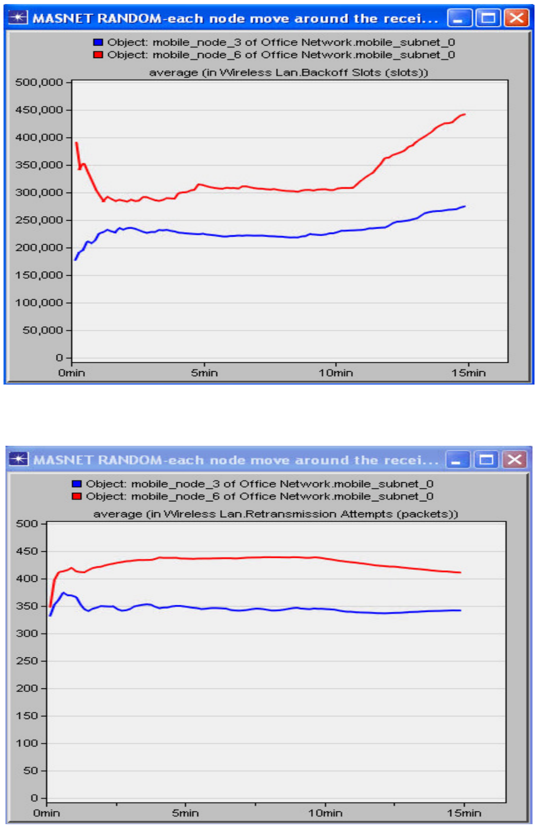


Fig.9. Average Back off slot time and retransmission attempts for the two selected nodes with higher traffic

As shown in Figure 9, the results illustrate that the back off slot time has increased dramatically for both nodes and that the difference becomes obvious as the distance to the receiver increases. This can be due to higher collision and more competition for accessing the channel resulting in higher back off's and retransmission attempts for both nodes. The closer the node is to the receiver, the higher chance it has to access the channel while moving around the receiver.

B) The fairness issue using RTS/CTS

Attribute	**Value**
Transmission power (W)	0.0003
Packet Size (bits)	Exponential (1024)
Internal Node's Speed (m/s)	0.2

Traffic generated per node	Exponential (0.1)
RTS threshold (bytes)	256
Internal Node's Speed (m/s)	1
Subnet speed (m/s)	1

Table 6: Common parameters

In this case, we added the RTS/CTS option to each node to see the efficiency of this method on the network. Low traffic has been used in this scenario. From Figure 10, we can see that the retransmission attempts have been decreased for Node 3, which demonstrated the effectiveness of the handshaking method. Same results occurred for other nodes inside the network showing the good efficiency of the handshaking method.

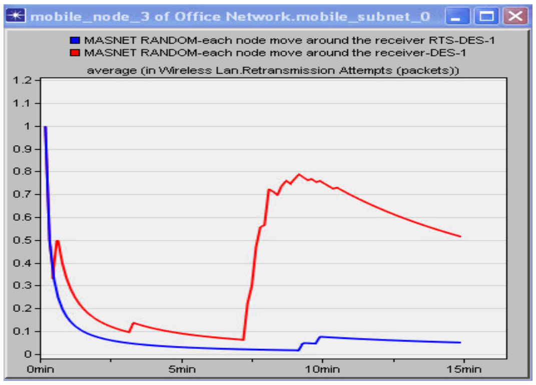


Fig.10. Comparison of the average retransmission attempts for Node 3 using CSMA/CA only (RED); and using CSMA/CA with RTS/CTS (BLUE)

C) *The network performance using RTS/CTS*

We also studied the performance of whole network using the RTS/CTS access mechanism. The same configurations were used as the above simulation. Results depict that the delay of the whole network decreases due to the prevention of the collisions and allowing the nodes to have their data delivered in a shorter amount of time.

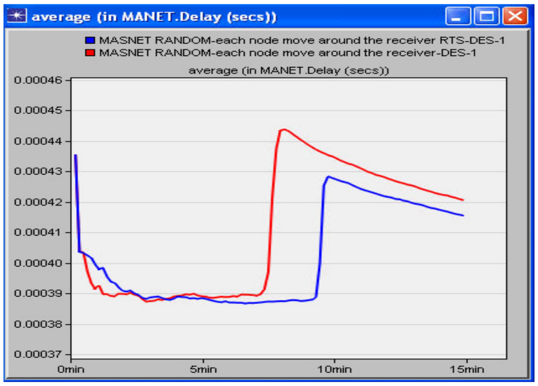


Fig.11. Comparison of average network delay: Red plot—without RTS/CTS; Blue plot-- with RTS/CTS

4. CONCLUSION

In this paper, the performance of the Mobile Ad-hoc networks is investigated using the IEEE 802.11 MAC protocol. We have shown that Mobility can affect the network based on different factors. We studied the effect of varying the speed of the nodes, and their location on the network. We have also studied the fairness and the effect of the RTS/CTS handshaking process on the performance of the nodes inside the network. Our results show that the performance of the network varies as the mobile nodes move inside the network. We illustrated that the performance of the network improves as the traffic increases when a sufficient transmission ranges of nodes is provided. We have also shown that Mobility will cause the nodes to have longer back off times and retransmission attempts in order to deliver their information to the destination. The RTS/CTS handshaking method demonstrated its efficiency on the mobile nodes when the number of collisions becomes more and more.

5. REFERENCES

[1] K. Xu, M. Gerla, and S. Bae, "Effectiveness of RTS/CTS handshake in IEEE 802.11 based ad hoc networks," Ad Hoc Networks, Elsevier, vol. 1, no. 1, pp. 107-123, 2003.

[2] T. S. Ho and K. C. Chen, "Performance Analysis of IEEE 802.11 CSMA/CA Medium AccessControl Protocol," in Proc. IEEE PIMRC '96, pp. 407-411, 1996.

[3] Khurana, S.; Kahol, A.; Jayasumana, A.P, Effect of Hidden Terminals on the Performance of IEEE 802.11 MAC Protocol," LCN '98 proceedings, pp 12-20, 1998.

[4] Samir R. Das , Robert Castañeda, Jiangtao Yan , Rimli Sengupta, Comparative performance evaluation of routing protocols for mobile, ad hoc networks. In 7th Int. Conf. on Computer Communications and Networks (IC3N), pages 153–161, October 1998.

[5] Das, S.R, Perkins C.E, Royer, E.M., Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks in IEEE Proceedings, pp 3 – 12, 2000.

[6] Jagadeesan S, Manoj BS, Murthy CSR. Interleaved carrier sense multiple access: an efficient MAC protocol for ad hoc wireless networks. Proceedings of IEEE ICC'03, May 2003.

[7] Online Documentation, "OPNET Modeler," http://www.opnet.com/, Date visited: January 2011.

[8] IEEE 802.11 Working Group, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, 1997.

[9] H. Jasani, N. Alaraje, Evaluating the Performance of IEEE 802.11 Network using RTS/CTS Mechanism," in IEEE EIT 2007 Proceedings.

[10] P. Karn, MACA—a new channel access method for packet radio, in: Proceedings of the ARRL/CRRL Amateur Radio 9th Computer Networking Conference September 22, 1990.

[11] V. Bhargavan, A. Demers, S. Shenker, L. Zhang, MACAW—A Media Access protocol for wireless Lans, in: Proceedings of the ACM SIGCOMM, 1994, pp. 212–225

[12] Online Documentation, "The Network Simulator - ns-2," http://www.isi.edu/nsnam/ns/, Date visited: January 2011.

Comparison of DSR, AODV and DSDV Routing Protocols in Fully and Partially Connected VANET

Sayed Mohammad Mehdi Feiz[1] and Ali Movaghar[2]
[1] Dept. of Computer Engineering, Sharif University of Technology - Kish Campus, Kish Island, I.R.IRAN
[2] Department of Computer Engineering, Sharif University of Technology, Tehran, I.R.IRAN

Abstract - *Vehicular Ad hoc Networks (VANETs) are special case of Mobile Ad hoc Networks (MANETs). The design of routing protocols in VANETs is an important and necessary issue. In this paper, DSDV, DSR and AODV routing protocols are simulated when the connection sessions are increased. Finding indicates utilization of DSR protocol is more applicable for environment that number of connection sessions is high. The simulation results of DSDV are not desirable in comparison with other two reactive routing protocols.*

Keywords: DSR; AODV; DSDV; MANETs; VANETs; routing protocols

1 Introduction

Vehicular Ad hoc Networks are special case of Mobile Ad hoc Networks. According to this definition, some characteristics of MANETs can be used for VANETs. In many ways VANETs are also similar to MANETs. As an illustration, both networks are multi-hop mobile networks having dynamic topology. There is no central entity, and nodes themselves route data across the network. Both MANETs and VANETs are rapidly deployable without need of any infrastructure.

MANETs and VANETs, both are mobile networks. However, the mobility pattern of VANET nodes is restricted to move on specific paths such as roads and hence not in random direction. This gives VANETs some advantage over MANETs as the mobility pattern of VANET nodes is predictable. MANETs are often characterized by limited storage capacity, low battery and processing power. VANETs, on the other hand, do not have such limitations. Sufficient storage capacity and high processing power can be easily made available in vehicles. Moreover, vehicles also have enough battery power to allow for long range communication. Another difference is highly dynamic topology of VANETs because vehicles may move at high velocities. This makes the lifetime of communication links between nodes quite short. Node density is also unpredictable; during rush hours the roads are crowded with vehicles, whereas at other times lesser vehicles are there.

VANETs consist of two basic components: vehicles and infrastructures. According to this, communication between components divides into Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I). Different types of vehicular communication cause diverse type of vehicular applications. The applications in VANET consist of safety-oriented application, traffic management applications, traffic coordination and traffic assistance, traveler information support, and comfort applications [1,13]. Public safety applications are very important application in VANET. The main purpose of safety applications is to avoid the accident of vehicles in the road and survive drivers' life. In safety applications, vehicle create warning message and send it to other drivers for informing the dangerous event in the road. The safety applications divide into passive safety, active safety as well as proactive safety and warning applications. Traffic management applications are focused on improving traffic flow, thus reducing both congestion as well as accidents resulting from congestion, and reducing travel time. Traffic management applications include traffic monitoring, traffic light scheduling, and emergency vehicles. Traffic coordination and traffic assistance have been the main research topics of many Inter Vehicle Communication (IVC) projects. Clearly these applications require close-range IVC with tight real-time constraints and can be implemented with either Sparse Roadside to Vehicle Communication (SRVC) or Ubiquitous Roadside to Vehicle Communication (URVC) system. Traveler information support applications provide updated local information, maps, and in general messages of relevance limited in space and time. These messages mainly focus on the local information and road warnings. Local information including local updated maps, the location of gas stations, parking areas, and schedules of local museums can be downloaded from selected infrastructure places or from other local vehicles. Road warnings comprise of many types such as ice, oil, or water on the road, low bridges, or bumps. In fact, traveler information support applications are the mixture of V2V and V2I communication. The last type of vehicular application is comfort applications. The main focus of comfort applications is to make travel more pleasant. This class of applications may be motivated by the desire of passengers to communicate with either other vehicles or ground-based destinations such as Internet hosts or the public service telephone network. Actually comfort application is an infotainment for passenger [2].

Nodes in VANETs are vehicles and infrastructures. Nodes can communicate with each other at any time and without any restriction, except for connectivity limitations and subject to security provisions. VANETs have two kinds of routing: unicast and multicast routing. If two nodes are adjacent, unicast routing is used. Under other circumstances, if there is no direct link between the source and the destination, multicast routing is used. Multicast routing approaches have further divided into six categories: flooding, tree-based, mesh based, overlay-based, backbone-based and stateless [14]. This paper review some flooding (broadcast) based routing protocols. Researchers attempt to adapt MANETs routing in VANETs environment because VANET is similar to MANET in many aspects. Routing protocols in mobile ad hoc network are mainly classified into topology-based and position-based approaches.

Topology-based routing protocols, which can be further divided into proactive, reactive and hybrid approaches, use the information about the links that exists in the network to perform packet forwarding. Proactive routing protocols utilize some traditional routing strategies such as DSDV [3], OLSR [4], and TBRPF [5]. They maintain and update information on routing between all nodes in a given network at all times. Route updates are periodically performed regardless of network load, bandwidth constraints, and network size. The main drawback of these protocols is that the maintenance of unused paths may occupy a significant part of the available bandwidth if the topology of the network changes frequently [6]. Reactive routing protocols, including AODV [7], DSR [8], and TORA [9], maintain only the routes that are currently in use, thereby reducing the burden on the network when only a small subset of all available routes is in use at any time. Hybrid routing protocols combine local proactive routing and global reactive routing strategy in order to achieve a higher level of efficiency and scalability. The salient example of hybrid routing protocols is ZRP [10].

Position-based routing protocols require additional information about the geographical position of the participating nodes. Each node determines its own position through the use of GPS or other type of positioning services. These type of routing protocols has two parts; location service which is used by the sender of a packet to determine the position of the destination and to include it in the packet's destination address. The second part is forwarding strategy sending packet to one or more one hop neighbors. Position-based routing protocols are suitable for networks that are highly dynamic and their topology may change frequently. Position-based routing algorithms do not require establishment or maintenance of routes, which means that nodes have neither to store routing tables nor to transmit messages to keep routing tables up to date. The prominent examples of position-based routing are LAR [11] and GPSR [12].

Feiz and Movaghar simulated AODV, DSR and DSDV routing protocols based on the effects of mobile speed and number of connection [2]. The objective of this paper is to extend the simulation of them by increasing the number of connection sessions. The simulation performance parameters are similar to the original paper.

The rest of this paper is organized as follows. In section 2, we describe the simulation environment and performance parameters. Section 3 presents the simulation results of routing protocols based on four performance parameters. Finally, we conclude our study in section 4.

2 Simulation environment

The simulation environment is described as follows.The simulator which is used a network simulator version 2 (ns2). Simulation area of this study is a partial part of the road. The pause time of the nodes is zero to stress the mobility of nodes.

The routing protocols are compared according to four metrics: packet delivery ratio, normalized routing load (NRL), packet loss ratio and end to end delay.

TABLE 1 Simulation environment for AODV, DSDV and DSR

Simulator	NS2
Simulation Area	500m × 200m
Number of Nodes	50
Pause Time	0.0 s
Maximum Speed of Node	30 m/s
Simulation Time	200 s
Maximum Connections	10, 20, 30, 40, 50
Traffic Type	CBR
Rate	4 packets/s

3 Simulation results

We performed analysis with respect to the number of nodes connected together. We ran our simulation with speed fixed at 30 m/s, pause time at zero, and the number of nodes at 50. We varied the number of connection sessions between 10 and 50.

The simulation results are shown in Fig. 1 through Fig. 4. Fig. 1 illustrates the packet delivery ratio of three routing protocols when the number of connection sessions is increased. Packet delivery ratio is the ratio of the data packets delivered to the destinations and the data packets generated by the CBR source. AODV and DSR have similar results until the number of connection sessions is 30. In 40, the packet delivery ratio of two protocols decreases steeply. The decline of AODV is more than DSR. It is due to use of different approaches for maintaining route [2]. The packet delivery ratio of AODV and DSR is increased again when the number of connection sessions is reached to 50. DSDV has lower packet delivery ratio in comparison with AODV and DSR. However, the packet delivery ratio of DSDV is higher than other two routing protocols in 40. It is due to the nature of proactive routing protocols. When the number of connection sessions is increased, routing table has various alternatives for selecting the route towards destination.

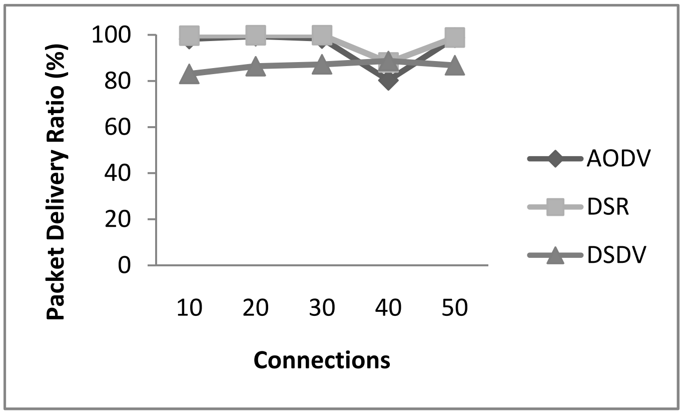


Figure 1 Effect of connection sessions on ratio of packet delivered

Fig. 2 depicts the normalized routing load when the connection sessions are increased. Normalized routing load (NRL) is the ratio of the total number of routing packets transmitted or forwarded at the network layer to the total number of CBR packets received at the destination at the application layer. DSR has the best result in comparison with other routing protocols. For example, the normalized routing load of DSR falls into zero when the number of connection sessions reaches to 50. DSDV has downward trend in normalized routing load when the number of connection sessions is increased. The result of AODV is between DSR and DSDV. The NRL of AODV and DSR, reactive routing protocols, reach at a peak when the connection sessions are 40. Consequently, DSR has lower NRL than other routing protocols.

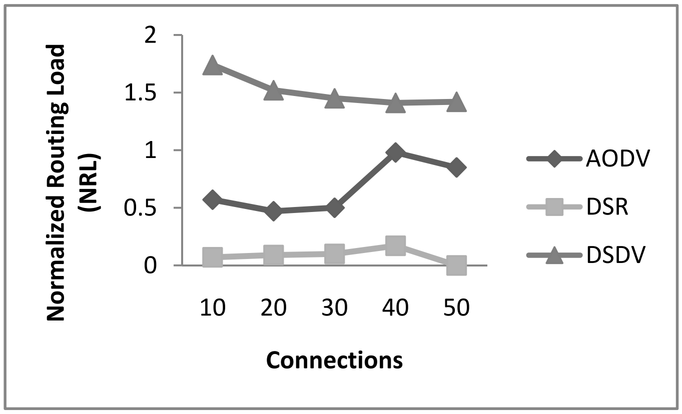


Figure 2 Effect of connections on normalized routing load

Fig. 3 shows the packet loss ratio of three routing protocols when the number of connection sessions is increased. Packet loss ratio is the ratio of packets loses due to reject or drop in the network. Packet loss can be caused by a number of factors including signal degradation over the network medium, oversaturated network links, corrupted packets rejected in transit, faulty networking hardware, faulty network drivers or normal routing routines. DSDV has undesired results in comparison with AODV and DSR. The results show that new proactive routing protocols can be suitable for environments which the connections are high. The discrepancy in the results of reactive routing protocols is use of various approaches for maintaining the routes. Consequently, DSR is a better choice for scenario which the connection sessions of nodes are high.

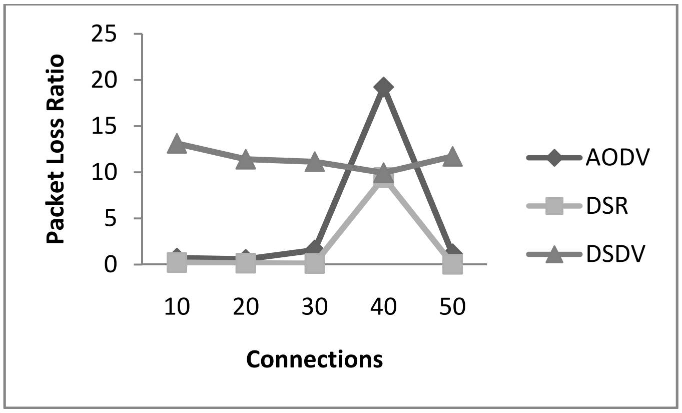


Figure 3 Effect of connections on packet loss ratio

Fig. 4 depicts the end to end delay of routing protocols when the connection sessions are increased. The diagrams of delay of three routing protocols are close together until the number of connections reach to 20. Since then, the delay of AODV and DSR is increased and reached a peak in 40. The delay of DSDV, proactive routing protocol, is gradually. Consequently, the delay of proactive routing protocols is lower than reactive routing protocols when the connection sessions are increased.

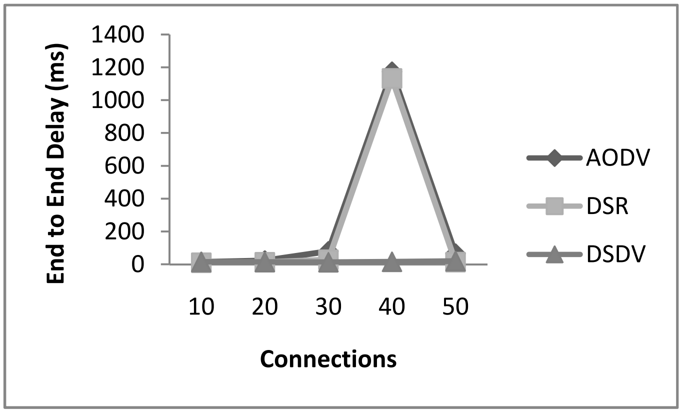


Figure 4 Effect of connections on end to end delay

4 Conclusion

This paper compares reactive and proactive routing protocols in the environment which the connection sessions are so high. DSR and AODV, reactive routing protocols, have some similarity such as these are an on demand protocols. The main deference of these routing protocols is two aspects. First of all, DSR is a source routing protocol but AODV is a distance vector routing protocol. The second difference is the approach of storing routes. AODV uses routing table but DSR uses cache. The discrepancy of results between AODV and DSR is originated from different approach of storing route. DSR has had desirable results in comparison with AODV. DSDV, traditional proactive routing protocol, shows some optimization when the number of connection sessions is increased. New proactive routing protocols may be had favorable results for VANET with number of connection is so high. The variation of results in DSDV is very low in four performance parameters. Sudden increase of performance parameters in reactive protocols is due to the time for achieving to the stable state in route maintaining approaches. Afterwards, the performance parameters are decreased when the maintaining approach reaches to the stable state. The simulation results show that MANETs routing protocols are not completely satisfied specifications of VANETs. The design of new routing protocols is an emerging issue for VANETs.

5 References

[1] C-M. Huang and Y-C. Chang, Telematics Communication Technologies and Vehicular Networks: Wireless Architectures and Applications,1st ed., Hershey, NY: Information Science Publishing, 2010, pp. 229–251.

[2] S. M. M. Feiz and A. Movaghar, "Characteristics of secure routing in vehicular ad hoc network (VANET)," ICWN 2010, pp. 128-134, 2010.

[3] C.E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in Proc. of ACM SIGCOMM '94 Symposium on Communication, Architectures and Protocols (1994), pp. 234–244, Aug 1994.

[4] T. H. Clausen and P. Jacquet, "Optimized link state routing (OLSR)," Network Working Group, RFC 3626, Oct 2003.

[5] R. G. Ogier, F. L. Templin, and M. G. Lewis. "Topology dissemination based on reverse path forwarding (TBRPF)," Network Working Group, RFC 3684, Feb 2004.

[6] S. R. Das, R. Castaneda, and J. Yan, "Simulation based performance evaluation of mobile ad hoc network routing protocols," ACM/Baltzer Mobile Networks and Applications (MONET) Journal, vol. 5, Issue 3, pp.179-189, Sep 2000.

[7] C. E. Perkins, E. M. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing," Network Working Group, RFC 3561, July 2003.

[8] D. B. Johnson, D. A. Maltz, and Y-C. Hu, "The dynamic source routing (DSR) protocol for mobile ad hoc network," IETF MANET Working Group, Internet Draft, July 2004.

[9] V.D. Park and S. Corson, "Temporally-ordered routing algorithm (TORA)," version 1 functional specification (Internet-draft), Mobile Ad-hoc Network (MANET) Working Group, IETF 1998.

[10] Z.J. Haas and M.R. Pearlman, "The zone routing protocol (ZRP) for ad hoc networks," Mobile Ad-hoc Network (MANET) Working Group, IETF 1998.

[11] Y. B. Ko and N. H. Vaidya, "Location aided routing (LAR) in mobile ad hoc network," wireless networks, vol. 6, pp. 307–321, 2000.

[12] B. Karp and H. T. Kung, "GPSR: greedy perimeter stateless routing for wireless networks," International Conference on Mobile Computing and Networking (MobiCom), pp. 243–254, 2000.

[13] M. Emmelmann, B. Bochow, and C. C. Kellum, Vehicular Networking : Automotive Applications and Beyond, 1st ed., Chichester, West Sussex, UK: Jhon Wiley & sons Ltd, 2010, pp. 1-28.

[14] C. D. Cordeiro, H. Gossain, and D. Agrawal, "Multicast over Wireless Ad hoc Networks: Present and Future Directions," Network, IEEE, vol.17, pp. 52-59, 2003.

Improvement of Medium Access Control Protocol to Reduce Latency in Ubiquitous Wireless Sensor Networks

Myung-Sub Lee[1]**, Chang-Hyeon Park**[2]
[1]Div. Of Computer Technology, Yeungnam College of Science & Technology, Daegu, Korea
[2]Dept. Of Computer Engineering, Yeungnam Univ., Gyeongsan, Gyeongbuk, Korea

Abstract – *In several ubiquitous sensor networks, sensor nodes often face contention for medium access and service. Unfortunately, traditional carrier sense multiple access with collision avoidance (CSMA/CA) protocols such as the IEEE 802.11 Distributed Coordination Function (DCF) do not handle such constraints adequately, and this weakness finally leads to an increase in the latency and degradation of the throughput as the scale of a network is increased. Therefore, we present a more efficient method for medium access in a real-time ubiquitous sensor network. The proposed MAC protocol is similar to conventional CSMA/CA protocols, except that it does not use a time-varying contention window from which a node randomly selects a transmission slot. To reduce the latency for the delivery of event data from sensor nodes, a fixed-size contention window with a non-uniform probability distribution of transmitting in each slot is selected. Through a simulation using ns-2, a widely used network simulation package, it is shown that the proposed method can reduce the latency considerably as compared to conventional IEEE 802.11 MAC protocols for a sensor network with up to 256 nodes. It is also shown that the proposed MAC scheme realizes a latency similar to that realized by a decentralized CSMA-based MAC protocol for real-time ubiquitous sensor networks that are sensitive to latency.*

Keywords: Sensor Network, Contention Window, CSMA/CA, DCF, 802.11, MAC

1 Introduction

In wireless networks, multiple wireless nodes share a channel to transmit data; however, this can lead to contention for channel access. Therefore, the medium access control (MAC) protocol, which is commonly used in networks to provide channel access control mechanisms, is essential for arbitrating this problem. This paper thoroughly analyzes the operating mechanism and characteristics of ubiquitous sensor networks from the viewpoint of developing a more suitable MAC protocol. One of the most important issues in sensor networks is improving the battery life; in fact, several researches have focused on this issue over the last few years [1-4].

The MAC protocol for ubiquitous sensor networks differs from conventional MAC protocols in the following respects.

● *Most sensor nodes are event-driven, and they contend with other nodes to share a transmission channel.*

● *Because all sensor nodes do not simultaneously respond to an event, the number of responding nodes in the sensor network is time-variant.*

● *Frequent occurrences of data transmission delays increase the energy consumption of a sensor node and simultaneously reduce the overall performance of the sensor network.*

Therefore, this research focuses on designing a method that can effectively decrease the data transmission delay in a sensor network to the greatest extent possible. This is realized by developing a new MAC protocol by taking into account the three issues listed above. It is demonstrated that with a suitably modeled and simulated sensor network, the proposed protocol reduces the energy consumption of wireless sensors and effectively improves the performance of real-time data transmission.

When several nodes (N) that are requested to transmit sensing data by the base station are concurrently competing to occupy the transmission channel, the nodes attempt to transmit the requested data (R) to the destination node by avoiding collisions to the greatest extent possible. Whereas conventional MAC protocols solve this problem by optimizing the throughput for the case in which all the nodes transmit the report data ($R = N$), the MAC protocol proposed in this paper realizes optimized performance of the ubiquitous sensor networks for the case in which only some nodes transmit the report data ($R < N$). Each node in the conventional carrier sense multiple access (CSMA) protocol chooses a slot randomly based on the probability of the uniform distribution in the current contention window (CW) in order to perform the back-off process, which decreases the value of the chosen slot every constant time; the node finally transmits data when this value becomes zero. This method can effectively prevent collisions; however, if the number of nodes is very high, collisions start to occur between nodes. In such a situation, the network cannot function normally because unavoidable transmission delays are caused.

In ubiquitous sensor networks, the amount of initially generated report data is generally quite less, and therefore, the initial slots selected at this time are in a low-contention state. As a result, hardly any collisions occur between slots, and the network latency is low. Considering the abovementioned factors, the MAC protocol proposed in this research differs from conventional ones in the following respects. The contention window size is regularly optimized so as to minimize the latency. In addition, the geometric probability distribution is specifically designed to replace the uniform probability distribution that is conventionally used in order to differentiate the selection probability during the process of selecting the transmission slot.

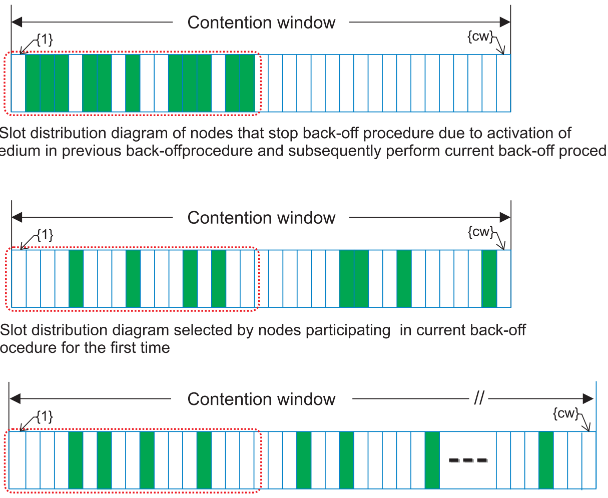
2 Modified MAC protocol

Whenever a collision occurs between nodes in CSMA-based protocols such as 802.11 [5-6], BMAC [7], SMAC [8-9], and MACAW [10], the binary exponential back-off (BEB) method is used to increase the size of the contention window of the colliding nodes twice. In other words, this method increases the size of the contention window of colliding nodes to a size that is sufficiently adapted to the value of the current active nodes in order to minimize collisions between the nodes. However, this method involves the following problems. First, if the number of active nodes (N) that are ready to transmit sensed data increases, say, in a situation in which several sensors simultaneously wait for an idle period, considerable time is required to increase the contention window to a size that is sufficient to accept all of them. Second, if the size of the contention window is increased more than necessary due to previous traffic collisions despite there being only a few active nodes, the bandwidth used in succeeding back-off procedures is dissipated unnecessarily. Therefore, most CSMA protocols can be considered to be inefficient because they focus only on enabling all active nodes to transfer the data by avoiding collisions.

Therefore, this paper proposes a more efficient dynamic collision avoidance scheme with a contention slot selection system that sets the size of the contention window to the minimum (32 slots) in order to avoid the problems caused in conventional protocols by a collision between nodes; this scheme is based on a non-uniform probability distribution so that it can actively respond to the number of active nodes. In the proposed protocol, the method for selecting a transmission slot differs most significantly from that in conventional CSMA-based wireless MAC protocols in that it minimizes the overlapped selection rate of a slot by fundamentally selecting a slot according to the differentiated probability.

2.1 Problem involved in conventional method to select a contention slot

All slots in a contention window are selected with the same probability in the conventional 802.11 MAC protocol. Therefore, in a competing state in which several sensor nodes attempt to access the medium, nodes that are newly participating in the current contention cycle or nodes that are participating again due to a collision in the previous cycle select a slot with the same probability in a contention window having a changed size. Such a slot selection mechanism gradually lowers the selection probability of an empty slot with time despite the change in the size of the contention window.



①Slot distribution diagram of nodes that stop back-off procedure due to activation of medium in previous back-offprocedure and subsequently perform current back-off procedure

②Slot distribution diagram selected by nodes participating in current back-off procedure for the first time

③Slot distribution diagram selected by nodes participating again in current back-off procedure due to collision in previous back-off procedure

Fig. 1 Variation of slot selection probability in general BEB method

If the back-off procedure is assumed as shown in Fig. 1 first, the slot of the smallest number ($WinSlot_{i-1}$) is assigned a channel to successfully perform data transmission. The remaining slots after this slot are forced to participate again in the next back-off procedure because they do not complete the back-off procedure; in this case, these slots naturally move to the first half of the contention window. However, because the nodes that are newly participating in the back-off procedure and those that are performing the back-off procedure again due to collision in the previous contention both enter this part of the contention window, the contention inevitably increases. Therefore, if the conventional random scheme that selects a transmission slot among several competing slots according to a uniform probability distribution is used, the probability of collision between nodes can increase considerably.

As a result, if a ubiquitous sensor network is intermittently in a high-load state, the overlapped selection rate of competing slots will generally increase due to the limitation of the size of the contention window and the uniform slot selection probability. In particular, the phenomenon in which competing slots are concentrated in the first half of the contention window will be caused as the back-off procedure is proceeding; as a result, collisions will occur frequently between nodes, and this is the most significant cause of the repeated increase in the size of the contention window. Because of such a structural problem, the back-off time of sensor nodes competing to access the medium generally increases, and the transmission delay time and power consumption rate of each sensor rapidly increases; this results in the overall performance of the sensor network deteriorating rapidly.

2.2 Modification of method to select contention slot

In order to solve the abovementioned problem, this paper proposes a method that differentiates the probability distribution for selecting a slot. The fundamental concept for solving this problem is that not only is the number of collisions reduced to the greatest extent possible but also the probability of selecting a slot in order to minimize the back-off time is relatively increased depending on the location in the contention window.

For this purpose, we first consider a method that finds an empty slot that other sensor nodes do not select in a state in which the size of the contention window is fixed. As shown in Fig. 2, if an arbitrary slot succeeds in transmitting in the contention window used for back-off, slots in the preceding locations should never be selected. In other words, the probability distribution function must be designed such that only those slots that are located after the slot that has successfully transmitted can be intentionally selected. The probability distribution function of such a property can be derived by multiplying the probability with which preceding slots cannot be selected by that with which succeeding slots can be selected based on an arbitrary slot for all slots, as shown in Fig. 2.

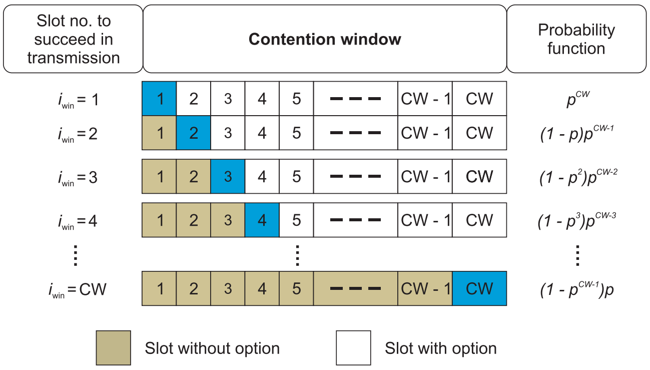


Fig. 2 Probability of selecting a slot to minimize contention between nodes

Investigating the designed probability distribution function in greater detail indicates that the best performance can be obtained if the $\{1\}th$ slot is selected when there is no contention in the current contention window (first stream in figure, $i_{win} = 1$). It may be preferable to select the $\{2\}th$ slot if the $\{1\}th$ slot has already been selected by another node. The fact that the $\{i\}th$ slot is selected in such a slot selection scheme implies that all the slots located before it have already been selected by other nodes. Therefore, the $\{i+1\}th$ slot is selected in order to transmit data without a collision for the minimum delay time.

In order to maintain such an optimum selection method, if the probability distribution function is derived using the probability with which preceding slots are not selected and that with which succeeding slots are selected based on an arbitrary slot, it could be said that the probability $f(i)$ with which each sensor node selects the $\{i\}th$ slot within the range of the contention window $(i \in [1, CW])$ follows a geometric distribution with a parameter *p*, and therefore, the probability mass function can be given by the following equation.

$$f(i) = \begin{cases} (1 - p^{i-1})p^{CW-i+1}, & i = 1, \cdots, CW \\ 0, & otherwise \end{cases} \qquad (1)$$

As a result, the proposed probability mass function $f(i)$ represents a form that increases geometrically with i in the range of $0 < p < 1$, as shown in Fig. 3, implying that the probability with which a certain slot is selected by nodes is relatively higher in the first half of the contention window than in the second half.

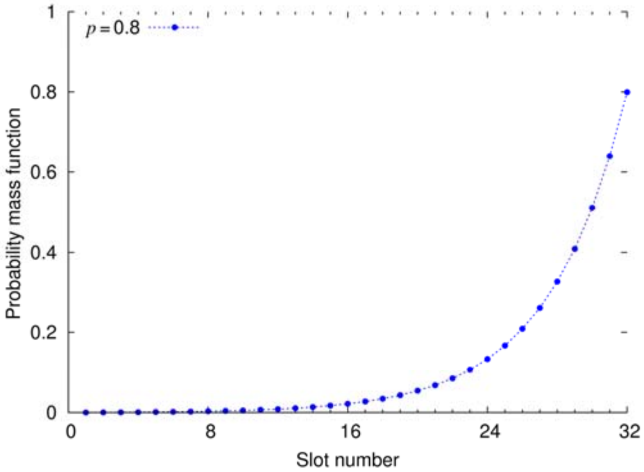


Fig. 3 Probability distribution depending on slot number in contention window

Here, p is a distribution parameter that indirectly represents the probability that the slot is empty; this parameter is determined by the number of active nodes (N).

Here, calculate the probability S_i with which the $\{i\}th$ slot is selected through a virtual decision process. Suppose that the number of active nodes at the first stage is N_1 and a node among these selects the $\{1\}th$ slot. If no node selects the $\{1\}th$ slot, the second stage reduces N_1 to N_2 and selects the $\{2\}th$ slot. If the $\{2\}th$ slot is not continually selected, the third stage decreases N_2 to N_2; in this manner, this process is repeated for each stage. If all slots $[1, \dots, i-1]$ are not empty in the entire process, it can be assumed that N is finally decreased to N_i.

In summary, the selection probability at the $\{i\}th$ slot is the highest when $N_i = N_1$, and N_i has the property that it is constantly reduced from N_1 to 1 as the stages proceed. Therefore, when N_i is relatively large and S_i is small, the probability with which a sensor only selects the $\{i\}th$ slot in the condition in which all the slots prior to this slot are not empty can be given as follows.

$$N_i S_i (1 - S_i)^{N_i - 1} \cong N_i S_i e^{-N_i S_i} \qquad (2)$$

If $N_i S_i$ is assumed to be constant in the above equation, the probability with which the $\{i\}th$ slot is selected ($N_i S_i e^{-N_i S_i}$) gradually decreases as we proceed toward the last slot. In other words, in order to efficiently deal with several nodes (N) that are competing to access the medium using a small contention window that has a fixed size, it is necessary to select a scheme in which the number of nodes accessing the medium reduces at a constant rate (Δ).

$$\frac{N_{i+1}}{N_i} = \Delta \quad (0 < \Delta < 1) \qquad (3)$$

The following condition is satisfied if $p = \Delta$ in equations (1) and (3).

$$N_i S_i \cong N_{i+1} S_{i+1} \qquad (4)$$

Here, if it is considered that each sensor independently selects a slot and the selected slots compete continuously until the sensor that selects the slot with the smallest number succeeds in transmitting through the back-off process, the probability of selecting a slot for each sensor can be derived from equation (1) as follows.

$$\frac{S_i}{S_{i+1}} = \frac{(1 - p^{i-1}) p^{CW-i}}{(1 - p^i) p^{CW-i}} \cdot p \approx p \qquad (5)$$

This could be developed for the slots before $\{CW\}th$ ($[i = 1, \dots, CW - 2]$) by the following equation.

$$\frac{S_1}{S_2} \cdot \frac{S_2}{S_3} \cdot \dots \cdot \frac{S_{CW-1}}{S_{CW}} = p^{CW-1} \qquad (6)$$

The result given below can finally be obtained by applying this equation to equation (4) and perform the following procedure.

$$\frac{N_1}{N_2} \cdot \frac{N_2}{N_3} \cdot \dots \cdot \frac{N_{CW}}{N_{CW-1}} = p^{CW-1}$$

$$\therefore \frac{N_{CW}}{N_1} = p^{CW-1} \qquad (7)$$

As mentioned above, if it is established that $N_{CW-1} = 1$ for the $\{CW\}th$ slot to be selected by only an active sensor, then $\frac{1}{N_1} = p^{CW-1}$. This is finally given by the equation below.

$$\therefore \; p = N_1^{\frac{-1}{CW-1}} \qquad (0 < p < 1) \qquad (8)$$

Then, if it is assumed that $N = N_1$ using equation (8), the optimum probability of selecting a slot, p, can be calculated. In other words, if a medium is accessed by using a contention window having 32 slots in a network consisting of 256 sensor nodes, the value of p is determined to be approximately 0.8 ($p = 256^{\frac{-1}{32-1}} \approx 0.08362$).

2.3 Comparison of properties of modified MAC protocol

The MAC layer of TinyOS, B-MAC [7], is a type of CSMA protocol that uniformly selects fixed-window-based contention slots. The MAC protocol for the sensor network is designed based on a contention window of fixed size because relatively good performance is maintained in an actual environment despite the simplicity of the design. On the other hand, there is a disadvantage in that the scalability to a sensor network that is intermittently in a high-load state is low.

MACAW [10], a MAC protocol for wireless LANs, exploits the BEB method but does not share channel state information. This protocol restarts contention for the next transmission because the size of the contention window is initialized to the minimum value if one of the nodes succeeds in transmitting. However, an overhead is incurred when the medium is accessed because the sensor nodes competing to access the medium are concentrated in a certain interval and the size of the contention window changes considerably. MACAW solves this problem using a learning method that does not newly reset the size of the contention window and instead decreases the size of the contention window used in the previous contention as a size for the next contention if a packet is successfully transmitted (MILD: multiplicative increase, linear decrease).

The 802.11 specification solves the fairness problem of service using a memory technique. A sensor node participates in a contention, where one of the slots in the contention window is randomly selected with uniform probability, and the value of the selected slot is set in a countdown timer. The countdown is stopped when the medium is busy, and it is continued when the medium is idle. If the value of the countdown timer becomes zero (i.e., the timer expires), the corresponding sensor node starts to transmit. When the transmission is completed, the size of the contention window is initialized to the decided minimum value. As a result, the bandwidth dissipates because the sensor node needs to determine an adequate size for the contention window.

In order to solve such a problem, this paper proposes differential probability of selection MAC (DPSMAC), a novel MAC protocol that exploits the fixed size of a contention window and the random slot selection technique with non-uniform probability. The proposed DPSMAC protocol is advantageous in that it minimizes collisions between nodes and reduces the delay time, and it also maintains the fairness of service relatively and constantly despite the simplicity of the protocol's structure as compared to conventional 802.11 MAC protocols. <Table 1> shows a comparison of the properties of contention-based MAC protocols when applied to a ubiquitous sensor network.

<Table 1> Comparison of contention-based CSMA protocols

Protocol	Learning technique	Memory technique	Contention window	Probability distribution
802.11	○	×	Variable	Uniform distribution
MACAW	×	○	Variable	Uniform distribution
BMAC, SMAC	×	×	Fixed	Uniform distribution
Proposed DPSMAC	×	×	Fixed	Geometrical distribution

3 Simulation Results

In this section, we show the simulation results for the proposed DPSMAC protocol using ns-2 version 2.30 [11], a network simulator. In the simulation, all sensor nodes are considered to be located within a flat area served by a common base station, and a sensor node responding to a certain event among them transmits a small-sized report packet via the base station.

The experiment compared the total delay time of the DPSMAC protocol redesigned under the $p = 0.8$ condition when the number of contention slots is 32 with those in the SMAC and MACAW protocols [7-9][12]. The performance evaluation to measure the delay time and throughput was carried out around 20 times on average after setting different random initial values in a condition in which RTS/CTS was deactivated and the data packet of the sensor nodes was set to have a size of 40 bytes.

3.1 Packet throughput of proposed DPSMAC protocol

First, an experiment was carried out to measure the packet throughputs of the proposed scheme, SMAC, and MACAW, after saturating the wireless medium by varying the network load. In addition, in order to show that the proposed protocol functions normally even in an event-driven network environment, the experiment modeled the constant bit rate (CBR) flows that were simultaneously generated by 32 sending nodes in an ad-hoc steady state network environment, and the packet throughputs were measured.

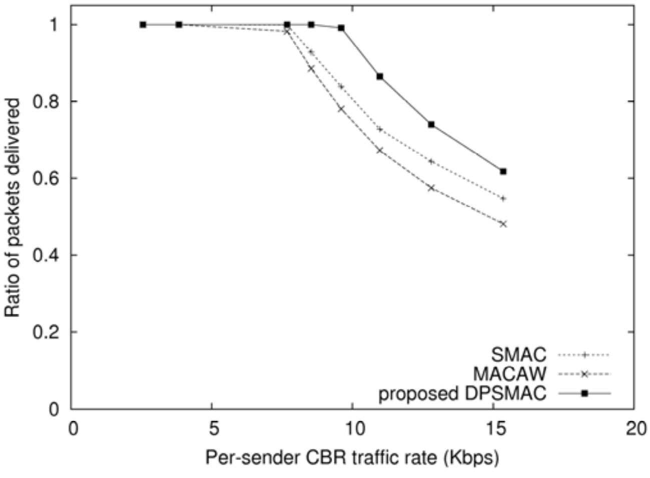


Fig. 4 Packet transmission rates of sensor nodes for CBR traffic

Fig. 4 shows that the packet throughput of the proposed protocol is superior to those of the other MAC protocols being compared as the CBR traffic is increased. It also indicates that the packet throughput of the protocols reduces slightly as the traffic is increased. It is considered that this reduction is not because of collision between transmitted packets but because the time required for the back-off procedure is comparatively increased because the slot selected in the contention window is relatively located toward the rear side.

First, in the proposed protocol, it might be known that packets collide to a comparatively lesser extent, and the location of a slot that succeeds one that transmits data in the contention window is moved toward the first half of the contention window due to the optimization of the slot selection probability distribution despite the increase in the number of sensors participating in the transmission. Such a fact can generally be considered to result in a reduction in bandwidth concentration when a transmission medium is accessed.

3.2 Transmission delay time of proposed DPSMAC protocol

An experiment was carried out to measure the transmission delay time of the proposed protocol for an event load generated at constant intervals. Then, the delay caused by the software system installed on the sensor node or the deviation caused by the electronic properties of the sensors is considered to the greatest extent possible in order to more accurately measure the variations in the delay time. In order to reflect this, the experiment added a random time of 0~1ms to the time at which each sensor sent its own event information.

The result of the transmission delay time measured in the experiment depending on the change in the number of sensor nodes is shown in Fig. 5.

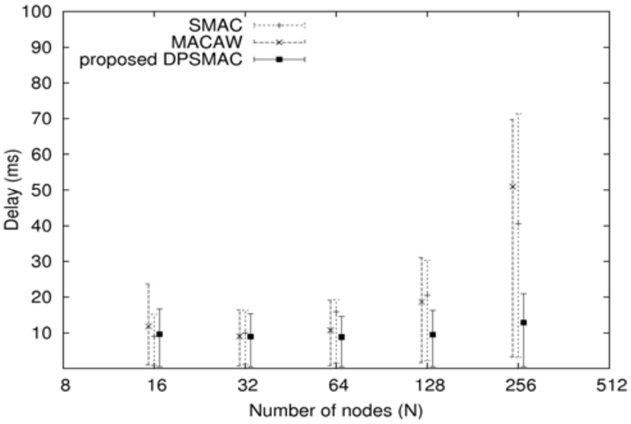


Fig. 5 Transmission delay times depending on number of sensor nodes

In Fig. 5, after arranging the packets sent from each sensor node by the order of arrival at the base station in terms of the percentile, the transmission delay time of the packets is respectively extracted at the first, middle, and last 90% locations to obtain the bottom, middle, and top points of the error bar, respectively.

The bottom of the error bar indicates that the minimum contention window size of SMAC and BMAC is sufficiently large to quickly solve the contention problem between nodes when there exist only a few sensors; however, it can be inferred that their contention window size should be continuously increased in advance as the number of sensors is increased in order to realize successful initial data transmission.

On the other hand, it can be confirmed that the proposed protocol requires constant time to solve the contention problem because the contention window size is fixed irrespective of the number of sensor nodes. In addition, the proposed protocol exhibits improved performance in that the total delay time required for data transmission for all sensor nodes is independent of the number of sensor nodes. Although the result measured the total transmission delay time when the sensor nodes were sending data, it is demonstrated that the proposed protocol maintains a performance that is at least equal to those of other contention-based MAC protocols for sensor networks.

3.3 Fairness test for medium access

Finally, it is considered whether or not the MAC protocol proposed in this paper fairly assigns bandwidth for sensor nodes to access the station in order to effectively solve the problem of starvation of service that can occur when some sensor nodes collide. The corresponding experiment simulated an even number of sensor nodes that could perform the role of both traffic sources and sinks within the range of wireless propagation. The size of the data packet of a sensor node was set to 1500 bytes, and RTS/CTS was activated for data exchange between nodes. To simulate a load environment in which the capacity of the wireless sensor network was exceeded, the congestion of packets transmitted from each sensor was also considered. Fig. 6 shows the throughput of each sensor node measured for 10 s.

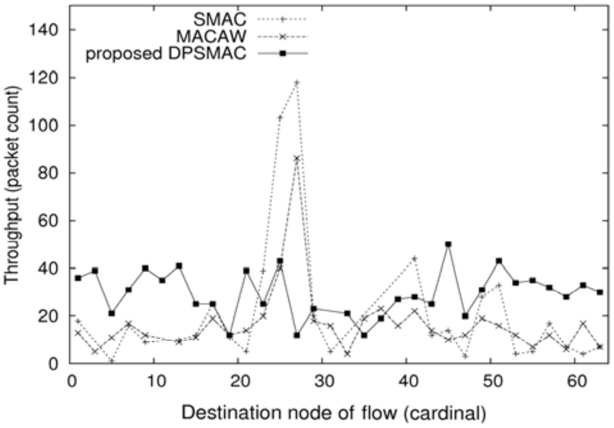


Fig. 6 Fairness comparison depending on medium access

As shown in the figure, the fact that the throughput of the proposed scheme is not concentrated in any certain sensor node but is generally evenly distributed indicates that it maintains relatively fair medium access as compared to conventional sensor MAC protocols.

Because practical ubiquitous sensor networks contain nodes sending redundant information, it can be considered that the proposed MAC protocol realizes an efficient network that is comparable to the distributed fair scheduling scheme despite the simplicity of the protocol's structure.

4 Conclusion

This paper proposed an efficient medium access technique suitable for a large-scale ubiquitous sensor network environment with high spatial density that can enable sensor nodes to effectively transmit information; this technique employs a modification of the conventional contention window technique. The proposed MAC protocol is highly adaptable in that it can constantly maintain the transmission rate to the greatest extent possible even if the environment changes frequently and unexpectedly. A simulation confirmed that the proposed MAC protocol functions normally even in practical conditions in which only some sensors transmit in response to a request to transmit data from the base station of a ubiquitous sensor network, whereas the other nodes are inhibited from transmitting.

First, the main concept of the proposed protocol is that the probability of selecting a slot is fundamentally differentiated to choose the distribution function with a geometrical probability that increases depending on the slot's location in order to select a slot located at the front part if possible when the sensor nodes select the transmission slot in a contention window. In addition, whereas most conventional CSMA/CA [13] based MAC protocols exploit a variable contention window technique, the proposed protocol exploits a fixed-size contention window to reduce the delay time.

In the future, the proposed MAC protocol should be applied to an actual event-based ubiquitous sensor network environment for a more practical performance evaluation, and the results of the performance should be analyzed and compared with those obtained through simulation.

5 Reference

[1] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E. (2002). A survey on sensor networks, IEEE Communications Magazine, 40(8), 102–116.

[2] Demirkol, I., Ersoy, C., Alagöz, F. (2005). MAC protocols for wireless sensor networks: a survey, IEEE Communications Magazine, 44, 115–121.

[3] Rajaravivarma, V., Yi, Y., Teng, Y. (2003). An overview of wireless sensor network and applications, in: Proceedings of the 35th Southeastern Symposium on System Theory, March 2003, pp. 432–436, Morgantown, WVa, USA.

[4] Abu-El Humos, A. (2005). Low latency and energy efficient MAC protocols for wireless sensor networks, Ph.D. Dissertation, Florida Atlantic University, Boca Raton, Florida.

[5] IEEE Std. 802.11-1999, Part 11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Reference number ISO/IEC 8802-11:1999(E), IEEE Std. 802.11, 1999 edition, 1999.

[6] Gast, M. (2005). 802.11 Wireless Networks: The Definitive Guide, second ed., Sebastopol, CA: O'Reilly Media, Inc.

[7] Polastre, J., Hill, J., Culler, D. (2004). Versatile low power media access for wireless sensor networks, in: Proceedings of the Second ACM Conference on Embedded Networked Sensor Systems (SenSys), November 2004, pp. 95–107, Baltimore, MD, USA.

[8] Ye, W., Heidemann, J., Estrin, D. (2004). Medium Access Control with coordinated adaptive sleeping for wireless sensor networks, IEEE/ACM Transactions on Networking, 12(3) (June 2004) 493–506.

[9] Ye, W., Heidemann, J., and Estrin, D. (2002). An energy efficient MAC protocol for wireless sensor networks, in: Proceedings of IEEE INFOCOM, vol. 21, no. 1, 2002, pp. 1567–1576.

[10] Bharghavan, V., et al. (1994). MACAW: A Media Access Protocol for wireless LANs, in: Proceedings of ACM SIGCOMM '94, vol. 24, no. 4, 1994, pp. 212–225.

[11] USC ISI, ns-2 Notes and Documentation, 2007.

[12] Woo, A. & Culler, D. (2001). A transmission control scheme for media access in sensor networks, in: Proceedings of ACM MobiCom '01, July 2001, pp. 221–235, Rome, Italy.

[13] Bianchi, G. (2000). Performance analysis of the IEEE 802.11 Distributed Coordination Function, IEEE Journal on Selected Areas in Communication, 18(3), 535–547.

Hardware Implementation of the Energy-Efficient Hybrid Key Management Protocol for Wireless Sensor Networks

E. Taqieddin[1], M. Zawodniok[2], S. Jagannathan[2], and A. Miller[2]

[1]Department of Network Engineering and Security, Jordan University of Science and Technology, Irbid, Jordan

[2]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO, USA

Abstract - *Dynamic key distribution has been proposed as an alternative for the pre-deployment of keys in wireless sensor networks. The keys are generated through the collaboration of the nodes in the network. Each participating node sends a partial key to a Head Cluster Head that later calculates the sub-network key and distributes it. The advantage of this approach is that the keys change frequently and adapt to the changes in the created clusters within the network. Furthermore, the malicious disclosure of a key will have limited effect. In previous work, Energy-efficient Hybrid Key Management (EHKM) was proposed for dynamic key distribution. The protocol's validity was shown through simulation. In this work, we implemented EHKM and an extended version of it Extended EHKM (EEHKM) on the UMR/SLU motes to get a better understanding of how the protocol would behave under the actual constraints of hardware that are not shown by simulation.*

Keywords: Energy, Delay, Security, Key Distribution

1 Introduction

The numerous applications of Wireless Sensor Networks (WSN) have encouraged the research in this promising field. Some examples of such applications are military operations, humanitarian relief, and medical services. A WSN consists of a set of nodes that collaborate with each other to guarantee proper communication between any node in the network and the Base Station (BS).

Unlike wired networks, the nodes in a sensor network have no fixed infrastructure and are limited in communication range. Moreover, they lack the abundant resources available in wired networks. This dictates the use of energy efficient and resource friendly routing algorithms.

The information in WSN is transmitted using RF channels which may pose a security threat. This calls for security features to guarantee the confidentiality and authenticity of the data delivered to the BS. This is especially needed in situations where the nodes are deployed in hostile environments.

The corner stone of security is the proper distribution and safe handling of the cryptographic keys. Any compromise of a node or the way the keys are distributed can affect part or the whole network. One approach for key distribution is to use public/private key pairs. Although this option would provide reasonable security, it does not serve well in WSN due to the lack of a trusted certification authority, the high energy consumption, and computational expensiveness.

The use of shared keys can provide the needed security while maintaining the limited energy and computational resources. One approach is to load the encryption keys into the nodes' memory before deployment. This may be done by using a global key shared by all the nodes in the network [4], [5]. Alternatively, the nodes can be pre-deployed with unique key pairs for every pair of nodes in the network. The former approach may be defeated if one or more nodes are compromised which gives the attacker access to the global key. The latter, on the other hand, is not susceptible to such a threat (a node compromise can only affect the traffic of that node) but it is resource consuming. Each node in a network of N nodes should store (N-1) keys for the pair-wise communications. In [7], it was shown that the number of keys increases exponentially with the size of the network. A tradeoff between the two approaches is to load the nodes with key rings consisting of sets of random keys selected from a larger pool of keys. The ides is that any two nodes in the network will share the same key with a certain probability p that depends on the number of keys in the key ring and the total keys in the pool [6], [8].

The use of dynamic key agreement is also proposed in the literature [1], [13]. Dynamic keys have the advantage of reduced storage requirements compared to pre-deployed keys. Moreover, the frequent changes of the keys limit the access of an intruder to them. In [13], a protocol for dynamic key generation is presented in which a cluster key is computed starting from the leaf nodes up to the Cluster Head (CH). Every node in the group uses the partial keys of its children as inputs of a function ($\alpha^{k_1 \oplus k_2} \bmod p$) where p is a prime number, α is the primitive root of p and k_1, k_2 are the partial input keys of the function. The CH computes the cluster key and securely broadcasts it to the group.

In [1], a mechanism called Energy-efficient Hybrid Key Management (EHKM) is given in which separate keys with different purposes are maintained by each node. These keys are either pre-deployed or dynamically generated using a method similar to that given in [13]. The goal of EHKM is to diversify the levels of security in the network since not all the data on the network need to be handled with the same level of security. EHKM enables the nodes in the network to operate in a low-security energy-efficient mode using static keys, while dynamically creating keys for high security sub-networks.

Most of the key management approaches mentioned above rely on simulations using NS2 or GloMoSim. This may be useful in cases where the performance metrics of two protocols are compared against each other. Nevertheless, these simulations do not give a fair assessment of the behavior of the protocols when run on hardware. Energy consumption, memory constraints, collisions and topology can be modeled in a simulator but the value of simulating them would be only as good as the model given to the simulator. The motivation of this research is to implement and evaluate the performance of EHKM and Extended-EHKM (EEHKM) on hardware. This study serves two purposes. First, it identifies the effects of hardware constraints on the overall performance of the protocol. Secondly, it gives a foundation for the improvement of the simulators. The actual testing results can be fedback into the simulator through a modified network model that closely resembles that of the hardware in terms of energy consumption and delay.

2 Overview of EHKM

The hybrid nature of EHKM is a result of running two different key management schemes to handle the different keys along with their varying levels of security requirements. Both dynamic and group wide distribution approaches are used.

EHKM assumes that the nodes will be safe from being compromised for a time T_{min} after deployment. Furthermore, it is assumed that the nodes form a sub-network whenever a sensing event occurs and that the nodes know the number of nodes that are members of the same sub-tree or the same level in the tree. Finally, it is assumed that the nodes employ a self-organizing protocol for selecting the CH nodes such as OEDSR [3].

In EHKM, three different types of keys are used to support the various levels of security and network applications.

Group-wide keys: Two group-wide keys are pre-deployed into each node. One key is used for group wide communication between nodes that are not involved in the sub-network (K_1). The other key, K_2, is used for the exchange of pair-wise keys as will be explained later.

Individual key: This is a unique key shared between the node and the BS for private communications (K_3). This key is also pre-deployed.

Sub-network key: This key is dynamically calculated whenever the nodes in the network form a sub-network.

At the beginning of the operation of the network, the nodes start the process of sharing a new random pair-wise key (K_r) to be used instead of K_2 that is common to all nodes. This needs to be replaced since the whole network may be compromised if a successful attack on at least one node takes place.

To share K_r, the nodes transmit KEY_HELLO messages that contain the node ID, the new key encrypted using K_2 and a MAC of <node ID, k_r>. Upon reception of the KEY_HELLO message, a node stores the node ID and the new key in a table for future use. Thus, at the end of this process each node will hold the key to be used to encrypt the packets by simply referring to the key table using the destination node ID. After the period of T_{min} each node must clear K_2 from its memory.

The sub-network key management protocol of EHKM is designed in a way to maximize the efficiency of the energy consumption in the network. Here, the nodes collaborate with each other to calculate a sub-network key. The main idea is that all the CH nodes collaborate to select the CH with the highest average energy reserves in its cluster, i.e, each CH calculates the average energy of all the nodes that belong to its sub-tree according to

$$\text{avg\_energy}(i) = \sum (E_{ij}) / n_i \qquad (1)$$

Where i is the cluster index and j is the node index within cluster i.

These values are advertised to all other CHs to select the one with the highest average energy as the Head Cluster Head (HCH) according to

$$\text{HCH} = \text{CH}(\max(\text{avg\_energy}(i))) \quad (2)$$

The HCH in cooperation with the other CHs generates a dynamic key using partial keys from the nodes in the network. The HCH checks whether there are sufficient nodes in its sub-tree to generate the desired key. In case there are less than m nodes then the HCH has to use keys produced by the members in the sub-tree of the CH with the second highest average energy. This process is repeated until m nodes are chosen to provide the needed partial keys.

The next step is to initiate the creation of the partial keys using a START_ALGORITHM message sent from the CH to the nodes in the sub-tree. The message contains the cluster ID and the depth metric which dictates the level that the message should reach within the sub-tree before a partial key is created. If all the nodes in the sub-tree are required to generate a partial key then the CH sets the depth field to -1. This is actually done in all sub-trees contributing the m keys except for the last one in which the depth field is set to

$$d = m - \sum n_{other\_clusters} \qquad (3)$$

Each node that receives the START_ALGORITHM message checks whether the depth field is -1 or not. If it is -1, then it sends the packet to the leaf nodes. Once the leaf nodes receive the message, they generate their partial keys and send them up the hierarchy to finally reach the HCH. For the nodes in the last sub-tree in which the depth (d) is set to some positive value, the node checks how many other nodes are at the same level within the cluster and subtracts that number from d. If new value of d is negative then it does not need any more contributions from the nodes below in the hierarchy and sends its partial key. Otherwise, it updates the value of d in the START_ALGORITHM message and forwards it.

After the HCH gets all the m partial keys from the participating clusters, it calculates the sub-network key and sends it out to the nodes after encrypting it using the new key (K_r) of the destination node. This key will be transmitted m times and in each time it will be encrypted with the key K_r of one of the nodes that participated with a partial key.

Following is a pseudo-code of the dynamic key calculation [1]

```
I. After creation of a sub-network

Set cluster_heads = {Each cluster head}
For each cluster head
  Set avg_energy(cluster)=
average(energy_in_nodes_cluster)
  Set $n_{cluster}$ = number of nodes in cluster
  Set HCH ={Cluster head: max(avg_energy(cluster_heads)}
  Set chosen_HCH = {HCH}
  cluster_heads = {cluster_heads} – {HCH}
  Set m_temp = m - $n_{HCH}$
  while m_temp ≥ 0
  CH_temp={cluster head:max(avg_energy(cluster_heads)}
  chosen_HCH = {chosen cluster heads} ∪{CH_temp}
  m_temp = m_temp – $n_{CH\_temp}$
  Set n_chosen = 0
  Set depth = -1
while |chosen_HCH| > 1
  CH_temp = first_element(chosen_HCH)
  CH_temp broadcasts start_algorithm {Cluster_ID || depth}
   n_chosen = n_chosen + $n_{CH\_temp}$
  chosen_HCH = {chosen_HCH} – {CH_temp}
  depth = m – n_chosen
  CH_temp = only_element(chosen_HCH)
  CH_temp broadcasts start_algorithm{Cluster_ID || depth}
```

```
II. After hearing a start_algorithm message
If node is a leaf
  Start_algorithm( ); return
Else
  depth = received_depth – $n_{ij}$ ($n_{ij}$ is the number of nodes in
cluster i on level j)
If depth ≤ 0
  Start_algorithm( ); return
Else
  Broadcast start_algorithm{Cluster_ID || depth}
```

3 Hardware implementation of EHKM

This section gives an overview of the hardware implementation of EHKM. We present a description of the capabilities, limitations and support for networking applications as well as the actual implementation details on the BS and the UMR/SLU motes.

3.1 Overview of hardware and associated limitations

The hardware used was chosen to be energy-conservative, performance-oriented and of small form-factor. For a low-power consumption, fast 8-bit processing and ease of interface to peripheral hardware components, the Silicon Laboratories® 8051 variant was chosen with the external RAM, UART interface and A/D sensing, the microcontroller is capable of performing various tasks done by sensor nodes. The Maxstream XBee^{TM} RF module was used for communication.

To implement the algorithm on hardware, many limitations had to be taken into consideration. The speed, precision and storage requirement all became factors in making a decision on which microcontroller to use. For EHKM, the initial design goals were to limit the energy consumption while providing a reasonable level of security. This was a factor in our hardware implementation as the options of low-power consumption and powerful 8-bit processor architectures were limited. Other factors that were considered are the limitations on the available memory (since EHKM stores different types of keys) and processing speed (encryption/decryption can be time consuming in some cases).

3.2 Sensor node hardware

The Generation-4 Smart Sensor Node (G4-SSN) was used to perform the functionality of the sensor nodes in our hardware model. G4-SSN was originally developed at the University of Missouri – Rolla (UMR) and updated afterwards at St. Louis University (shown in Figure 1). This node has several abilities for sensing such as strain gauges, accelerometers, thermocouples, and general A/D sensing. It is also capable of performing processing tasks such as analog filtering, Compact Flash memory interfacing and 8-bit data processing at a maximum of 100 MIPS. Table 1 gives a summary of the specifications of the G4-SSN [2][3].

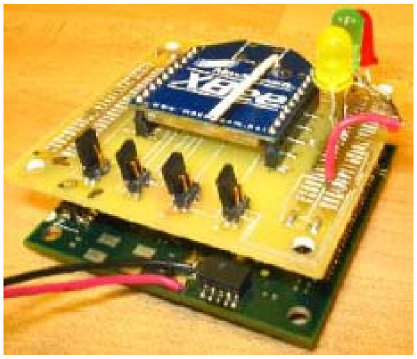


Figure 1. UMR/SLU G4-SSN

Table 1. G4-SSN specifications

	G4-SSN
I@ 3.3 V [mA]	35
Flash Memory [bytes]	128k
RAM [bytes]	8448 bytes
ADC Sampling Rate [kHz]	100@ 10/12-bit
Form-Factor	100-pin LQFP
MIPS	100

3.3 Implementation details

EHKM extends the software architecture that was developed in [3] by modifying the behavior of the nodes and the control packets they receive from the BS.

Figure 2 shows the software architecture of our hardware implementation which illustrates the three layer approach that is used in this development. The various layers provide flexibility since the layer specific details may be changed with minimal effect on the overall system. The Application layer is responsible of handling all sensor data processing. The Physical layer sets up the serial interface between the microcontroller and the radio module. The layer in between is responsible of the queuing, scheduling, routing, and message abstraction. Note that EHKM falls within the same layer but runs in parallel to the sub-layers.

Before implementation, we considered the possibility of further enhancements on the protocol. Our first observation was that it is resource consuming to distribute the sub-network key using the unicast pair-wise approach described above. Instead, we propose that a node encrypts the message containing the sub-network key using its own K_r. This message can then be sent as a broadcast message and all the nodes in the sub-network will be able to decrypt it using the key K_r of the HCH.

In addition to the existing packets in the routing protocol implementation in [3], we added several packets that are sent either from the BS to the nodes or among the nodes themselves. These packets are essential for the proper exchange of the various keys in the network. Following is an explanation of the purpose of each packet.

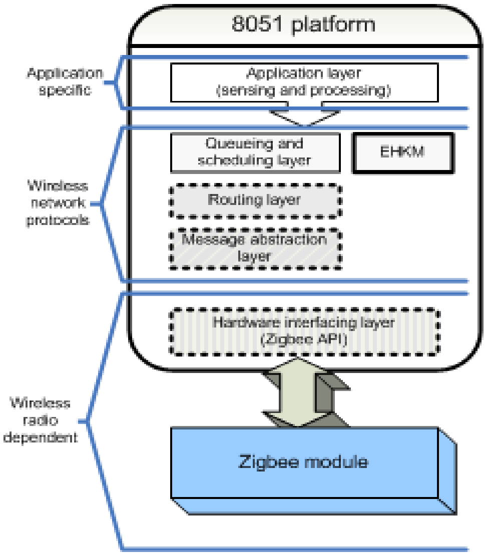

8051 platform
Application specific
Application layer (sensing and processing)
Queueing and scheduling layer
EHKM
Wireless network protocols
Routing layer
Message abstraction layer
Hardware interfacing layer (Zigbee API)
Wireless radio dependent
Zigbee module

Figure 2. Software architecture of EHKM

START_KEY_EXCHANGE: The BS initiates the key exchange between the nodes at the beginning of the operational lifetime of the network. This will mark the beginning of the period running for T_{min} seconds. The nodes proceed with their key exchange upon receiving this packet by sending a KEY_HELLO message.

KEY_HELLO: During the interval that starts when a START_KEY_EXCHANGE is received and ends T_{min} seconds later, the nodes in the network generate their new key K_r to replace the old pair-wise key K_2. The message will be sent to all one hop neighbors in a broadcast mode in the following format.

N_i → all one hop neighbor: $E_{K2}(\text{Node\_ID}, K_r)$

During T_{min}, the key K_2 is known by all the one hop neighbors and can be used to decrypt the message. After T_{min} the nodes will delete K_2 from the memory and use K_r Instead.

START_ALGORITHM: This triggers the partial key exchange discussed earlier.

PARTIAL_KEY_EXCHANGE: Each node generates a random partial key and transmits it to the node in the level above it in the sub-tree until it reaches the HCH that calculates the final sub-network key.

SUB_NETWORK_KEY: After the final key is calculated it is sent to the sub-network members in an encrypted packet. The message in EHKM will differ from that of the EEHKM. For EHKM, the packet is encrypted using the key K_r of the intended destination and transmitted as a unicast (once for each intended recipient) to node i (i = 0, 1,…, m). EEHKM, on the other hand encrypts the packet once using its own K_r and broadcasts it. The recipients of the message are responsible of decrypting it using K_r of the HCH.

Figure 3 gives a graphical explanation of how EHKM and EEHKM handle the various packets for their operation.

4 Hardware setup and results

We ran the Optimal Energy Delay Senor Routing protocol (OEDSR) on the network under test. The choice of the protocol has no effect on the operation of the key management technique and is only responsible of routing the various packets within the network. The sensor nodes employ 802.15.4 modules that transmit at data rate of 250 kpbs and interfaces to the node processor at 38.4 kbps. These modules run at a low-power of 1-mW whereas the module connected to the BS transmits at 100-mW to extend its range.

The experiments were conducted on a network consisting of 6 nodes. The nodes generated around 3.2 packets/sec with a packet length of 100 bytes of which 12 were header bytes. The experimental cases were

1) Nodes using OEDSR with no key management.

2) Nodes using OEDSR with EHKM.

3) Nodes using OEDSR with EEHKM.

The first experiment was run 10 times for each of the 3 cases to find the throughput and the change in the energy consumption.

From Table 2 we notice the increase in the overhead energy consumption. This is attributed to the added overhead of packets for the functionality of the key management schemes. The other contributing factor is the encryption of the packets. We note that EEHKM results in better energy management and this is attributed to the reduction of the repeated encryptions and transmissions that were eliminated from the original EHKM.

Table 2. Average percentage of routing energy and bit rate

	OEDSR w/o EHKM	OEDSR with EHKM	OEDSR with EEHKM
Avgerage percentage routing energy	22.5	32.4	29.1
Avgerage bit rate	6.28	5.84	6.15

As for the throughput, we studied the bit-per-second of data that were sent out. We see that OEDSR with no key management gives a higher data rate but EEHKM is relatively close to it while EHKM has the least throughput. This decrease in the throughput may be attributed to the added overhead and the bottlenecks around the CH nodes.

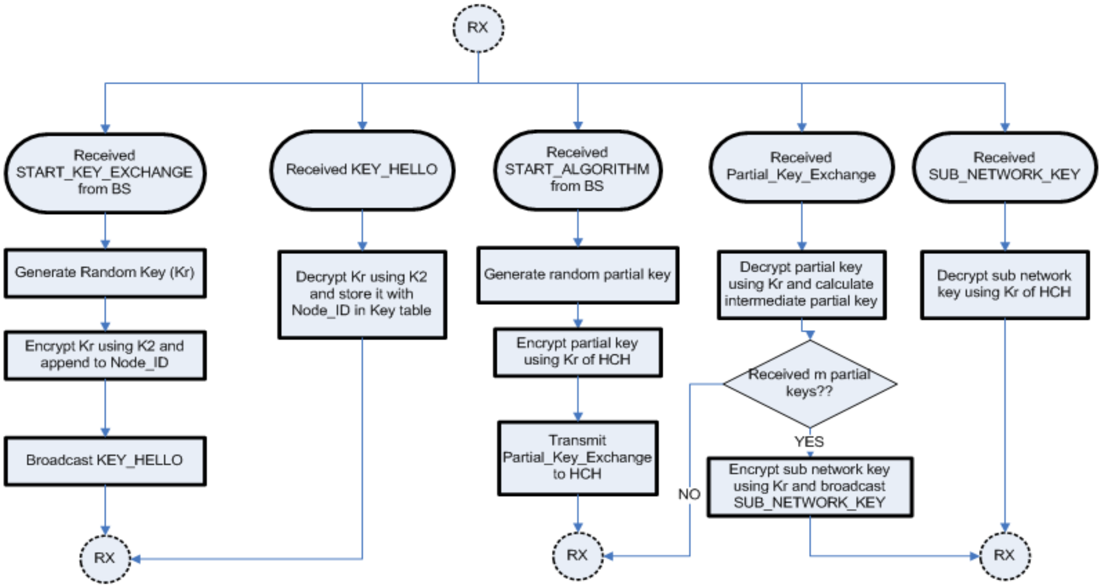

RX
Received START_KEY_EXCHANGE from BS
Generate Random Key (Kr)
Encrypt Kr using K2 and append to Node_ID
Broadcast KEY_HELLO
RX
Received KEY_HELLO
Decrypt Kr using K2 and store it with Node_ID in Key table
Received START_ALGORITHM from BS
Generate random partial key
Encrypt partial key using Kr of HCH
Transmit Partial_Key_Exchange to HCH
RX
Received Partial_Key_Exchange
Decrypt partial key using Kr and calculate intermediate partial key
Received m partial keys??
YES
NO
Encrypt sub network key using Kr and broadcast SUB_NETWORK_KEY
Received SUB_NETWORK_KEY
Decrypt sub network key using Kr of HCH
RX

Figure 3. Packet handling in EHKM implementation

Next, we studied the end-to-end delay to determine the effect of applying EEHKM. The results are given in seconds in Table 3.

We notice that the application of the encryption algorithm increased the average end-to-end delay by about 15%. This is reasonable given the computational requirements. The choice of a low cost encryption scheme such as the Tiny Encryption Algorithm (TEA) can reduce the delay.

Table 3. Average end-to-end delay comparison

Test	OEDSR w/o EHKM (Sec)	OEDSR with EEHKM (Sec)
Test # 1	0.1728	0.2013
Test # 2	0.1537	0.2711
Test # 3	0.1674	0.1792
Test # 4	0.1456	0.1628
Test # 5	0.1911	0.1813
Test # 6	0.1731	0.1623
Average	0.1673	0.1930

5 Conclusions

Two key management protocols were implemented on hardware for performance analysis. The study aimed at investigating the behavior of the protocol on hardware. The results of this work can be applied to the existing simulation model for simulations of different scenarios where a wide area or a high number of nodes is needed.

There is always a tradeoff between applying security in the network and the energy, delay, and throughput. This was the case in this study as we noticed a drop in the throughput and an increase in the end-to-end delay and overhead energy consumption. An increase of about 7% on the energy consumption and around 15% on the delay were observed. These values are relatively low and may be acceptable given the added security in the network.

The nodes demonstrated different behavior than that of what was seen in simulation. For example, the effect of contention was clearly apparent in some tests when all the nodes started transmitting at the same time.

Our future direction with this work is to improve the random key generation. Currently, our model runs a simple random generator which is clearly not a safe option from a cryptographic point of view. We will investigate the use of an advanced random generator such as the Pseudo Random Number Generator such as Fortuna. Another aspect to be studied is the priorities given to key management packets. Currently, all the control packets flow through a FIFO queue while we believe that a more suitable implementation is to assign higher priorities for such packets. Finally, we will develop attack scenarios and traffic analysis techniques to study the effectiveness of this protocol.

6 References

[1] Landstra, T., Zawodniok, M., Jagannathan, S., "Energy-Efficient Hybrid Key Management Protocol for Wireless Sensor Networks", International Journal of Network Security, 9(2), 2009, 121-134.

[2] Fonda, J., Zawodniok, M., Jagannathan, S., and Watkins, S. "Adaptive Distributed Fair Scheduling and its Implementation in Wireless Sensor Networks", Proceedings of the IEEE Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 2006, 3382-3387.

[3] Fonda, J., Zawodniok, M., Jagannathan, S., and Watkins, S. "Development and Implementation of Optimized Energy-Delay Sub-network Routing Protocol for Wireless Sensor Networks", Proceedings of the IEEE International Symposium on Intelligent Control, Munich, Germany, 2006, 119-124.

[4] Anderson, R., and Kuhn, M., "Tamper Resistance – A Cautionary Note", Proceedings of the second USENIX Workshop on Electronic Commerce, Oakland, CA, USA, November, 1996.

[5] Basagni, S., Herrin, K., Rosti, E., and Bruschi, D., "Secure Pebblenets", Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2001), Long Beach, CA, USA, October, 2001, 156 – 163.

[6] Blundo, C., Santis, A., Herzberg, A., Kutten, S., Vaccaro, U., and Yung, M. "Perfectly-secure key distribution for dynamic conferences", Advances in Cryptology, Proceedings of CRYPTO'92. LNCS 740, 1993, 471–486.

[7] Carman, D., Kruus, P., and Matt, B. "Constraints and Approaches for Distributed Sensor Network Security", NAI Labs Technical Report #00-010, September 2000.

[8] Chan, H., Perrig, A., Song, D. "Random Key Pre-distribution Schemes for Sensor Networks", Proceedings of the 2003 IEEE Symposium on Security and Privacy (SP'03), Oakland, CA, USA, May, 2003.

[9] Culler, D., Estrin, D., and Srivastave, M. “Overview of Sensor Networks”, IEEE Computer, 37, 8 (2004), 41 – 49.

[10] Du, W., Deng, J., Han, Y., Chen, S., and Varshney, P. A “Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge”, Proceeding of the 23rd Conference of the IEEE Communications Society (INFOCOM ‘04), Hong Kong, March, 2004.

[11] Eschenauer, L, and Gligor, V. “A Key-Management Scheme for Distributed Sensor Networks”, Proceedings of the 9th ACM Conference on Computer and Communication Security, November, 2002.

[12] Kim, Y., Perrig, A., Tsudick, G. “Tree-Based Group Key Agreement”, ACM Transactions on Information and System Security, 7(1): 60 – 96, February, 2004.

[13] Panja, B., Madria, S. and Bhargava, B., “Energy and Communication Efficient Group Key Management Protocol for Hierarchical Sensor Networks”, Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2006), Taichung, Taiwan, June, 2006.

[14] Price, A., Kosaka, K., Chatterjee, S., “A Secure Key Management Scheme for Sensor Networks”, Proceedings of the Tenth Americas Conference on Information Systems, New York, NY, August, 2004.

[15] Zhu, S., Setia, S., and Jajodia, S., “LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks”, Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS ’03), Washington D.C., October, 2003.

Intelligent Node Placement using GA (INPGA) Protocol in Wireless Sensor Networks

Zeynep Orman[1*], Ali Norouzi[1], Faezeh Sadat Babamir [2]

[1] Department of Computer Engineering, Istanbul University, Istanbul, Turkey
[2] Department of Computer Science, Shahid Beheshti University of Tehran, Tehran, Iran

Abstract: One of the main design issues for wireless sensor networks is the node placement (NP) problem. This paper investigates the optimization of Wireless Sensor Networks (WSNs) layout. All sensor nodes in the environment are required to be connected to high-energy level nodes which operate as relay from environment to base or ground to a satellite in order to transmit aggregated data. Long distance transmission by sensor nodes is not efficient in terms of energy since energy consumption is a super linear function of the transmission distance. In this study, the sensors are assumed to communicate in fixed range. Our framework for WSNs optimizes two competing objectives which are the total sensor coverage and the lifetime of the network. By using the genetic algorithm, our approach obtains a solution which contains a set of minimum number of nodes to cover the sensing range with optimum energy consumption.

Keywords: Wireless Sensor Network, Clustering, Routing Algorithm, Optimum consumption of energy, Node Placement.

I. INTRODUCTION

WSNs consist of large number of sensing devices, which are equipped with limited radio communication capabilities and limited computing features. They are used for many purposes in the present and the future. The realization of WSNs poses a lot of challenges in system and network design, algorithm and protocol design, and query language and database design. The primary issue under focus which is critical to the proper functioning of wireless sensor networks is the clustering techniques that can save energy consumption. Some applications just need the aggregated value to be sent to the sink (base station), so the sensors collaborate between each other to send more accurate information about their local locations. In order to help sensor nodes to aggregate data in efficient manner, node placement should be applied. Since energy consumption during the communication can be considered as major energy depletion parameters, the number of transmissions must be reduced as much as possible to achieve the extended battery life [1, 2]. On the other hand, the recent advances that witnessed to WSNs made it interested in the applications that need high-deployment, i.e. environmental monitoring applications, where the sensor nodes are left unattended in order to report the parameters of interest like humidity, temperature, light, etc. [11]. Because of the difficulty of recharging node batteries in the case of the high-density deployment, the energy efficiency became a major design goal in WSNs.

In recent years, there are many significant interests toward WSN especially in optimization of BS location problem, but research on BS is never done. In [4], Chakrabarty et al. devote research using Integer Programming (IP) , while Bulusu et al. [5], Dhillon et al. [6] and Howard et al. [7, 8] utilize different greedy heuristic rule to incrementally deploy the sensor. Zou et al. [9] inspect Virtual Force Methods (VFM) for the sensor deployment (VFM often apply in the deployment of robots [8]).

In this article, we present some optimal placement strategies and through numerical results, we show that the optimal node placement strategies provide significant benefit over a commonly used uniform placement scheme.

II. BACKGROUND

WSNs are used in several areas including: military, medical, environmental and household. But in all these fields, energy, data lost, delays have determining role in the performance of wireless sensor networks [3].

Therefore, in order to balance the energy consumption between the nodes, the node placement process should be applied to all sensor nodes.

2.1 Definitions

Before heading with the details of the research, we will first describe some basic concepts.

Clusters: They are the organizational unit for WSNs.

Cluster Heads: They are the organization leader of a cluster. They are often required to organize activities in the cluster. These tasks include but are not limited to data-

* Corresponding author. Phone:+90 212 4737070
E-mail address: ormanz@istanbul.edu.tr

aggregation and organizing the communication schedule of a cluster.

Base station: In WSNs, a base station is a wireless communication station installed at a fixed location to communicate and used to gather data from other nodes.

Node Placement: Nodes are deployed on the campus network to ensure the network can have the best performance.

A suitable node placement protocol, in addition to balancing the energy consumption between all nodes, provides a better network throughput under high load by reducing the channel contention and packet collisions. Figure 1 denotes an example of a node placement scheme.

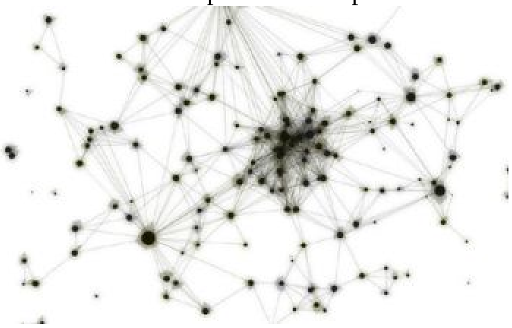


Figure 1: Node Placement Scheme

Generally, advantages of suitable sensor propagation in WSNs are [2]:

• **Scalability:** number of the deployed nodes in the network can be high, that is for the limited number of transmissions between the nodes.

• **Collision reduction:** because the cluster head (CH) works as a coordinator, the number of nodes that access the channel is limited and the communication between the cluster members and the cluster head is local.

• **Energy efficiency:** the periodic relocation causes the network to consume more energy. However, by periodic relocation, the duties of the CH are distributed among all other nodes and this leads to less energy consumption.

•**Low Cost:** Placing sensors at appropriate locations can prevent the excess costs.

• **Routing backbone:** the CH aggregates the collected data by cluster members and sends it to the sink. This means the network can be built with a little route-thru traffic and an efficient routing backbone.

2.2 Research Objectives

Application requirements can often determine the NP objectives. Some important objectives for network clustering are as follows:

• ***Load balancing:*** As CH's duties are more than the cluster members, it is important to relocate the network periodically in order to distribute the CH's duties to other nodes.

• ***Increased connectivity and reduced delay:*** Since the CH aggregates the collected data and sends it to the base station through other CHs, the inter-CH connectivity is an important requirement. However when the latency is important, intra-cluster connectivity also becomes an objective of the design.

• **Minimal cluster count:** Some CHs could be expensive, for example they can be laptop computers, robots or maybe a mobile vehicle, and so the designer wills to employ such expensive and vulnerable nodes as less as possible. Moreover, the small number of deploying such nodes could be due to the complexity nature of these nodes.

• **Maximal network longevity**: Minimizing the energy consumption for the intra-cluster communication, placing cluster members near to their respective CHs and distributing the load to other nodes by relocation will lead to maximize the network life.

III. NEW ALGORITHM

3.1 Genetic Algorithm

Genetic Algorithm is a search technique inspired by the evolution nature. This kind of algorithm operates stochastically and starts to generate new population with primary individuals. Every chromosome includes several gens which are binary, real, etc. number. After any generation, all chromosomes will be evaluated according to their fitness (in fact fitness is our criterion) to make new set of better chromosome. This new population gradually biases toward to an optimum solution.

3.2 WSN Modeling

We utilize a flat square surface where nodes can be monitored within the 'coverage' function to experience our approach. The sensors can communicate with each other within 'R'. All aggregated data are transmitted from upstream to downstream via hops. This assumption is just for convenience and surely does not prevent generalizability. Each node primarily has a specific energy level in its battery which will be arbitrary reduced by a transmission period.

3.3 Formulating

Because of the massive calculations and in order to gain a suitable network, we sever the problem into two sub-problems.

In the first step, we obtain minimum number of nodes which are required to monitor the network area using the genetic algorithm. Usually, we can obtain a couple of solutions which result in the same efficient network.

In the second step, using the prime and Dijkstra algorithms, the connectivity of nodes are evaluated. The two objectives are mostly considered are the coverage and the lifetime of the network. The covered area is calculated by the area of the union of the disks measured by radius 'R' centered at each connected sensor that is normalized by the total area and this is given with the Equation (1).

$$coverage = \begin{cases} \bigcup_{i=1}^{n} R^2_{x_i,y_i} \;\therefore |(x_i,y_i)-(x_{i-1},y_{i-1})| = (0,0) \\ \bigcup_{i=1}^{n} R^2_{x_i,y_i} - r^2_{x_i,y_i} \;\therefore o.w. \end{cases} \qquad (1)$$

$$where -> r_{x_i,y_i} = |(x_i,y_i)-(x_{i-1},y_{i-1})|$$

In 2005 [10], Quint et al. proposed the formula given by (2) to calculate the required energy for covering some purpose points which is called as the purpose function.

$$\min\_\ f = \sum_{i\in s} (\varepsilon + d_i) * y_i + \sum_{i\in D} NC_j * h_j \qquad (2)$$

In this equation, the constant ε is a required energy to setup a node and d_i represents the cost of routing between upstream through downstream. This mount is computed before launching the program by using the Dijkstra algorithm. This value is a kind of penalty for the remote nodes. NC is the surcharge of uncovering points; y_i and h_j show the activation status of node i and the covering status of point j, respectively.

$$fitness - function = \min\_\ f_{i,j} / coverage_{i,j} \qquad (3)$$

Equation (3) is our proposed fitness function which considers both the proposed coverage formula and the energy procedure. In this formula, we gain the ratio of the energy level and the mount of energy in order to design the appropriate network.

IV. SIMULATION AND EVALUATION

The considered space of WSN connectivity optimization is significantly non-linear, due to the binary nature of the communication medium between sensors; a little sensor movement will cause large effects in both objectives (even disconnecting the network). The GA was chosen to carry out the optimization since it has a good efficiency with non-linear objectives. Using the WSN simulator, we simulated our approach to show the efficiency of the network. The figures from 1 to 6 represent the algorithm performance step by step. We try to adapt the applied protocol to our approach to verify the sensor nodes, the power consumption and more importantly, the lifetime of the network. The right-hand sides of the figures are designed with a striped pattern which shows the uplink zone. Fig. 2 represents the network in the early monitoring. A green bit shows the data package and the series of red circles surrounded by gray circles are the network where gray circle is a sensor detection range [10].

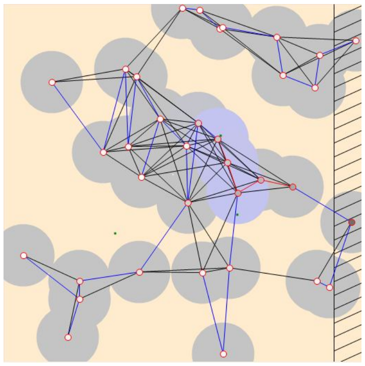


Fig.2: Network in the early monitoring

Fig. 3 represents the network by passing of time. As it is obvious, some sensor nodes are depleted for transmitting data. Several bluish show the active nodes that red edges are the active communication lines.

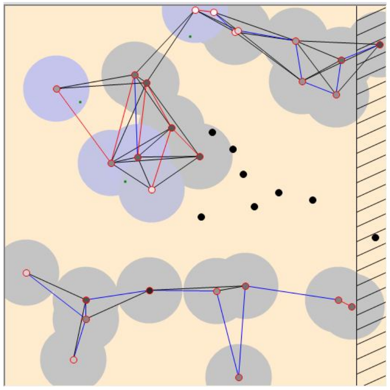


Fig.3: Some nodes are in-actives

Figure 4 contains small circles and this represents a node that is depleting energy. This network is disconnecting which few data can be monitored.

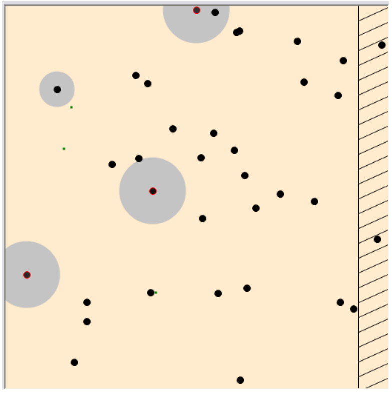


Fig.4: small circles show reducing the corresponding power

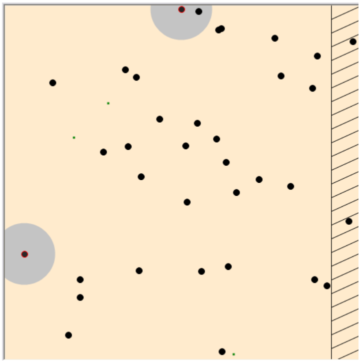


Fig.5: Network is disconnecting

Figure 5 shows the network with just two lived nodes. This network can cover the least area. Fig. 6 shows the dead network with no monitoring.

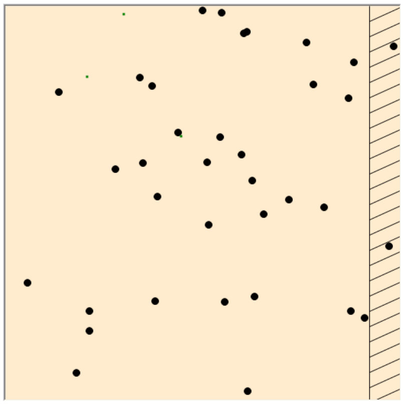


Fig.6: Dead network

Fig. 2 represents the network in the early monitoring.
As you see fig. 6 represents the dead network .These figures approve our improvement which can extend the lifetime of the network. Figure 3 shows the network in progress with some disconnected nodes which make the network bi-parted.

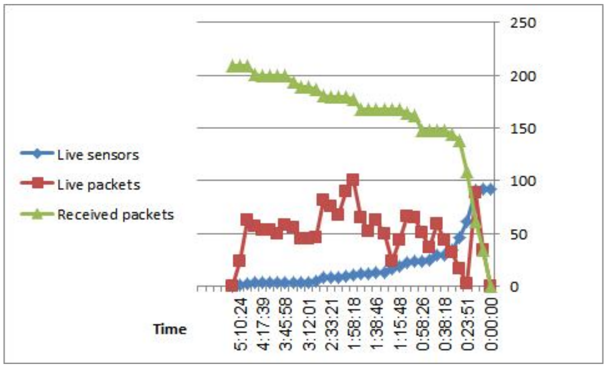


Fig.7: Comparison between number of available sensors, live and received packets existing in the network

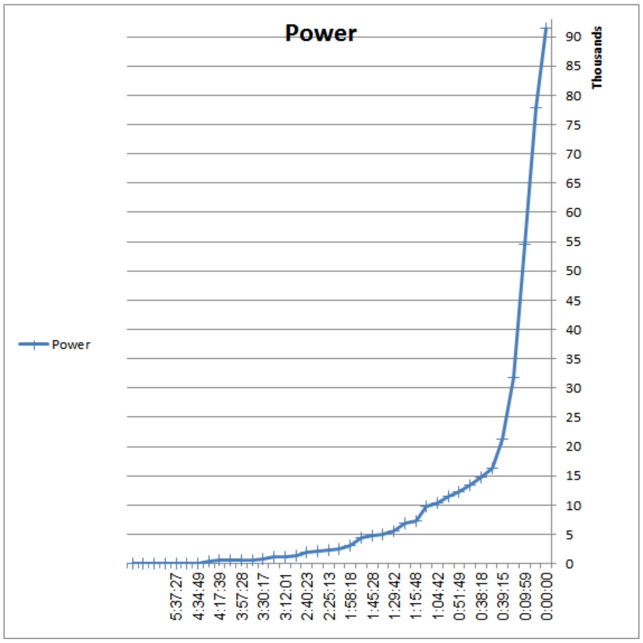


Fig.8: Comparison between mount of power and lifetime of network

Figures 7 and 8 show the average experiment associated with 200 packets. In Fig. 8 because of the heavy packets, the energy power of network is become low at time 0:39:15, while the transferring of packet data has an incremental rate and continues until 5:10:24 in Fig. 7. It shows that the node placement is optimally achieved such that even with sending heavy packets, the network optimally tolerates. Even when a node becomes dead, another near node can almost transmit the data so that finally the packets are received to the sink. Therefore, the proposed algorithm is an optimal one and extends the network lifetime.

V. CONCLUSION AND FUTURE WORK

In this paper, we present a fitness function according to the mount of covered area of the network to obtain an optimum solution where more area can be covered in spite of efficient energy consumption. These objectives result in suitable lifetime of the network. For a future work, we will try to achieve an algorithm which can select the nodes and the points for covering.

REFERENCES

[1] V. Mhatre, C. Rosenberg, D. Kofman, R. Mazumdar, and N. Shroff, “A minimum cost heterogeneous sensor network with a lifetime constraint,” IEEE Transactions on Mobile Computing (TMC), vol. 4, no. 1, pp. 4 –15, 2005.

[2] N. Trigoni, Y. Yao, A. Demers, J. Gehrke, and R. Rajaramany, “Wave scheduling: Energy-efficient data dissemination for sensor networks,” in Proceedings of the International Workshop on Data Management for Sensor Networks (DMSN), in conjunction with the International Confernece on Very Large Data Bases (VLDB), August 2004.

[3] A Norouzi, Dabaggian M,Ustunag B.berk “An improved ELGossiping Data Distribution Technique with Emphasis on Reliability and Resource Constraints in Wireless Sensor Network,IEEE 2010.

[4] K. Chakrabarty, S. S. Iyengar, H. Qi and E. Cho, “Grid coverage for surveillance and target location in distributed sensor networks,” *IEEE Transactions on Computers*, vol. 51, pp. 1448-1453, Dec. 2002.

[5] N. Bulusu, J. Heidemann and D. Estrin, “Adaptative beacon placement,” *Proc. Int. Conf. on Distributed Computing Systems*, pp. 489-498, Apr. 2001.

[6] S. S. Dhillon, K. Chakrabarty and S. S. Iyengar, “Sensor placement for grid coverage under imprecise detections,” *Proc. Int. Conf. on Information Fusion*, vol. 2, pp. 1581-1587, Jul. 2002.

[7] A. Howard, M. J. Mataric and G. S. Sukhatme, “An incremental selfdeployment algorithm for mobile sensor networks,” *Autonomous Robots Special Issue on Intelligent Embedded Systems*, vol. 13, no. 2, pp. 113- 126, 2002.

[8] A. Howard, M. J. Mataric and G. S. Sukhatme, “Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem,” *Proc. Int. Conf. on Distributed Autonomous Robotic Systems*, pp. 299-308, 2002.

[9] Y. Zou and K. Chakrabarty, “Sensor deployment and target localization based on virtual forces,” *Proc. IEEE Infocom Conference*, vol. 2, pp. 1293-1303, Apr. 2003.

[10] Norouzi Ali,Faeze Sadat Babamir, Berk Burak Ustundağ **“**A Genetic Algorithm based Approach Focused on Maintenance Methodology in Wireless Sensor Network”Elsevier,2010

[11] O.Younis, M. Krunz, and S.Ramasubramanian ,Node Clustering in Wireless Sensor Networks: Recent Developments and Deployment Challenges, DOI: 0890-8044/06 ,IEEE 2006

SESSION

WIRELESS NETWORKS

Chair(s)

TBA

Jointly Optimal Congestion Control, Network Coding and Power Control for QoS Multicast over DiffServ MAI-affected Wireless Networks

Enzo Baccarelli, Nicola Cordeschi and Valentina Polli
Dpt. of Information Engineering, Electronic and Telecommunication, "Sapienza" University of Rome, Italy

Abstract—*Recent advancements in network coding have shown great potential for efficient information multicasting in wireless packet networks in terms of both throughput and robustness. In this paper, we address the jointly optimal congestion control, network coding and self-adaptive distributed power control for DiffServ-based wireless networks. The target is to provide Quality of Service (QoS) support to multiple multicast multimedia sessions in the presence of Multiple Access Interference (MAI). To cope with the nonconvex nature of the addressed cross-layer optimization problem, we develop a two-level decomposition that is able to find the optimal solution by means of a suitable relaxed convex version of its comprising subproblems. Sufficient conditions for the equivalence of the primary (nonconvex) problem and its related (convex) version are provided, and a distributed, iterative algorithm for computing the solution of the resource allocation problem is developed. Actual performance and robustness against node-failures are numerically tested and compared with the ones of Dense Mode Protocol Independent Multicast (DM-PIM) routing algorithms.*

Keywords: QoS wireless multicast, cross-layer optimization, intra-session network coding, distributed power control.

1. Introduction

Network Coding (NC) extends the functionality of interior network nodes from simple storing/forwarding of packets to performing (in principle, arbitrary) of algebraic operation on the received payload data [1]. Starting from the seminal work in [2], several potential benefits of NC have been envisioned, including maximization of multicast throughput [2]–[5], robustness to link/node failures and/or packet losses [1], [6]. Lately, distributed random linear network coding schemes [7] have made actual implementation of NC feasible in the wireless domain. Hence, by referring to the emerging Next-Generation of wireless connectionless DiffServ networking architectures for the support of QoS-demanding multicast multimedia sessions, in this contribution we focus on jointly optimal congestion control, intra-session NC and power control when MAI cannot be canceled via suitable MAC scheduling, so that the resulting network wide cross-layer optimization problem is intrinsically *nonconvex*. By fact, most published work involving network coding in the wireless domain has been developed by focusing on cross-layer optimization [8], [9] and has given rise to a variety of contributions which differ mainly in how they deal with MAI. When orthogonal scheduling is allowed at the MAC layer (as in orthogonal Time and/or Frequency Division Multiple Access networks [3], [4], [10]), or link capacities and/or flow rates are directly assumed to belong to convex resource sets [11]–[13], cross-layer optimization problems are instances of convex optimization and can be readily solved.

On the contrary, when MAI effects cannot be removed through suitable MAC policies, the resulting resource allocation problems cannot, in general, be reduced to convex equivalent ones by means of either hidden convexity properties or particular algebraic transformations of the involved variables ([9], [14]). Nevertheless, there have been several attempts to develop manageable *approximations* of such nonconvex problems. In [15], a QoS power allocation problem for CDMA-based networks is proved to be convex if, and only if, the Signal to Interference plus Noise Ratio (SINR) can be expressed as a log-convex function of the QoS parameters. Recently, in [16], the authors derive conditions for the capacity function that are able to convexify the tackled joint power control, network coding and congestion control problem. Although conveniently solvable by common optimization tools, such convex *approximations* present limited application ranges, mainly because, due to the basic assumptions introduced in [15], [16], low SINRs may give rise to *negative* link-capacity values.

This considerations provide a strong motivation to investigate the possibility to compute the *exact* (i.e., *nonapproximate*) solution of MAI-affected nonconvex network-wide resource allocation problems by solving suitably designed *convex* problems. To this end, in this work we embed session utility, flow-control, QoS intra-session network coding, MAC design and power control in a wide cross-layer problem, named the Primary Optimization Problem (POP). Hence, by leveraging on the POP's structural properties, we develop a two-level solution that combines the performance advantages claimed by the cross-layer approach with the convenience of an optimization-driven decomposition [9], and that is able to give the *exact* solution of the nonconvex POP under a suitable set of conditions we provide in Sect.4. Moreover, the proposed decomposition leads to a distributed asynchronous self-adaptive iterative jointly optimal congestion control, network coding and power control algorithm, which demands *limited* information exchange among *neighbouring* nodes, and exhibits good convergence properties.

The remainder of this paper is organized as follows. In Sect.2, we describe the considered DiffServ multicast MAI-affected power-limited networking scenario. Sect.3 focuses on the analytical formulation and the constraint description of the POP. The proposed two-level decomposition solution and its structural properties are detailed in Sect.4. The development of the distributed resource allocation algorithm is shown in Sect.5, whereas numerical results and conclusion are provided in the final Sect.6.

About the adopted notation, $\mathbf{A} \equiv [a(v,l),\, v = 1,\ldots,V;\, l = 1,\ldots,L]$ indicates a $(V \times L)$ matrix with the (v,l)-th entry equal to $a(v,l)$, $\overrightarrow{e_i}$ is the i-th unit vector of $\mathbb{R}^V$, $\log(\cdot)$ is the natural logarithm, $f^{-1}(y)$ is the inverse of the scalar function $y \equiv f(\cdot)$, and A denoted the cardinality of set $\mathcal{A}$.

2. Wireless Network Model

The considered wireless network is described by a directed graph $\mathcal{G} \equiv (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{L}$ is the set of feasible links. A directed link l going from the transmit node $t(l)$ to the receive one $r(l)$ is feasible if the gain $g(t(l), r(l))$ of the corresponding physical channel is *strictly* positive, e.g., when the receive node $r(l)$ falls within the transmission range of $t(l)$. Let $\mathbf{A} \equiv [a(v,l)]$ be the $(V \times L)$ node-link incidence matrix that describes the feasible topology[1] of the network graph $\mathcal{G}$, that is,

$$a(v,l) \triangleq \begin{cases} 1, & \text{if node } v = t(l), \\ -1, & \text{if node } v = r(l), \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

and let $\mathbf{A}_s \equiv [a_s(v,l)] \triangleq \max\{\mathbf{A}, \mathbf{O}_{V \times L}\}$ be the corresponding source matrix. We rely on a network fluid model [17], where $F \geq 1$ multicast sessions, each one identified by the source/flow/destination-set triplet: $(s_i \in \mathcal{V}, f_i \in \mathbb{R}_0^+, \mathcal{D}_i \subseteq \mathcal{V}, i = 1, \ldots F)$, distribute their traffic flows across multiple paths. Furthermore, we define the overall sink set as $\mathcal{D} \equiv \bigcup_{i=1}^{F} \mathcal{D}_i$. To each flow f_i (measured in Information Unit per second (IU/s)) corresponds a link-flow vector $\overrightarrow{x_i}$, whose l-th component, $x_i(l)$, indicates the flow portion carried by the l-th link, so that this last is loaded with a total flow $x_T(l) \equiv \sum_{i=1}^{F} x_i(l)$. Furthermore, as in [4], [8], [12], [13], [16], intra-session network coding is considered as a viable means to improve network efficiency and, as a consequence, we have that [3]

$$x_i(l) = \max_{j=1,\ldots,D_i} \{x_{ij}(l)\} \tag{2}$$

where $x_{ij}(l)$, named the j-th subsession flow, is the part of $x_i(l)$ intended for destination $d_j \in \mathcal{D}_i$.

In accordance with the DiffServ architecture, we assume each session to belong to a different service class, which demands for specific QoS requirements and priority levels. Without loss of generality, we label the multicast sessions with increasing IDentity numbers (IDs) that correspond to decreasing priority levels, so as to assign smaller IDs to the sessions requiring higher priority. Due to the presence of network coding and multiple service classes, each output port of every interior node is equipped with F intra-session encoders, F parallel queues (one for each QoS-level) and a single server, which statistically multiplexes the outgoing flows according to an assigned priority-based discipline [17].

Hence, since the i-th session is handled in accordance with the i-th QoS class, the delay function $\Delta_i(C, x_1, \ldots x_F)$ adopted to measure the average queue-plus-transmission delay induced by each single-hop depends on the session-ID i, the link capacity C, and on *all* the conveyed traffic flows $\{x_1, \ldots x_F\}$. Hence, we assume that the following (mild) basic properties are retained by each $\Delta_i(\cdot)$: *i)* it is continuous with respect to its $F+1$ variables; *ii)* for any assigned set of variables $\{C, x_1, \ldots x_F\}$, it is increasing in the session-ID i, so that the average delay introduced by the link increases for increasing session-ID; *iii)* for any assigned i and $\{x_1, \ldots x_F\}$, it is strictly decreasing in C; *iv)* for any assigned i and C, it is nondecreasing in $\{x_1, \ldots x_F\}$; finally, *v)* for any assigned i, $\Delta_i(\cdot)$ is *jointly convex* in the $F+1$ variables $(C, x_1, \ldots x_F)$. Due to the Kleinrock's independence condition and Jackson's Theorem, these (mild) assumptions may be reasonably considered met in the considered connectionless networking scenario, where each source-destination coded route may be modeled as the cascade of several queueing systems, whose corresponding input traffics are the aggregation of multiple flows belonging to different paths [17].

Due to the possible nomadic behaviour of the networking nodes, we assume that each feasible link acts as a block-fading channel, whose gain may be periodically measured by the receive node. Besides fading, topological and MAC-related parameters, as well as other system depending parameters (e.g., cross-correlation properties of the utilized access codes, beamforming coefficients, etc.), affect the connection between two nodes. To capture their comprehensive effect, we define $\mathbf{G} \triangleq [g(k,l)]$ as the $(L \times L)$ matrix that collects the (nonnegative) gains between each transmit-receive pair:

$$g(k,l) \triangleq g\left(t(k), r(l)\right), \quad k, l = 1, 2, \ \ldots L\,.$$

Entries along the main diagonal of $\mathbf{G}$ (i.e., the coefficients $\{g(k,k)\}$) describe the gains of the feasible links, whereas the remaining (possibly, nonzero) entries $\{g(k,l),\, k \neq l\}$ are MAI coefficients that measure the interference among different links. Thus, for each link $l \in L$ with transmit power $P(l)$ (W), we can express the corresponding SINR(l) measured at the receive node $r(l)$

[1] We explicitly remark that such matrix serves the only purpose to capture the feasible network connectivity, whereas the actual topology of the network, i.e., the activated links with their relative capacities, will be the final solution of the considered POP of Sect.3.

as

$$\mathrm{SINR}(l) \equiv \frac{\Gamma(l)\, g(l,l)\, P(l)}{\sum_{k=1,\, k\neq l}^{L} g(k,l)P(k) + N(l)}\,, \tag{3}$$

where $\Gamma(l) > 0$ is the so-called SINR-gap accounting for the target Bit Error Rate (BER), and the denominator in (3) is the corresponding receive noise $N(l)$ (W) plus MAI power. In our framework, the function $\Psi_l(\mathrm{SINR}(l))$ that measures the capacity $C(l)$ (IU/s) of the l-link, is assumed *nonnegative*, *continuous* and *strictly increasing* for $\mathrm{SINR}(l) \geq 0$, with $\Psi_l(0) \equiv 0$. We underline that *none* of the convexity and/or log-convexity assumptions advanced by previous works on power-control of MAI-affected networks (see [8], [10], [13], [16]) on the capacity function is here introduced.

All the mentioned per-link parameters may be gathered in the following $(L \times 1)$ column vectors: $\overrightarrow{x_T}$ (total flow), $\overrightarrow{x_{ij}}$ (subsession flow), $\overrightarrow{\mathrm{SINR}}$ (SINR), $\overrightarrow{\Gamma}$ (SINR-gap), $\overrightarrow{C}$ (capacity) and $\overrightarrow{P}$ (power).

3. The Multicast Primary Optimization Problem (POP)

Let $\overrightarrow{f} \equiv [f_1, \dots, f_F]$ (IU/s) be the vector collecting the F multicast flows forwarded from each source node $s_i \in \mathcal{V}$ to its corresponding destination nodes $\{d_j \in \mathcal{D}_i\}$. Thus, the Primary Optimization Problem (POP) we introduce is a wide, generally nonconvex, cross-layer problem. Its ultimate goal is to compute the set of network variables $\{\overrightarrow{f}, \overrightarrow{x_1}, \dots, \overrightarrow{x_F}, \overrightarrow{P}, \overrightarrow{\Gamma}, \mathbf{G}\}$ which minimizes a given network cost function $\Phi(\cdot)$, while meeting a suitable set of session-dependent constraints arising from the Application, Transport, Network, MAC and Physical Layers. Specifically, the POP we consider is formally stated as follows:

$$\min_{\overrightarrow{f}, \overrightarrow{x_1}, \dots, \overrightarrow{x_F}, \overrightarrow{P}, \overrightarrow{\Gamma}, \mathbf{G}} \Phi\left(\overrightarrow{f}, \overrightarrow{x_1}, \dots, \overrightarrow{x_F}, \overrightarrow{C}\right) \tag{4.1}$$

$$\text{s.t.: } \mathbf{A}\overrightarrow{x_{ij}} - f_i(\overrightarrow{e_{s_i}} - \overrightarrow{e_{d_j}}) = \overrightarrow{0_V}, j = 1, ...D_i, i = 1, ..., F, \tag{4.2}$$

$$x_T(l) - \eta(l)C(l) \leq 0,\ l = 1, ..., L, \tag{4.3}$$

$$x_{ij}(l) - Div(i)f_i \leq 0, l = 1, ..., L, j = 1, ..., D_i, i = 1..., F, \tag{4.4}$$

$$C(l) - C_{max}(l) \leq 0,\ l = 1, ..., L, \tag{4.5}$$

$$\sum_{l=1}^{L} \varepsilon(l)C(l) - C_{ave} \leq 0, \tag{4.6}$$

$$\sum_{l=1}^{L} J_i\Big(C(l), x_i(l)\Big) - H_t(i) \leq 0,\ i = 1, ..., F, \tag{4.7}$$

$$B_{min}(i) - f_i \leq 0,\ i = 1, ..., F, \tag{4.8}$$

$$\sum_{l=1}^{L} \Delta_i\Big(C(l), x_1(l), .., x_F(l)\Big) - \nabla_t(i) \leq 0,\ i = 1, ..., F, \tag{4.9}$$

$$D_i(f_i, \sum_{l=1}^{L} \Delta_i(C(l), x_1(l), .., x_F(l))) - \sigma_D^2(i) \leq 0,\ i = 1, ..., F, \tag{4.10}$$

$$f_i, x_{ij}(l) \geq 0,\ j = 1, ..., D_i, l = 1, ..., L\ ,\ i = 1, ..., F, \tag{4.11}$$

$$\Gamma(l) - \Gamma_{max}(l) \leq 0,\ l = 1, ..., L, \tag{4.12}$$

$$g(l,l) - G_{max}(l) \leq 0,\ l = 1, ..., L, \tag{4.13}$$

$$-G_{min}(k,l) + g(k,l) \leq 0,\ l, k = 1, ..., L,\ k \neq l, \tag{4.14}$$

$$\sum_{l=1}^{L} a_s(v,l)P(l) - P_{max}(v) \leq 0,\ v \notin \mathcal{D}, \tag{4.15}$$

$$g(k,l), P(l), \Gamma(l) \geq 0,\ l, k = 1, ..., L. \tag{4.16}$$

Delving into the reported POP constraints, we note that, in addition to the usual flow conservation laws in (4.2) (which, due to the presence of network coding, apply to each subsession, i.e., to each single source-destination pair [3]), the flow vectors $\overrightarrow{x_{ij}}$ and $\overrightarrow{x_T}$ have to comply with the constraints in (4.3)-(4.4) that upper limit link utilization. The reason for such bounds is manifold. First, a proper tuning of the link utilization factor $\eta(l)$ guarantees the existence of a primary conflict-free scheduling policy that may avoid self-interference over the network-coded paths (see [18] for details). Second, since the f_i's are only *average* measures of the forwarded flows, setting a working condition of $\eta(l) < 1$ may prevent exceeding capacity events due to traffic-volume fluctuations. Third, the i-th link diversity factor $Div(i) \in (0, 1]$ controls the *minimum* number of distinct paths to be employed by the i-th source to each destination $d_j \in \mathcal{D}_i$: when $Div(i) < 1$, every $\overrightarrow{x_{ij}}$ is strictly multipath, so that it provides improved reliability and exhibits failure-tolerant properties (see the numerical results of Sect.6).

Constraints in (4.5)-(4.6) may arise from economical restrictions applied by the Network Administrator on the capacity planning of the links [4]. These constraints fix a maximum link-capacity $C_{max}(l)$ as well as a maximum average network capacity cost C_{ave} when the price-rate of $C(l)$ is set to $\varepsilon(l)$. Similarly, the (convex) function $J_i(C(l), x_i(l))$ in (4.7) measures the cost to route the i-th session over the l-th link and may be used to build up suitable session-dependent overlay networks on top of the assigned graph $\mathcal{G}$. Per-session QoS requirements are detailed in (4.8)-(4.12). Specifically, in addition to minimum connection bandwidth $B_{min}(i)$ (IU/s) and maximum delay $\nabla_t(i)$, we also account for a constraint on the maximum (average) distortion $\sigma_D^2(i)$ tolerated by the end-users joining the i-th session. This bound is media-application specific: as pointed out in [19], subjective QoS may be measured by a proper, convex, distortion function $D_i(\cdot, \cdot)$ that depends on both i-th bandwidth and delay. At the MAC Layer, we upper limit the achievable gains of the feasible network links (4.13) and lower limit the minimum allowed MAI coefficients (4.14). As a consequence, only when all the G_{min} MAI coefficients in (4.14) vanish, orthogonal access is, indeed, *feasible* for the considered POP. A maximum per-link BER is set through the corresponding maximum gap $\Gamma_{max}(l)$ in (4.12). Finally, at the Physical Layer, we impose a maximum power budget per transmit node (4.15), while constraints (4.11), (4.16) assure the nonnegativity of all variables.

Formally, as in [10], [11], the objective $\Phi(\cdot)$ function in (4.1) is a real-valued, jointly convex function of the link-capacities $\overrightarrow{C}$, end-to-end forwarded flows $\overrightarrow{f}$ and the link-flows $\overrightarrow{x_i}$, continuously differentiable up to second order. Since the nondifferentiability of the maximum function in (2) may affect the differentiability of $\Phi(\cdot)$ (and, likewise, of all the POP's constraints involving x_i), we replace (2) with the upperbound given by the corresponding $\mathcal{L}^n$-norm (see [16]):

$$x_i(l) \equiv \max_{j=1,\ldots D_i} x_{ij}(l) \leq \Big(\sum_j (x_{ij}(l))^n \Big)^{1/n} , \qquad (5)$$

which converges to (2) for large[2] n, while preserving convexity. The objective function in (4.1) may be used to capture user satisfaction (e.g., flow maximization), network operator's goals (e.g., efficient resource allocation, load-balancing and fairness) or a proper trade-off of both [9].

3.1 Unicast, Multicast and Multiple Unicast applications

The formulation of the POP in (4) refers to the general case of a multiple multicast problem with intra-session network coding and multiple QoS classes. Depending on the actual number of sources/destinations and, most importantly, sessions, the POP directly adapts to unicast, multiple unicast and multicast (with/without NC) scenarios. Application of the POP to unicast and single-session (coded) multicast is straightforward, since they can be obtained by directly setting $D_i = F = 1$ and $F = 1$, respectively. Routing-based multicast and multiple unicast without NC can be described by replacing the expression in (2) with the following one:

$$x_i(l) = \sum_{j=1}^{D_i} x_{ij}(l) , \qquad (6)$$

since in the routing case, each $x_{ij}(l)$ corresponds to an independent flow (see [3]).

4. The proposed solving approach

The multicast POP in (4) is, generally, a *nonconvex* optimization problem, despite the fact that its constraints are linear or convex. This is a direct consequence of the *nonconvexity* of the relationship tying powers and link-capacities (see (3) and following text). These latter are not actually part of the POP variables set, but act as *coupling* parameters between Transport/Network $\{\overrightarrow{f}, \overrightarrow{x_1}, \ldots, \overrightarrow{x_F}\}$ and MAC/Physical $\{\overrightarrow{P}, \overrightarrow{\Gamma}, \mathbf{G}\}$ variables via the corresponding feasible capacity region $\mathcal{C}$. Therefore, *neither* guaranteed-convergence iterative algorithms *nor* closed-form solutions are available to evaluate the optimum $\{\overrightarrow{f}^*, \overrightarrow{x_1}^*, \ldots, \overrightarrow{x_F}^*, \overrightarrow{P}^*, \overrightarrow{\Gamma}^*, \mathbf{G}^*\}$ of the resulting MAI-affected nonconvex POP. However, by leveraging on the coupling role of link-capacities, we propose a two-level decomposition where: *i)* the upper level tackles the Flow Network Coding Problem (FNCP), to find the optimal link-capacity vector $\overrightarrow{C}^*$; and, *ii)* the lower level computes the capacity bounds for the FNCP on the basis of the Physical/MAC constraints and, then, it solves the corresponding Efficient Resource Allocation Problem (ERAP). This last aims at computing the capacity target vector $\overrightarrow{C}^*$, while minimizing resource consumption.

To formally define the abovementioned feasible capacity region $\mathcal{C}$ of the multicast POP, let

$$\Pi \triangleq \left\{ \left(\overrightarrow{P}, \overrightarrow{\Gamma}, \mathbf{G} \right) : \text{(4.12)-(4.16) simultaneously met} \right\}, \qquad (7)$$

be the convex region of the $(L^2 + 2L)$-dimensional Euclidean space comprising all the triplets $\left(\overrightarrow{P}, \overrightarrow{\Gamma}, \mathbf{G}\right)$ meeting the MAC and Physical layers constraints. Furthermore, let

$$S \triangleq \left\{ \overrightarrow{\text{SINR}} \triangleq [\text{SINR}(1), \ldots, \text{SINR}(l)]^T \right\}, \qquad (8)$$

be the related L-dimensional set of *feasible* SINR vectors obtained by a componentwise application of the scalar expression in (3) to the elements of the set Π in (7). The resulting POP capacity region $\mathcal{C}$ can be now formally defined as

$$\mathcal{C} \triangleq \left\{ \overrightarrow{C} \in \left(\mathbb{R}_0^+\right)^L : \exists\, \overrightarrow{\text{SINR}} \in S \,:\, C(l) \leq \Psi_l(\text{SINR}(l)),\ \forall\, l \right\}. \qquad (9)$$

On the basis of (9), the FNCP is an optimization problem in the $\{\overrightarrow{f}, \overrightarrow{x_1}, \ldots, \overrightarrow{x_F}, \overrightarrow{C}\}$ variables, so formulated:

$$\min_{\overrightarrow{f}, \overrightarrow{x_1}, \ldots, \overrightarrow{x_F}, \overrightarrow{C}} \Phi\left(\overrightarrow{f}, \overrightarrow{x_1}, \ldots, \overrightarrow{x_F}, \overrightarrow{C}\right), \qquad (10.1)$$

$$\text{s.t. : POP constraints in (4.2)–(4.11)}, \qquad (10.2)$$

$$\overrightarrow{C} \in \mathcal{C}. \qquad (10.3)$$

We define the corresponding ERA problem as

$$\min_{\overrightarrow{P}, \mathbf{G}, \overrightarrow{\Gamma}} \varphi(\overrightarrow{P}, \mathbf{G}), \qquad (11.1)$$

$$\text{s.t. : POP constraints in (4.12)–(4.16)}, \qquad (11.2)$$

$$\text{SINR}^*(l)/\text{SINR}(l) - 1 \leq 0, \quad l = 1, \ldots, L, \qquad (11.3)$$

where $\text{SINR}^*(l) \triangleq \Psi_l^{-1}(C^*(l))$, and $C^*(l) \in \overrightarrow{C}^*$ is the capacity value of link l, that is solution of the FNCP in (10). Furthermore, $\varphi(\overrightarrow{P}, \mathbf{G})$ in (11.1) is a cost function introduced in order to pick up the most resource-efficient allocation when the feasibility problem defined by the constraints (11.2)-(11.3) allows multiple solutions. The aboveintroduced subproblems are coupled by the optimal capacity vector $\overrightarrow{C}^*$ and the feasible capacity region $\mathcal{C}$,

[2] We have numerically ascertained that $n = 10$ suffices to guarantee a final accuracy within 1%.

and retain the two following basic properties (whose proof can be found in [20]).

Proposition 1: When $\varphi(\overrightarrow{P}, \mathbf{G})$ is posynomial in $\{\overrightarrow{P}, \mathbf{G}\}$, the *ERAP* becomes an instance of *geometric programming* and, therefore, solvable by convex optimization. □

Proposition 2: Let us assume $\mathcal{C}$ in (10.3) be defined as in (9). Thus, the multicast POP in (4) admits the same solution of the combined FNC-plus-ERA problem in (10)-(11). □

A direct consequence of *Proposition 2* is that, whenever the feasible capacity region $\mathcal{C}$ in (9) is a convex set, both the combined FNC-plus-ERA problem and the POP are convex and their optima coincide. Unfortunately, this condition is met *only* for log-convex capacity functions (see [15], [16]), or when the POP allows orthogonal access (i.e., when all $\{G_{min}(k,l)\}$ in (4.14) vanish). Nonetheless, in [20] we prove that the proposed two-level decomposition retains the following *fundamental* properties, that make it valuable for the solution of nonconvex POPs.

Proposition 3: Let us consider a convex outer-bound $\mathcal{C}_0$ of the feasible capacity region $\mathcal{C}$, i.e., $\{\mathcal{C} \subseteq \mathcal{C}_0\}$, and let $\overrightarrow{C_0}^*$ be the link-capacity vector obtained by solving the resulting $\mathcal{C}_0$-relaxed *FNCP*[3]. Thus, the following properties hold:

1. when the $\mathcal{C}_0$-relaxed *FNCP* is unfeasible, then also the *POP* is unfeasible;
2. when the $\mathcal{C}_0$-relaxed *FNCP* is feasible but the *ERAP* is unfeasible (i.e., $\overrightarrow{C_0}^* \notin \mathcal{C}$), then *no* conclusion may be drawn about the feasibility/unfeasibility of the POP;
3. when the $\mathcal{C}_0$-relaxed *FNCP* and the *ERAP* are both feasible (i.e., $\overrightarrow{C_0}^* \in \mathcal{C}$), then the POP is feasible and, most importantly, $\overrightarrow{C}^* \equiv \overrightarrow{C_0}^*$. □

Proposition 3 provides both *formal* and *practical* support to the proposed two-level decomposition, by guaranteeing that it is, indeed, possible to solve the *nonconvex* multicast POP by means of two coupled *convex* problems. Such capability is, however, dependent on the occurrence of Case 3 of *Proposition 3*, which, in practice, is higher when $\mathcal{C}_0$ is tighter to $\mathcal{C}$. Since tight outer bounds are generally very complex to be characterized and their evaluation may lead to NP-hard problems [4], in practice, we are interested in the $\mathcal{C}_0$ leading to the best accuracy-vs.-complexity tradeoff (see Sect.6).

As pointed out in *Proposition 3*, feasibility of the ERAP guarantees the feasibility of the POP and, moreover, the *coincidence* of its solution to that of the FNC-plus-ERA problem. Therefore, the following condition (proved in [20] and derived for the feasibility of the ERAP) acts as *sufficient* condition for both the POP feasibility and the solutions equivalence: $\overrightarrow{C}^* = \overrightarrow{C_0}^*$.

Proposition 4: Let $\mathbf{J}$ be the $(L \times L)$ matrix whose (k,l)-th entry is defined as

$$J(k,l) \triangleq \begin{cases} -1, & k = l, \\ \frac{G_{min}(k,l)\Psi_k^{-1}(C_0^*(k))}{G_{max}(k)\Gamma_{max}(k)}, & k \neq l. \end{cases}$$

Thus, *if and only if* there exists a $((V+L) \times 1)$ nonnegative vector $\overrightarrow{\beta}$ that satisfies the following (matrix) equation:

$$\begin{bmatrix} \mathbf{J} \\ \mathbf{A_s} \end{bmatrix}^T \overrightarrow{\beta} + \overrightarrow{1}_L = \overrightarrow{0}_L, \qquad (12)$$

the *ERAP* is feasible. □

5. The Distributed Self-Adaptive Resource Allocation Algorithm

In principle, being convex optimization problems, under the Slater's qualification conditions, both the $\mathcal{C}_0$-relaxed FNCP and the ERAP can be solved via the Karush-Kuhn-Tucker (KKT) optimality conditions. However, the particular structure of the ERAP makes the solution of its dual problem conveniently suitable for distributed implementations. In detail, the Lagrangian function associated to (11), when $\varphi(\overrightarrow{P}, \mathbf{G}) \equiv \sum_{l=1}^{L} P(l)$, can be expressed as [20]:

$$\begin{aligned} L(\overrightarrow{z}, \overrightarrow{y}, \mathbf{W}, \overrightarrow{\lambda}) = & \sum_{l=1}^{L} e^{z_l} + \\ & + \sum_{l=1}^{L} \lambda_{1l} \Big(\mathrm{SINR}^*(l)\, e^{-z_l - y_l - w_{ll}} \Big[\sum_{k \neq l}^{L} e^{z_k + w_{kl}} + N(l) \Big] - 1 \Big) + \\ & + \sum_{v=1}^{V} \lambda_{2v} \Big(P_{max}(v)^{-1} \sum_{l=1}^{L} a_s(v,l) e^{z_l} - 1 \Big) + \\ & + \sum_{l=1}^{L} \lambda_{3l} \Big(e^{y_l} - \Gamma_{max}(l) \Big) + \sum_{l=1}^{L} \lambda_{4l} \Big(e^{w_{ll}} - G_{max}(l) \Big) + \\ & + \sum_{l=1}^{L} \sum_{k \neq l}^{L} \lambda_{5kl} \Big(e^{-w_{kl}} + G_{min}(k,l) \Big), \end{aligned} \qquad (13)$$

where $z_l \triangleq \log(P(l))$, $y_l \triangleq \log(\Gamma(l))$, $\mathbf{W}(k,l) \triangleq [w_{k\,l} = \log(g(k,l))]$, and $\overrightarrow{\lambda} \triangleq [\overrightarrow{\lambda}_1 \overrightarrow{\lambda}_2 \overrightarrow{\lambda}_3 \overrightarrow{\lambda}_4 \overrightarrow{\lambda}_5]^T$ is the vector collecting all the Lagrangian multipliers associated to the constraints in (11.2)-(11.3).

From the analysis of (13) (see [20] for details), it results that all variables and multipliers are *local quantities*, so that the ERAP may be solved through an iterative *distributed* asynchronous algorithm that requires only *per-link* measurement/processing and limited message passing among *neighbouring* nodes. Furthermore, such algorithm simply relies on gradient-based updates for each variable u, according to [20]:

$$u^{(k+1)} = u^{(k)} - a^{(k)} \nabla_u L\left(\overrightarrow{z}^{(k)}, \overrightarrow{y}^{(k)}, \mathbf{W}^{(k)}, \overrightarrow{\lambda}^{(k)} \right), \qquad (14)$$

where $k = 0, 1, \ldots$ is a discrete iteration index, whereas $\{a^{(k)}\}$ is a suitable step-size sequence we may adaptively tune according to the relationship developed, for example, in [21].

[3] This problem is obtained by replacing $\mathcal{C}$ in (10.3) with its outer bound $\mathcal{C}_0$.

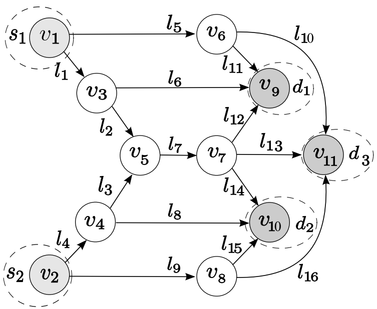


Fig. 1: Butterfly network with two active multicast sessions sharing the same QoS requirements and destination sets [8].

Table 1: Main Simulation Parameters

$N = 0.01$mW	$P_{max} = 2$mW	$\sigma_D^2 = 0.2$	$\Gamma_{max} = 0.5$
$G_{min} = 10^{-2}$	$Div = 1$	$\eta = 0.8$	$\nabla_t = 30\mu$s
$C_{ave} = 50$Mb/s	$\varepsilon = 1$	$G_{max} = 1$	$C_{max} = 4$Mb/s

Table 2: Example of practical relevance of *Proposition 3*

Case A	$\mathcal{C}_0$-FNCP unfeasible	POP unfeasible
Case B	MAI-free	POP feasible
Case C	medium MAI	POP feasible
Case D	high MAI	POP undetermined

6. Numerical Results and Conclusion

In this section, we test the performance of the proposed resource allocation algorithm with respect to conventional IP multicast routing algorithms, and, then, show how it is able to self-adapt to node failure events. We consider the butterfly topology in Fig.1 [8], where two sources (located at nodes v_1, v_2) have to send information to a common destination set $\mathcal{D} \equiv \{d_1, d_2, d_3\}$ with the same QoS requirements.

In the carried out numerical tests, we adopt the Shannon-Hartley's logarithmic formula: $C(l) \equiv B \log_2(1 + \text{SINR}(l))$ (Mb/s) to measure the capacity of the l-th link, with bandwidth $B \equiv 1$ MHz. This function successfully complies with all the previously reported assumptions on $\Psi_l(\cdot)$ and guarantees *nonnegative* capacities, even for *vanishing* SINRs. Furthermore, we consider per-session link-delays measured as in [17]: $\Delta(C, x_{_T}) \equiv ([C - x_{_T}]^{-1} + C^{-1})$ (s). As in [19], $D(f) = \exp(-f)$ is employed as distortion function in (4.10) and the total power consumption is selected as the ERAP cost function, i.e., $\varphi(\overrightarrow{P}, \mathbf{G}) \equiv \sum_l P(l)$.

In the carried out tests, we directly adopt the following simple-to-characterize box-type outer bound for the capacity region:

$$\mathcal{C}_0 \equiv \left\{ \overrightarrow{C} \in \left(\mathbb{R}_0^+\right)^L : \forall l = 1, \dots L, \right.$$
$$\left. 0 \leq C(l) \leq \Psi_l \left(\frac{\Gamma_{max}(l) G_{max}(l) P_{max}(t(l))}{N(l)} \right) \right\} . \tag{15}$$

Such bound coincides with the actual capacity region $\mathcal{C}$ for the MAI-free case and we anticipate that it has proven *sufficient* to guarantee the occurrence of Case 3 of *Proposition 3* in *all* the presented numerical tests. The basic simulation parameters are detailed in Table 1. Interfering links are the ones ending into a common receive node, and G_{min} in Table 1 indicates their minimum gain[4]. Nonuniform resource availability is considered: links indexed with 5, 6, 7, 8 and 9 in Fig.1 are assigned $C_{max} = 5$Mb/s. Furthermore, we set $f_1 = f_2 = f$ and define a common (total) bandwidth requirement for each destination $B_{min}(i) = 2$Mb/s, $i = 1, 2$.

About actual relevance of *Proposition 3*, Table 2 shows how variations in the system parameters in Table 1 impact on the three cases detailed in *Proposition 3*. Specifically, if a too demanding per-session QoS requirement (such as $\nabla_t = 5\mu$s) is advanced, the $\mathcal{C}_0$-FNCP results to be unfeasible (Case A), and so the POP (see *Proposition 3*.1). MAI-free operating conditions (Case B) and medium interference gains (Case C: $G_{min} \leq 0.04$) guarantee the feasibility of the POP and the *optimality* of the solution found by the proposed approach (see *Proposition 3.3*). Finally, when the interference gains grow beyond 0.04 (Case D), the bound $\mathcal{C}_0$ in (15) becomes too loose and, according to *Proposition 3.2*, prevents us from drawing any conclusion about the feasibility/unfeasiblity of the POP. In this regard, we point out that all the following reported numerical results refer to the POPs' *exact* solutions.

Let us analyze the flow distribution when: *i)* the objective in (10.1) is the maximization of the total flow arriving at each destination, i.e., $\Phi(\overrightarrow{f}, \overrightarrow{x_1} \dots, \overrightarrow{x_F}, \overrightarrow{C}) = -f$; and *ii)* the resulting POP is solved either with routing (i.e., as a multiple unicast problem) or with network coding (see Sect.3.1)[5]. The available paths to the destinations d_2, d_3 (d_1 is not shown since it is symmetrical to d_2), their nodes composition and the corresponding Routing-based (R) and Network Coding-based (NC) flows (in Mb/s) are detailed in the first four columns of Table 3. From the reported values, it is apparent the advantage of a network coded solution: the total conveyed flow $f_{_{NC}} = 7.25$Mb/s *more than doubles* the routing one $f_{_R}$, which is only equal to 3.20Mb/s. As shown in the last columns of Table 3, network coding is *still* beneficial when failures occurred at nodes v_6, v_8. However, since the failure makes the number of available path decrease, in this case the throughput gain due to network coding is limited to the 37%.

The network coding gain, as well as the convergence behaviour and the adaptivity to node failures of the overall distributed resource allocation algorithm of Sect.5, may be appreciated through the plots in Fig.2. These latter report the time-evolution (with respect to the iteration

[4] Orthogonal links have identically zero interfering gains.

[5] Having a common destination set allows to code together the flows of s_1 and s_2 and still be consistent with the assumption of intra-session network coding (see [2]).

Table 3: Path-flow distribution for destinations d_2, d_3

	path	composition	flows (v_6, v_8 on) R	NC	flows (v_6, v_8 off) R	NC
d_2	$\mathcal{P}_1$	$\{v_2, v_8, v_{10}\}$	0.95	2.40	0.00	0.00
	$\mathcal{P}_2$	$\{v_2, v_4, v_{10}\}$	0.58	2.40	1.00	1.61
	$\mathcal{P}_3$	$\{v_2, v_4, v_5, v_7, v_{10}\}$	0.07	0.00	0.00	0.37
	$\mathcal{P}_4$	$\{v_1, v_3, v_5, v_7, v_{10}\}$	1.60	2.45	1.00	1.22
d_3	$\mathcal{P}_5$	$\{v_2, v_8, v_{11}\}$	1.40	2.40	0.00	0.00
	$\mathcal{P}_6$	$\{v_2, v_4, v_5, v_7, v_{11}\}$	0.20	1.22	1.00	1.60
	$\mathcal{P}_7$	$\{v_1, v_6, v_{11}\}$	1.40	2.40	0.00	0.00
	$\mathcal{P}_8$	$\{v_1, v_3, v_5, v_7, v_{11}\}$	0.20	1.22	1.00	1.60
		total flow f	3.20	7.25	2.00	3.20

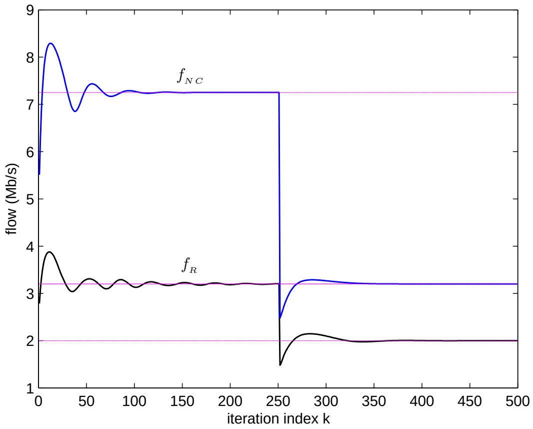


Fig. 2: Evolution of routing (f_R) and network coding (f_{NC}) total flows to destination d_2, in the presence of failures of v_6 and v_8 at $k = 250$.

index k) of the total flows f_{NC} and f_R to destination d_2. Good convergence to the optimal POP solutions (indicated by the horizontal lines in Fig.2) is achieved in about 50 iteration cycles, whereas quick reactivity to the failures (which occur at $k = 250$) is supported by the fact that the optimum is approached (with an error below 10%) within 20 iterations.

Conventional multicast based on the DM-PIM routing algorithm [22], that floods information across the minimum-hop distribution-tree, is considered for performance comparison. All simulation parameters have been kept unchanged, and both the proposed and the DM-PIM's algorithms are required to support the same flow f = 2Mb/s to each destination. In the case of the network in Fig.1, the minimum-hop distribution tree comprises of links $\{l_1, l_2, l_3, l_4, l_7, l_{12}, l_{13}, l_{14}\}$, and the corresponding total power consumption and maximum delay are equal to 2.88mW and 7.2μs, respectively. Whereas the DM-PIM's strategy optimizes resources by minimizing the number of involved links, the POP solution proves that, a more even flow-distribution across the network can save more than the 75% of power, while even gaining the 6% in delay with respect to the DM-PIM one. These performance gains support the actual effectiveness of the proposed distributed resource allocation algorithm with intra-session NC.

References

[1] T.Ho, D.S.Lun, *Network Coding - An Introduction*, Cambridge Press, 2008.

[2] R.Ahlswede, Ning Cai, S.Y.R.Li, R.W. Yeung, "Network information flow," *IEEE Tr. on Information Theory*, vol.46, no.4, pp.1204-1216, Jul. 2000.

[3] Y.Wu, P.A.Chou, S.Y.Kung, "Minimum-Energy Multicast in Mobile ad-hoc Networks using Network Coding", *IEEE Tr. on Communications*, vol.53, no.11, pp. 1906-1918, Nov. 2005.

[4] Y.Wu, P.A.Chou, Q.Zhang, K.Jain, W.Zhu, S.Y.Kung, "Network Planning in Wireless ad-hoc Networks: A Cross-Layer Approach", *IEEE J.Sel.Areas in Commun.*, vol.23, no.1, pp.136-150, Jan. 2005.

[5] A.Eryilmaz, D.S.Lun, "Control for Inter-session Network Coding", Proc. *IEEE NetCod*, Jan. 2007.

[6] R.Koetter, M.Medard, "An algebraic approach to network coding", *IEEE/ACM Tr. on Networking*, vol.11, no.5, pp.782-795, Oct. 2003.

[7] P.A.Chou, Y.Wu, K.Jain, "Practical network coding", Annual Allerton Conference on Communication, Control, and Computing, Oct. 2003.

[8] J.Yuan, Z.Li, W.Yu, B.Li, "A cross-layer optimization framework for multicast in multi-hop wireless networks", *International Conference on Wireless Internet*, WICON 2005, pp. 47-54, July 2005.

[9] M.Chiang, S.H.Low, A.R.Calderbank, J.C.Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, vol.95, no.1, pp.255-312, Jan. 2007.

[10] D.S.Lun, N.Ratnakar, R.Koetter, M.Medard, E.Ahmed, Lee Hyunjoo, "Achieving minimum-cost multicast: a decentralized approach based on network coding," Proceedings *IEEE INFOCOM 2005*, vol.3, pp. 1607- 1617 vol. 3, March 2005.

[11] Y.Wu, M.Chiang, S.Y.Kung, "Distributed Utility Maximization for Network Coding Based Multicasting: A Critical Cut Approach", Proc. *IEEE NetCod* 2006.

[12] Y.Xuan, C.T.Lea, "Network-Coding Multicast Networks With QoS Guarantees", *IEEE/ACM Tr. on Networking*, vol.19, no.1, pp.265-274, Feb. 2011.

[13] L.Clien, T.Ho, S.H.Low, M.Chiang, J.C.Doyle, "Optimization Based Rate Control for Multicast with Network Coding", Proceedings *IEEE INFOCOM 2007* pp.1163-1171, May 2007.

[14] C.W.Tan, D.Palomar, M.Chiang, "Exploiting hidden convexity for flexible and robust resource allocation in cellular networks", *IEEE INFOCOM 2007*.

[15] H.Boche, S.Stanczak, "On the convexity of feasible QoS regions," *IEEE Tr.on Information Theory*, vol.53, no.2, pp. 779-783, Feb. 2007.

[16] Xi Yufang, E.M.Yeh, "Distributed Algorithms for Minimum Cost Multicast With Network Coding", *IEEE/ACM Tr. on Networking*, vol.18, no.2, pp.379-392, April 2010.

[17] D.P.Bertsekas, R.Gallagher, *Data Networks*, Second Ed. Englewood Cliffs, NJ, Prentice Hall, 1992.

[18] B.Hajek, G.Sasaki, "Link scheduling in polynomial time," *IEEE Tr. on Information Theory*, vol.34, no.5, pp.910-917, Sep 1988.

[19] J.Yan, K.Katrinis, M.May, B.Plattner, "Media and TCP-friendly Congestion Control for scalable video streams", *IEEE Tr. on Multimedia*, vol.8, no.2, pp.196-206, Apr.2006.

[20] E.Baccarelli, N.Cordeschi, V.Polli, "Distributed self-adaptive QoS Traffic Engineering for MAI-affected Power-limited wireless networks", available at http://infocom.uniroma1.it/~enzobac/technicalreport.pdf

[21] H.J.Kushner, J.Yang, "Analysis of adaptive step-size SA algorithms for parameter tracking", *IEEE Tr. on Automatic Control*, vol.40, no. 8, pp 1403-1410, Aug.1995.

[22] S.Deering, D.Estrin et al., "The PIM architecture for Wide area multicasting", *IEEE/ACM Tr.on Networking*, vol.4,no.2, pp.153-162, Apr.1996.

A Practical Perspective of Wireless Network Coding

Marium Jalal Chaudhry[1,2], Timo Hämäläinen[1], Jyrki Joutsensalo[1], Farhat Saleemi[2], Kari Luostarinen[3]
1 Department of Information Technology, University of Jyväskylä, Finland
2 Department of Electronic Engineering, Lahore College For Women University, Lahore, Pakistan
3 Metso Paper Inc. Jyväskylä, Finland

***Abstract*—Wireless networks are one of the most essential components of the communication networks. In contrast to the wired networks, the inherent broadcast nature of wireless networks provides a breeding ground for both opportunities and challenges ranging from security to reliability. Moreover, energy is a fundamental design constraint in wireless networks. The boom of wireless network is closely coupled with the schemes that can reduce energy consumption. Network coding for the wireless networks is seen as a potential candidate scheme that can help overcome the energy and security challenges while providing significant benefits. This paper presents a basic model for formulating network coding problem in wireless setting. Since the optimal solution to wireless network coding is NP (non-deterministic polynomial-time)-hard, we intend to explore the impact of different parameters with random and non-random solutions. We present extensive simulation to show the strength of random network coding scheme in general wireless network scenarios. This study provide a comprehensive insight into the limits of wireless network coding.**

I. Introduction

Wireless networks are one of the most essential components of the communication networks. In contrast to the wired networks, the inherent broadcast nature of wireless networks provides a breeding ground for both opportunities and challenges ranging from security to reliability. Moreover, energy is a fundamental design constraint in wireless networks. The boom of wireless network is closely coupled with the schemes that can reduce energy consumption. Network coding for the wireless networks is seen as a potential candidate scheme that can help overcome the energy and security challenges while providing significant benefits.

In reference to its applications in wireless network, the typical setting of the network coding problem of consist of a server and a set of clients. Server has set of all packets needed by clients. Each client needs a packet(or packets) called *required packets*and might have *overheard packets* due to broadcast nature of wireless networks during some previous transmission. Server can make use of the *overheard* packets at the clients to reduce the overall number of transmissions. The goal of the wireless network coding problem is to reduce the number of transmissions as explained in the following example.

Consider an example of a simple wireless network shown in Fig 1 consisting of a server and four clients. In this example a server(wireless tower) needs to satisfy 4 wireless clients by transmitting 4 packets p_1, p_2, p_3, p_4 if it does not use network coding but using network coding the transmissions is reduced to half i.e. $p_1 + p_2$ and $p_3 + p_4$.

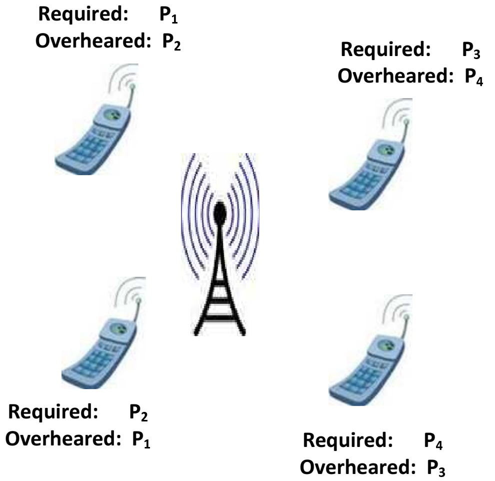


Fig. 1. An example of network coding in wireless networks

Unfortunately optimal solution to network coding problem for wireless network is NP-Complete i.e. a polynomial time solution does not exist. Therefore finding the minimum number of transmissions and the coded packet combinations that can help in reducing the number of transmissions is not possible in a realistic time for any general input. Due to hardness of finding exact solution to the problem, a set of algorithms is developed to address the issue which, in most of the practical situations, guarantees a lesser number of transmissions as compared to traditional schemes although not the least one.

This paper presents a basic model for formulating network coding problem in wireless setting. Since the optimal solution to wireless network coding is NP-hard [1], we intend to explore the impact of different parameters with random and non-random solutions. We present extensive simulation to show the strength of random network coding scheme in general wireless network scenarios. This study provide a comprehensive insight into the limits of wireless network coding.

II. Previous Work

The network coding was firstly introduced by Ahlswede et al. [2] in 2000. Figure 2 represents the famous example given by Ahlswede et al. showing the necessity of using network coding i.e., coding at the intermediate nodes of the network in order to achieve multicast capacity. The main result by Ahlswede et al. [2] was to show that network coding can help to send the data traffic at the same rate as the min-cut between the sender and the receiver. The work on network coding was further explored by Koetter and Medard [3] giving

Corresponding author Marium Jalal Chaudhry mariumjalal@yahoo.com

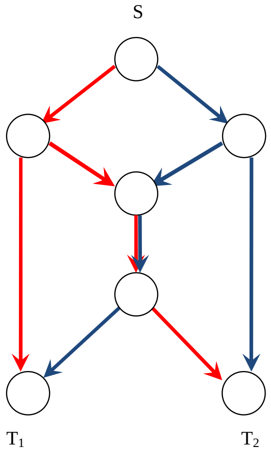


Fig. 2. An example of a butterfly network [2]

the first algebraic framework for linear network coding over any given network. The results of the authors in [3] also ties network coding with the robustness. Koetter and Medard [3] also studied their proposed solutions for more practical cyclic networks that incorporate delays. Authors in [4] have also presented a sound mathematical model for finding the network coding solution by expressing the global coding coefficients as transfer matrices, which was also shown to have an interesting relationship with a mapping to the bipartite matching and then to network flows. Single-source multicast network was studied in detail by Sanders et al. [5] and Jaggi et al. [6]. They not only presented the bounds on the required field size but also presented the first deterministic algorithm for the network coding solutions. Fragouli et al. [7] presented an easier and efficient heuristic solution for the network coding problem by relating it to the graph coloring problem. Rasala Lehman and Lehman [8] and Feder et al. [9] studied how the characterization of any networks impacts the feasibility of the proposed solution. Network Coding problem over undirected networks was explored by Li and Li [10] bounding the network coding gain by a factor of 2.

III. Organization

This paper presents our basic model for formulating network coding problem into a graph theoretic problem in Section IV followed by the suggested performance metric. Then we formally define the problem statement for wireless network. We describe two algorithms in Section V. One is random network coding and other is non-random network coding. We then present extensive simulation studies in Section V-E. Finally we conclude in Section VI.

IV. Generic Network Model For Wireless Networks

The basic model presented in this chapter is specific for single-hop wireless channel with one server s and a set of x terminals $T = \{t_1, \ldots, t_x\}$ [1], [11], [12]. The server has a pool $P = \{p_1, p_2, \ldots, p_n\}$ of n packets that are required by clients and server has to transmit either all or a subset of these packets to the clients such that all clients receive whatever they require. We define two sets for each client $t_i \in T$:

1) $Overheard(t_i) \subseteq P$ - the set of packets overheard by terminal t_i;
2) $Required(t_i) \subseteq P$ - the set of packets required by terminal t_i.

The coefficient of each packet is an element of a finite field $GF(q)$. We assume packets of one bit each without loss of generality as large packets can always be divided into smaller packets of one bit size that can be send separately with the network coding applied to each of the smaller packets individually [13], [11].

The server knows the packets in $Overheard(t_i)$ and $Required(t_i)$ set for all $t_i \in T$ the can transmit any packet from pool P as well as can transmit any linear combinations (over $GF(q)$) of packets in that P. Each transmission i is specified by an encoding vector x_i. The problem is to find the set of encoding vectors $\Phi = \{g_i\}$ of minimum rank that allows each terminal to decode the packets requested.

We assume, without loss of generality, that for each terminal $t_i \in T$, there exists at least one packet $p_i \in P$ such that p_i belongs to the Required set $Required(t_i)$ of terminal t_i. We also assume that for each client $Required(t_i) \cap Overheard(t_i) = \emptyset$.

To show advantage of network coding over traditional broadcast scheme two parameters of performance is define for any underlying network model.

A. Coding Gain

The gain in terms of number of ratio of minimum number of transmissions needed for a source to satisfy all terminals using simple broadcast over network coded broadcast is termed as coding gain G is defined as :

$$G = n/m \tag{1}$$

where m is number of message sent using traditional method and n is number of messages sent using network coding.

B. Problem Statement

For a wireless network with a single server s and a set of terminals T and a pool of packets P and each client has at least one packet in its Required set and Overheard set, minimize the number of transmissions to satisfy each and every client.

V. Practical Aspects of Wireless Network Coding

Network Coding is proved to be NP-complete in wireless network scenarios i.e. finding solution in polynomial time is not possible. Working in wireless setting is constrained by a energy efficient scheme therefore the parameter of concern is to reduce the number of transmissions required to fulfill demand of all users. Wireless network have inherent overhearing setup that lays the basics of network coding and helps in achieving much better results than serving each user separately.

We study two basic type of algorithms for network coding in wireless networks in terms of the several important parameters.

- Random network coding
- Non random network coding

A. Random Network Coding

This is a very simple technique for solving the network coding problem for minimizing the number of transmissions required to satisfy all terminals. Simple and easy but proved to be much efficient in most of the practical scenarios which are studied during simulations and presented in Section V-E.

In random network coding Server send out random linear combinations of memory contents or the packets form the pool P at every transmission opportunity. Terminal nodes can decode the desired packets if :

- Received Packet contains the desired packet
- Received Packet is a coded packet of the desired packet with packets in overheard packet of the client.

Random network coding technique is different from the traditional networking approaches in which server need to broadcast each and every packet separately.

B. Non-Random Network Coding

In this scheme server transmit packets that are coded deterministically from the $Required(t_i)$ and $Overheard(t_i)$ sets of each terminal.

C. Simulation Setup

For the purpose of each experiment in the Section V-E we randomly generate an instance of the wireless network coding as follows. We generate a server with 100 packets $p_1, p_2, \cdots, p_{100}$. For some experiments we generate n clients with a randomly generated *Required-set* and *Overheard-set*, where n is an integer and varies from 1 to 100. The distribution of packets in *Required-set* and *Overheard-set* for each client is selected by a probabilistic distribution which is either uniform or Gaussian distribution. Moreover the cardinality of *Required-set* and *Overheard-set* is also selected with uniform random distribution ranging from 1 to n.

D. Simulation Parameters

We have studied the effect of following parameters on the wireless network:

- **Cardinality of Required-set:** Number of packets Required by a client.
- **Cardinality of Overheard-set:** Number of packets overheard by a client
- **Number of clients**
- **Distribution of packets in both Required and Overheard sets:** Packet required by each client can be chosen uniformly or according to Gaussian distribution.

E. Results and Observations

We have extensively studied the wireless network under different settings of the simulation parameters defined in Section V-D. The results are as follows:

- Fig 3 shows the results for 5 clients and with random cardinality of required and Overheard sets, uniform distribution of packets and with non-random Coding technique.
- **Observation:**Coding gain can go as high as 5 for 15% of cases studied and on average gain is about 2.25.
- Fig 4 shows the results for 5 clients and with random cardinality of required and Overheard sets, uniform distribution of packets and Random Coding technique.

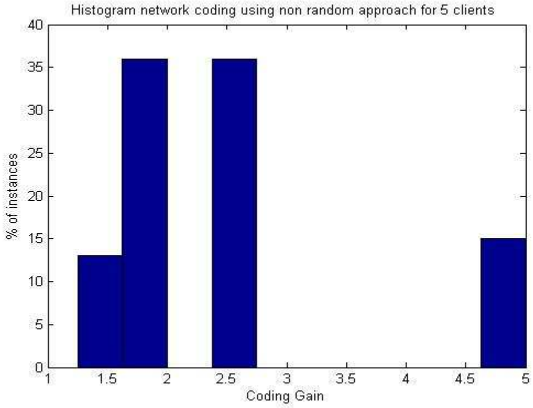


Fig. 3. For 5 clients and with random cardinality of required and Overheard sets, uniform distribution of packets and with non-random Coding technique.

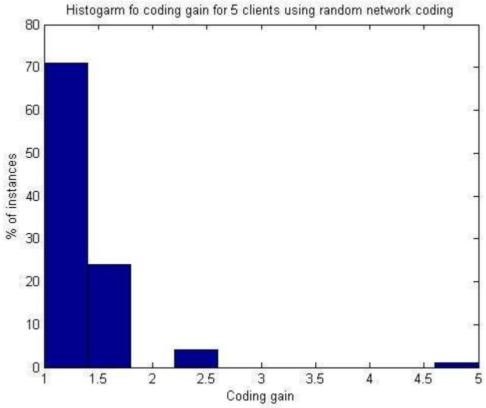


Fig. 4. For 5 clients and with random cardinality of required and Overheard sets, uniform distribution of packets and Random Coding technique.

- **Observation:**Coding gain can go as high as 5 for 2% of cases studied and on average gain is about 2.
- Fig 5 shows the results for 10 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is non-random
- **Observation:** Coding gain can go as high as 3.25 for 5% of cases studied and on average gain is about 1.75.
- Fig 6 shows the results for 10 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is random
- **Observation:**Coding gain can go as high as 1.6 for 2% of cases studied and on average gain is about 1.2.
- Fig 7 shows the results for 20 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is non-random
- **Observation:** Coding gain can go as high as 1.8 for 3% of cases studied and on average gain is about 1.4.
- Fig 8 shows the results for 20 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used

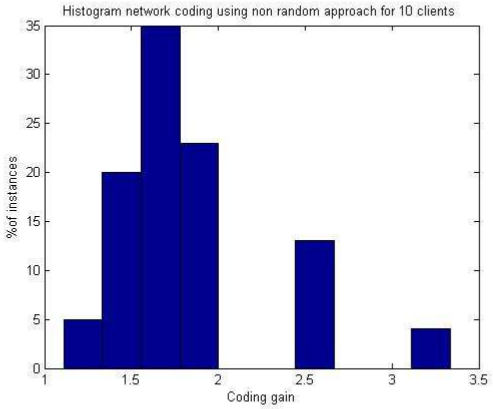


Fig. 5. For 10 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is non-random.

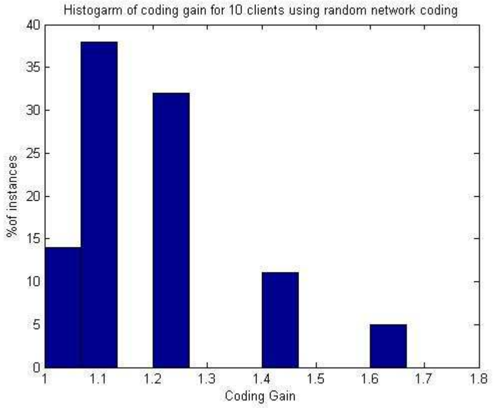


Fig. 6. For 10 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is random.

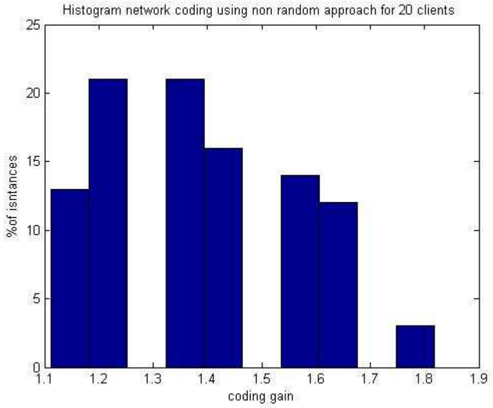


Fig. 7. For 20 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is non-random.

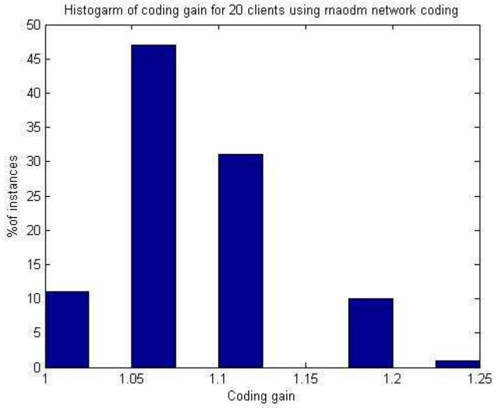


Fig. 8. For 20 clients and with random cardinality of required and Overheard sets and distribution of packets chosen is uniform. Coding technique used is random.

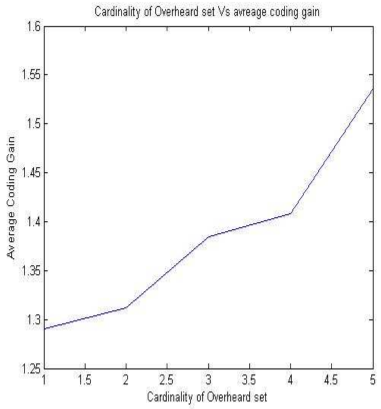


Fig. 9. For 5 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is random.

is random

- **Observation:**Coding gain can go as high as 1.25 for 2% of cases studied and on average gain is about 1.08.
- Fig 9 and Fig 10 shows the results for 5 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform, Coding technique used is random and non-random respectively.
- **Observation:** Average coding gain increases with increase in number of packets in overheard set. As compared to coding gain with random technique, non-random gives much higher coding gain with sharp rise.
- Fig 11 and Fig 12 shows the results for 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is random and non-random respectively.
- **Observation:** Average coding gain increases with increase in number of packets in overheard set. As compared to coding gain with random technique, non-random gives much higher coding gain with sharp rise.
- Fig 13 and Fig 14 shows the results for 1 up to 10

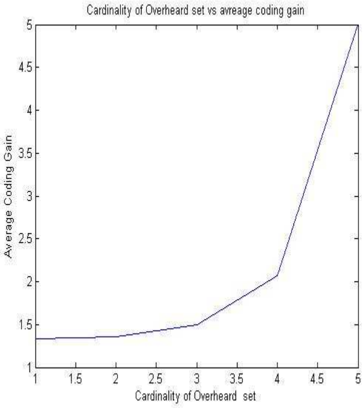


Fig. 10. For 5 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is non-random.

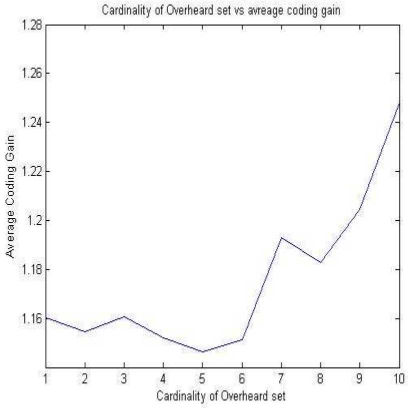


Fig. 11. For 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is random.

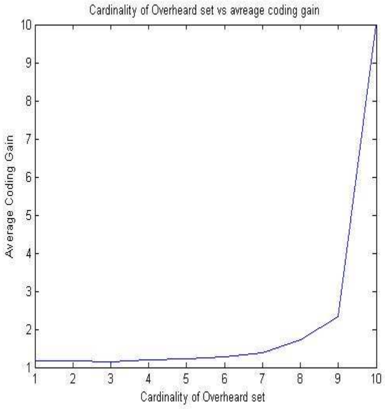


Fig. 12. For 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is non-random.

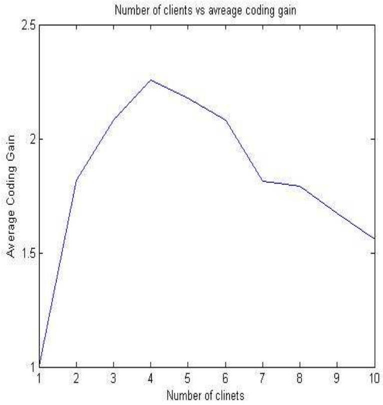


Fig. 13. For 1 up to 10 clients and with fixed cardinality of Overheard set and required set is random and distribution of packets chosen is Gaussian. Coding technique used is non-random.

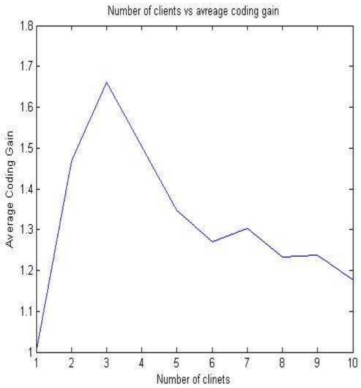


Fig. 14. For 1 up to 10 clients and with fixed cardinality of Overheard set and required set is random and distribution of packets chosen is Gaussian. Coding technique used is random.

clients and with fixed cardinality of Overheard set and required set is random and distribution of packets chosen is Gaussian. Coding technique used is non-random and random respectively.

- **Observation:**Average coding gain increases with increase in number of packets in overheard set shows more network coding gain. As compared to coding gain with random technique, non-random gives much higher coding gain.
- Fig 15 and Fig 16 shows the results for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is non random. Fig 15 shows coding gain whereas Fig 16 shows time taken.
- **Observation:**Average coding gain decreases with increase in number of clients. More packets in overheard set shows more network coding gain.
- Fig 17 and Fig 18 shows coding gain and time taken respectively for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is Gaussian. Coding technique used is non random.

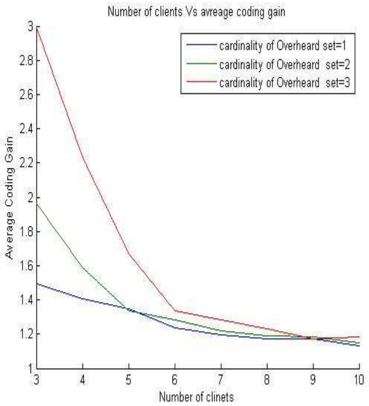


Fig. 15. Coding Gain for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is non random.

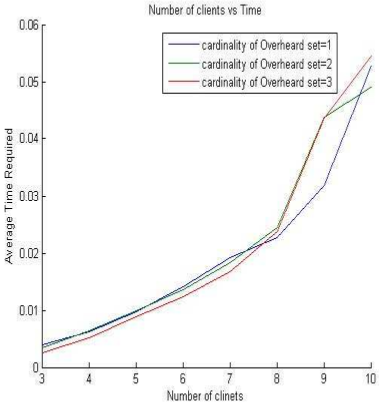


Fig. 16. Time taken for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is uniform. Coding technique used is non random.

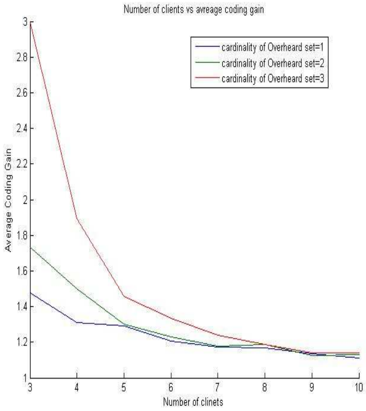


Fig. 17. Coding Gain for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is Gaussian. Coding technique used is non random

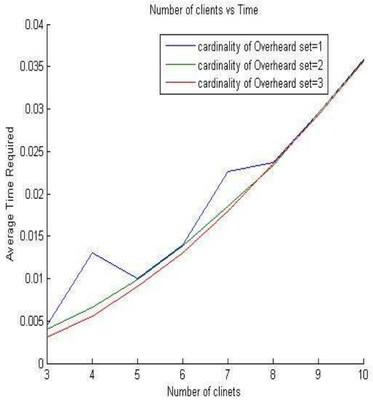


Fig. 18. Time taken for 3 up to 10 clients and with fixed cardinality of Overheard set, random cardinality of required set and distribution of packets chosen is Gaussian. Coding technique used is non random

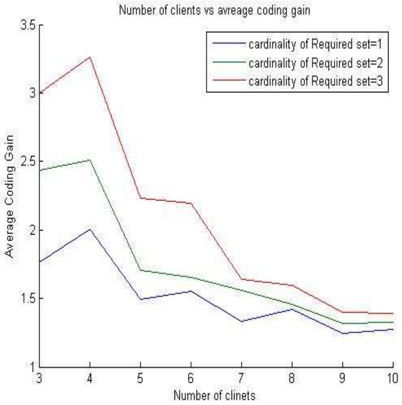


Fig. 19. Coding Gain for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is uniform. Coding technique used is non random.

- **Observation:** Average coding gain decreases with increase in number of clients. More packets in overheard set shows more network coding gain.
- Fig 19 and Fig 20 shows coding gain and time taken respectively for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is uniform. Coding technique used is non random.
- **Observation:** Average coding gain decreases with increase in number of clients whereas server takes more time to code as number of clients increases. More packets in Required set shows more network coding gain and more time.
- Fig 19 and Fig 20 shows coding gain and time taken respectively for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is Gaussian. Coding technique used is non random.
- **Observation:**Average coding gain decreases with increase in number of clients. More packets in Required set shows more network coding gain.

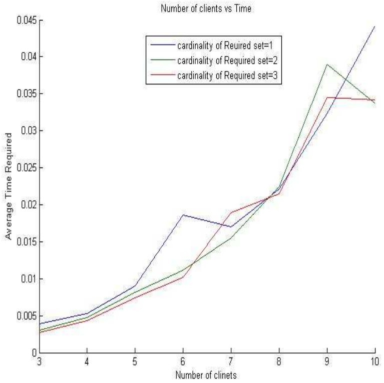


Fig. 20. Time taken for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is uniform. Coding technique used is non random.

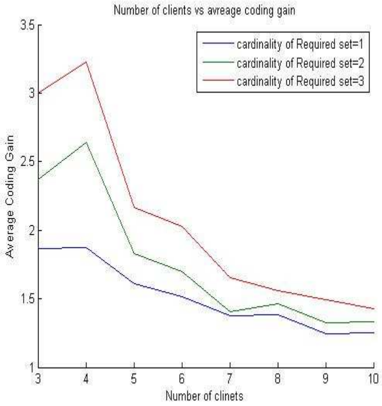


Fig. 21. Coding Gain for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is Gaussian. Coding technique used is non random.

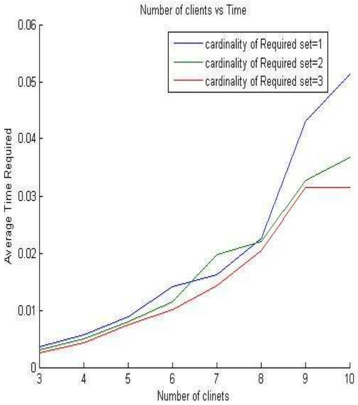


Fig. 22. Time taken for 3 up to 10 clients and with fixed cardinality of Required set, random cardinality of Overheard set and distribution of packets chosen is Gaussian. Coding technique used is non random.

VI. Conclusion

We studied random and non-random network coding for wireless network scenario. As solving the wireless network coding problem optimally is shown to be NP-hard, we concentrate on the random coding solutions for simulation section. We consider several practical scenarios and parameters that might affect the overall gain in terms of saving the number of transmissions. The results shows how the distribution of packets in *Required-set* and *Overheard-set*, the cardinality of packets in the *Overheard-set* of a client, and the number of clients affect the gain of using network coding schemes over traditional schemes. The experiments shows that in all the cases network coding shows positive gains which is most prominent when the number of clients and the cardinality of packets in *Overheard-set* is higher. The experiment results strongly supports the usefulness of the wireless network coding.

In future we would like to extend the results to the scenario when the channel condition is not ideal (i.e. lossless). We would like to explore the affect and benefits of the wireless network coding schemes under the faulty channels like bit flip channels as well as completely Byzantine channels.

References

[1] S.Y. El Rouayheb, M.A.R. Chaudhry, and A. Sprintson. On the minimum number of transmissions in single-hop wireless coding networks. *Information Theory Workshop, 2007. ITW '07. IEEE*, pages 120–125, Sept. 2007.

[2] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung. Network Information Flow. *IEEE Transactions on Information Theory*, 46(4):1204–1216, 2000.

[3] R. Koetter and M. Medard. An Algebraic Approach to Network Coding. *IEEE/ACM Transactions on Networking*, 11(5):782 – 795, 2003.

[4] T. Ho, D.R. Karger, M. Medard, and R. Koetter. Network coding from a network flow perspective. *Information Theory, 2003. Proceedings. IEEE International Symposium on*, pages 441–, June-4 July 2003.

[5] Peter Sanders, Sebastian Egner, and Ludo Tolhuizen. Polynomial time algorithms for network information flow. In *SPAA '03: Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures*, pages 286–294, New York, NY, USA, 2003. ACM.

[6] S. Jaggi, P.A. Chou, and K. Jain. Low complexity algebraic multicast network codes. *Information Theory, 2003. Proceedings. IEEE International Symposium on*, pages 368–368, June-4 July 2003.

[7] C. Fragouli, E. Soljanin, and A. Shokrollahi. Network coding as a coloring problem (Invited paper). In *IEEE Annual Conference on Information Sciences and Systems (CISS 2004).*

[8] April Rasala Lehman and Eric Lehman. Complexity classification of network information flow problems. In *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 142–150, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

[9] Ami Tavory, Meir Feder, and Dana Ron. Bounds on linear codes for network multicast. *Electronic Colloquium on Computational Complexity*, 10:2003, 2003.

[10] Z. Li and B. Li. Network coding in undirected networks. *in Proc. 38th Annu. Confe. Information Sciences and Systems (CISS)*, 2004.

[11] M.A.R. Chaudhry and A. Sprintson. Efficient algorithms for index coding. *Computer Communications Workshops, 2008. INFOCOM. IEEE Conference on*, pages 1–4, April 2008.

[12] Sachin Katti, Hariharan Rahul, Wenjun Hu, Dina Katabi, Muriel Médard, and Jon Crowcroft. Xors in the air: practical wireless network coding. *SIGCOMM Comput. Commun. Rev.*, 36(4):243–254, 2006.

[13] P.A. Chou, Y. Wu, and K. Jain. Practical network coding. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, volume 41, pages 40–49, 2003.

An Energy-Efficient Transmission Scheme for Cooperative MIMO Wireless Networks

Jain-Shing Liu
Dept. of Computer Science and Information Engineering, Providence University, Taichung, Taiwan, R.O.C.

Abstract—*Cooperative Multiple-Input Multiple Output (MIMO) schemes are recently shown to be able to reduce the transmission energy in distributed wireless sensor networks (WSNs). Given the merits of MIMO, in this work we aim to develop a joint routing, scheduling and stream control solution that can minimize energy consumption while satisfying a given end-to-end (ETE) traffic demand in scheduling-based multihop WSNs. To this end, we present a cross-layer formulation that can incorporate various power and rate adaptation schemes, and take into account an antenna beam pattern model and a signal-to-interference-and-noise (SINR) constraint at the receiver side. Specifically, we propose a fast column generation algorithm along with a transmission mode generating algorithm to get rid of the complexity of having to enumerate all possible sets of scheduling links. The formulation is verified by simulation experiments, and the results show its efficiency on the energy consumption with a low time-complexity.*

Keywords: Wireless Sensor Networks, Cross-Layer Design, Cooperative MIMO System, Column Generation.

1. Introduction

In WSNs, wireless communication is identified as the dominant power-consumption operation, which continuously intensifies the interest in the development of energy-efficient wireless transmission schemes. Apart from the energy issue, it is also widely known that most applications of WSNs such as target tracking and fire detection usually have their particular requirements on ETE QoS. To meet the diverse design goals, cross-layer optimization schemes are recently proposed to take into account the problems cross physical, medium access control (MAC) and network layers. Specifically, energy efficient wireless communication schemes can now be designed to break the limits of layering principle and to jointly solve the routing, scheduling and stream control problems so as to maximize the network performance.

On the ground of cross-layer design, a MIMO antenna system in the physical layer has the potential to offer multiple Degrees of Freedom (DOFs) for communications in a station while reducing interference and improving network throughput, which motivates many related research works on wireless communication [1], [2], [3]. In particular, it excites several cross-layer design schemes for improving throughput and/or fairness on the MIMO-based wireless networks [4], [5], [6]. However, the fact that MIMO could require complex transceiver circuitry and signal processing leading to large power consumption has been shown to preclude its application to energy-constrained WSNs. To overcome this difficulty, cooperative MIMO [7] and virtual antenna array [8] are proposed to achieve the MIMO capacity by using only single antenna stations to form a virtual MIMO (VMIMO) network, attracting various techniques to conduct such networks with different approaches.

Among these virtual MIMO techniques, collaborative beamforming (CB) [9] has been proposed as a communication scheme to fully utilize spatial diversity and multiuser diversity. Apart from the above, it is also considered that CB has the promise of greatly improving network performance by remarkably increasing the transmit power gain and by providing security and interference reduction due to less transmit power being scattered in unintended directions. Given these merits, the related works proposed usually focus on its beampattern characteristics [9], [10] to know its potential on the physical layer. While analyzed by these related works in detail, the CB approach is seldom, if ever, considered by a cross-layer optimization approach that explicitly takes into account its direct impacts on the virtual MIMO. For this reason, we develop here a cross-layer formulation that can realize these impacts so as to account for the energy-efficient communications in multihop WSNs while satisfying a given ETE traffic demand. Specifically, The cross-layer design involves power and rate adaptation at the physical layer, scheduling at the MAC layer and routing at the network layer, seamlessly integrating a SINR constraint for generating active sets of links. The objective of this problem is then to compute the most energy-efficient scheduling that can satisfy the ETE traffic demand for a set of source-destination pairs. To this end, we resolve the minimum energy scheduling problem (MESP) with a linear programming (LP) approach, which involves a fast column generation algorithm (FCGA) along with a transmission mode generating algorithm (TMGA) to overcome the problem due to a huge number of transmission modes to be selectively enumerated when towards the optimal solution to this problem.

2. System Model

2.1 Overview of Collaborative Beamforming

As shown in Fig. 1, the WSN under consideration is composed by N randomly located stations in (x, y) plane, and organized into k_c clusters. Each of the stations has a single

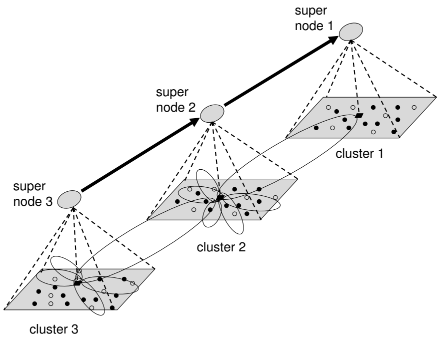


Fig. 1: Multi-hop virtual MIMO with collaborative beamforming.

antenna and operates in a half duplex mode. The rectangular coordinates of a station, (x_k, y_k), $k = 1, ..., N$, is conveniently represented by the polar ones of $\left(r_k = \sqrt{(x_k^2 + y_k^2)},\ \psi_k = tan^{-1}\left(\frac{y_k}{x_k}\right)\right)$. Let M_i, $i \in \{1, 2, ..., k_c\}$ be a cluster of stations located within the coverage range and $C_i \subset M_i$ be a set of collaborative stations to be selected from M_i.

To transmit, the source cluster head (CH), say CH_i, first broadcasts its data to the cluster members M_i. Then, the $n_i = |C_i|$ cooperative stations (CS) selected from M_i will transmit the data to the next cluster head (CH), say CH_j, $j \in \{1, 2, ..., k_c\} \backslash i$ if scheduled. Assume that the transmission target, say CH_1, is located at the direction of φ_1. In order to construct a main lobe towards CH_1, the carrier of each should be synchronized with initial phase $\Psi_k = -\frac{2\pi}{\lambda} d_k(\varphi_1)$, where λ denotes the wavelength and $d_k(\phi) \approx A - r_k \cos(\phi - \psi_k)$ is the Euclidean distance between the kth station and a point (A, ϕ) at the reference sphere $r = A$. With that, for the n_i stations, the far-field beampattern can be given by [9]

$$F(\phi / C_i) = \frac{1}{n_i} \sum_{k \in C_i} e^{j \frac{2\pi}{\lambda} r_k [\cos(\varphi_1 - \psi_k) - \cos(\phi - \psi_k)]} \tag{1}$$

In above, $F(\phi / C_i)$ is the antenna beampattern associated with C_i. As indicated in [11], the beampattern is similar to the antenna gain, but is different form the latter. That is, the pattern is a relative measure of where the transmit power is going whereas the gain is an absolute measure of how much of the transmit power is actually being transmitted in the direction ϕ. Specifically, the gain could be obtained by

$$G(\phi / C_i) = \frac{2\pi F(\phi / C_i)}{\int_0^{2\pi} F(\theta / C_i) d\theta} \tag{2}$$

2.2 Network and Communication Model

The multi-hop WSN in this context is represented by the graph of $G = (V, E)$, where V denotes a set of super nodes, in which a (super) node i, represents a cluster i (as shown in Fig. 1), and E denotes a set of links, in which a link $\{i, j\}$ represents a node i using the beamforming or multiple-input single-output (MISO) transmission to communicate with node j, or more precisely CH_j. For each link in E, the receive power is considered by [11]

$$P_{r_{\{i,j\}}} = \frac{P_{t_{\{i,j\}}} G_{t_{\{i,j\}}} G_{r_{\{i,j\}}}}{\gamma_{\{i,j\}}^{\alpha}} \left(\frac{\lambda}{4\pi}\right) \tag{3}$$

where $P_{t_{\{i,j\}}}$ denotes the transmit power of i, $G_{t_{\{i,j\}}}$ the corresponding transmit power gain, $G_{r_{\{i,j\}}}$ the receive power gain, $\gamma_{\{i,j\}}$ the Euclidean distance between i and j, and α the pathloss factor. In comparing with non-CB transmissions, it is important to note that the gain $G_{t_{\{i,j\}}}$ obtained from (2) is proportional to n_i^2 rather than n_i in the main lobe, which is the reason why CB is considered more beneficial. On the other hand, $G_{r_{\{i,j\}}}$ equals one here because the receiving CH_j is assumed to have the same isotropic antenna as the other member stations. Given P_r's and η_j that is the thermal noise at receiver CH_j, the SINR at receiver CH_j due to transmission from i in the presence of other transmissions will be

$$SINR_{\{i,j\}} = \frac{P_{r_{\{i,j\}}}}{\eta_j + \sum_{k \neq i,j} P_{r_{\{k,j\}}}} \tag{4}$$

The channel capacity of WSN associated with a link $\{i, j\}$ is a function of SINR on the channel. We assume that data is coded separately for each link and that the receivers consider unintended receptions as noises. In this case, each $\{i, j\}$ can be regarded as a single-user Gaussian channel. Consequently, with W denoting its bandwidth, the capacity of this link can then be obtained by the Shannon theory as [12]

$$LC_{\{i,j\}} = W \log_2(1 + SINR_{\{i,j\}}) \tag{5}$$

3. Problem Formulation

3.1 Energy Consumption of Virtual MIMO with Collaborate Beamforming

To model the energy consumption for the virtual MIMO with CB transmission (VMIMO-CB), we consider the following two energy components to be consumed. The first component is resulted from power amplifier, which is dependent on the transmit power P_t, and can be given by

$$P_{pa} = (1 + \alpha) P_t \tag{6}$$

where α is a factor depending on the drain efficiency of the power amplifier. The second component is resulted from the other communication circuits. As approximated in [7], the circuits for the transmitting mode include DAC, mixer (*mix*), active filter at transmitter (*filt*) and frequency synthesizer (*syn*), consuming their energies as

$$P_{ct} = P_{DAC} + P_{mix} + P_{filt} + P_{syn} \tag{7}$$

Except *mix* and *syn*, the circuits for the receiving mode include also low noise amplifier (*LNA*), intermediate frequency amplifier (*IFA*), active filter at receiver (*filr*) and ADC to contribute their energies. That is,

$$P_{cr} = P_{LNA} + P_{mix} + P_{IFA} + P_{filr} + P_{ADC} + P_{syn} \tag{8}$$

With the above, we note that VMIMO-CB in fact has two phases for data transmission. The first is intra-cluster communication phase, which involves a local exchange between the source node (represent by CH_i) and its collaborative neighbors (represented by C_i). In this phase, the energy consumption for the power amplifier is $P_{pa_{local}} = (1 + \alpha) P_{ts}$, where P_{ts} denotes the local transmit power. In addition, this phase requires that when CH_i broadcasts, all the other $n_i - 1$ members in C_i selected for CB should receive, which consumes

$$P_{CB_{local}} = P_{ct} + (n_i - 1) P_{cr} \tag{9}$$

Apart from that, the receive power is considered by (3) with $G_{t_{\{i,j\}}} = 1$, $G_{r_{\{i,j\}}} = 1$, and the assumption of $\alpha = 2$ and $\gamma_{\{i,j\}}$ being the diameter of a cluster. Then, the receive power of this link as well as those of the others with the same consideration is taken into (4) to obtain its $SINR$ value. Finally, with the resulted SINR and a given W, the data rate of this link in the first phase, $R_{b_{local}}$, is obtained by (5).

The second phase involved is for inter-cluster communication, which denotes a long-haul transmission from C_i to its next hop, say CH_j. In this phase, all stations in C_i collaboratively contribute P_{tl} for such a transmission, and overall consume $P_{pa_{long}} = (1+\alpha)P_{tl}$ energy for the amplifiers. In addition, unlike that for the first phase, there are n_i stations to be the transmitters for a single receiver CH_j, which contributes the energy consumption of

$$P_{CB_{long}} = n_i P_{ct} + P_{cr} \tag{10}$$

In this phase, the receive power is similarly considered by (3). However, $G_{t_{\{i,j\}}}$ is now obtained by the corresponding collaborative beamforming while $G_{r_{\{i,j\}}} = 1$ as before. With a longer distance, α is now assumed to be 3 with $\gamma_{\{i,j\}}$ being the distance between CH_i and CH_j. The above parameters as well as those from the other links are then taken into (3) to result in all the receive powers required by (4) to obtain the $SINR$ value for the link under consideration. Finally, with the SINR and a given W, the data rate of this link in the second phase, $R_{b_{long}}$, is obtained by (5) as well. Now, given the data rates for the two phases, we have the total rate as $R_b = (R^{-1}_{b_{local}} + R^{-1}_{b_{long}})^{-1}$. When taking all the above into account, and considering the data rates for the two phases, $R_{b_{local}}$ and $R_{b_{long}}$, we have the total energy consumption per bit for link $\{i,j\}$ as

$$E_b = \frac{P_{pa_{local}} + P_{CB_{local}}}{R_{b_{local}}} + \frac{P_{pa_{long}} + P_{CB_{long}}}{R_{b_{long}}} \tag{11}$$

Finally, we note that for a specific link $\{i,j\}$ under consideration, the above notations will be explicitly shown along with the subscript, For example, R_b and E_b are represented by $R_{b_{\{i,j\}}}$ and $E_{b_{\{i,j\}}}$, respectively, in the following sections.

3.2 Spatial-TDMA and Scheduling

In this work, it is considered that a contention-based scheme such as carrier sense multiple access (CSMA) is not well suited for providing QoS guarantees. On the contrary, it is widely accepted that its counterpart or a scheduling-based scheme such as time division multiple access (TDMA) can usually guarantee the performance of a network. In [13], the efficiency of TDMA is further improved by allowing its time-slots to be shared by simultaneously transmissions that are geographically separated, which is usually termed as Spatial-TDMA (STDMA). For the QoS guarantee, we adopt STDMA as the MAC layer, and seamlessly integrate the SINR constraint into the scheduling with its energy consumption defined as follows:

Definition 1: $T_M \subset E$ is a set of links that can be concurrently activated without violating the minimum SINR for communication ζ. That is, all the receivers of the concurrent links in T_M must have their SINR values higher than ζ. If T_M can satisfy this constraint, it is called a *transmission mode*.

Definition 2: A scheduling matrix is defined as an indexed collection, $\Gamma = \{T^1_M, T^2_M, ..., T^s_M\}$, where the index s could be an arbitrarily large finite number. A schedule S is called *feasible* if there exists a scheduling vector, $\mathbf{p}$ $= [p^1_{tm}, ..., p^s_{tm}]$ satisfying $\sum_{k=1}^{s} p^k_{tm} = 1$, and its element $p^k_{tm} \leq 0$, $1 \leq k \leq s$, denotes the duration that all the links in T^k_M can be simultaneously active in the periodically recurring time frames of STDMA.

Definition 3: The energy consumption for a transmission mode T^k_M in Γ is given by $E_{T^k_M} = \sum_{\forall\{i,j\}\in T^k_M} E_{b_{\{i,j\}}}$. Similarly, the energy consumption for a feasible scheduling S is defined as $E_S = \sum_{k=1}^{s} P^k_{tm} E_{T^k_M}$.

3.3 The minimum energy scheduling problem

Now, given the source-destination pairs of M end-to-end communication sessions, $\{s_m, d_m\}$, $1 \leq m \leq M$, our aim is thus to find the most energy-efficient scheduling that can jointly consider the routing, scheduling and stream control problems for the WSN operated under the physical layer of VMIMO-CB. To this end, we consider 1) a rate allocation $\mathbf{r}$ specifying the rate r_m for each session m, as the stream control variable, 2) a flow allocation vector $\mathbf{f^m}$ specifying the amount of traffic $f^m_{\{i,j\}}$ of session m routed through link $\{i,j\}$, as the routing variable, and 3) a transmission scheduling vector $\mathbf{p}$ specifying time fraction p^k_{tm} for each transmission mode T^k_M, as the scheduling variable. With the above, the Minimum Energy Scheduling Problem **[MESP]** can be formulated as follows:

$$\begin{array}{lll}
\textbf{Minimize} & E_S = \sum_{k=1}^{s} p^k_{tm} E_{T^k_M} & (a) \\
\textbf{subject to} & \sum_{\{i,j\}\in E^{out}_{s_m}} f^m_{\{i,j\}} - \sum_{\{i,j\}\in E^{in}_{s_m}} f^m_{\{i,j\}} = r_k, & \\
 & \quad 1 \leq m \leq M & (b) \\
 & \sum_{\{i,j\}\in E^{out}_{v}} f^m_{\{i,j\}} - \sum_{\{i,j\}\in E^{in}_{v}} f^m_{\{i,j\}} = 0, & \\
 & \quad 1 \leq m \leq M, \ \forall v \in V \backslash \{s_m, d_m\} & (c) \\
 & \sum_{m=1}^{M} f^m_{\{i,j\}} \leq \sum_{\forall T^k_M \in \Gamma : \{i,j\}\in T^k_M} p^k_{tm} \cdot R_{b_{\{i,j\}}}, & \\
(12) & \quad \forall\{i,j\} \in E & (d) \\
 & \sum_{k=1}^{s} p^k_{tm} = 1 & (e) \\
 & f^m_{\{i,j\}} \geq 0, & \\
 & \quad 1 \leq m \leq M, \ \forall\{i,j\} \in E & (f) \\
 & p^k_{tm} \geq 0, \qquad 1 \leq k \leq s & (g) \\
 & r_m \geq TL_m, \qquad 1 \leq m \leq M & (h)
\end{array}$$

In the set of constraints, (12-b) represents the conservation low for source nodes to ensure that the net amount of traffic going out of the source node of a session is equal to that of the end-to-end session rate, where $E^{out}_{s_m}$ ($E^{in}_{s_m}$) denotes the set of outgoing (incoming) links of source node s_m. (12-c) represents the conservation low for intermediate nodes to ensure that the amount of traffic of a session entering any intermediate node is equal to that exiting the intermediate node, where E_v^{out} (E_v^{in}) denotes the set of outgoing (incoming) links of node $v \in V \backslash \{s_m, d_m\}$. (12-d) gives the bandwidth constraint to make sure that the total traffic on a link is no more than the average link transmission rate. (12-e) gives the scheduling constraint, forcing that the summation of all elements in a transmission schedule vector is equal to 1. (12-f) and (12-g) simply represent the valid constraints for flow rate and transmission scheduling, respectively, and (12-h) gives the traffic demand TL_m for each session m. Finally, we note that without limits on the $f_{\{i,j\}}^m$'s involved, a session m can be routed through different links, $\{i,j\}$'s, towards its destination, which is usually called traffic splittable.

4. Minimum Energy Scheduling Computation

As can be seen in above, MESP is in general a linear programming problem. Its complexity lies in the computation of the set of all transmission modes. In fact, for the optimal solution, there may exist $s = 2^{|E|}$ such modes to be enumerated, which is not computationally efficient and should be solved with a method that can avoids the explicit enumeration. To this end, we adopt a *column generation* (CG) approach to decompose the original problem into a *master problem* and a *sub-problem*. The strategy of the CG decomposition is to operate iteratively on two separate, easier-to-solve problems [14]. The master problem will pass down a new set of cost coefficients to the sub-problem, and then receives a new column (i.e., a new transmission mode in this case) based on these cost coefficients from the sub-problem. Its purpose is to sequentially improve the upper bound by identifying new columns and adding them to the master problem.

Specifically, the master problem in this work is a restrict version of the original MESP, which considers only a subset of available columns or transmission modes. In other words, it uses only a sub-matrix $\Gamma^o \subset \Gamma$ with its index $s^o \leq s$ for this problem, and can be formulated as follows:

$$\textbf{[Master]}\text{:}\min\left(\sum_{k=1}^{s^o} P_{tm}^k E_{T_M^k}\right) \tag{13}$$

subject to the same set of constraints given in (12) while (12-d), (12-e) and (12-g) are modified with the following more concise representation:

$$\sum_{m=1}^{M} f_{\{i,j\}}^m \leq \sum_{\forall T_M^k \in \Gamma^o : \{i,j\} \in T_M^k} p_{t_m}^k \cdot R_{b_{\{i,j\}}}, \forall \{i,j\} \in E \tag{14}$$

$$\sum_k p_{t_m}^k = 1, p_{t_m}^k \geq 0, 1 \leq k \leq s^o \tag{15}$$

4.1 Sub-problem

Given the master problem, we now need to identify whether its result can be re-optimized by adding a new transmission mode to Γ^o. Denoting the dual variables corresponding to constraint (14) by $w_{\{i,j\}}$, we suggest to find out a transmission mode T_M^k that can maximize $\sum_{\{i,j\} \in E} w_{\{i,j\}} B_{e_{\{i,j\}}}$, where $B_{e_{\{i,j\}}} = \frac{1}{E_{b_{\{i,j\}}}}$ is the reverse representation of $E_{b_{\{i,j\}}}$, and denotes the number of bits per energy unit for link $\{i,j\}$. Obviously, $E_{b_{\{i,j\}}}$ involved (for the representation of $B_{e_{\{i,j\}}}$) is a complex nonlinear function as shown in (11). For the non-linear optimization problem, one option is to use a general purpose solver. Unfortunately, using such a solver is very time-consuming for nontrivial cases in usual. Here, since our aim is to solve the minimum energy problem involving the transmit power with a cross-layer approach, we simplify this function by considering that the local transmit power, P_{ts}, can be any value sufficient for the local communication, and $n_i, \forall i$, can be chosen a priori to be a fixed value also, leading to the transmit power for long-haul transmission, P_{tl}, involved in the second term of (11) to be the variable for the power adaptation. Hence, when ignoring the local transmission, we can approximate E_b in (11) by $\frac{P_{pa_{long}} + P_{CB_{long}}}{R_{b_{long}}}$. Further, by 1) ignoring the constant of $P_{CB_{long}}$ for the given $n_i, \forall i$, 2) replacing $P_{pa_{long}}$ with P_{tl} by ignoring the constant α, and 3) using $R_{b_{\{i,j\}}}$ to replace $R_{b_{long_{\{i,j\}}}}$ with the rate assumption just given, we have the following maximization for the sub-problem (corresponding to $B_{e_{\{i,j\}}}$):

$$\textbf{[Sub]}\text{:}\max\left(\sum_{\{i,j\} \in E} \frac{w_{\{i,j\}} R_{b_{\{i,j\}}}}{P_{tl_{\{i,j\}}}}\right) \tag{16}$$

However, the object function in above is still non-linear, and hard to be solved for non-trial networks. In addition, the non-linear problem would depend on the power/rate adaptation schemes to be employed with the restrictions on the VMIMO-CB transmission. For this problem, one may expect that a variable power and variable rate (VPVR) transmission method would be the most efficient scheme because it can allow each source node to vary its transmit power up to the maximum P_{max} and can result in arbitrary data rate for each active link. Nevertheless, allowing a wireless sensor station to transmit with arbitrary power and resulting in arbitrary data rate is impractical for implementation. In fact, a realistic physical design for such a station, e.g., the IEEE 802.15.4 compliant transceiver examined in [15], can support only several (8 in [15]) radio modes, and each mode is associated with a discrete transmit power. Hence, it is more proper to consider that each cluster or node in the network supports only a number of power levels. That is, a link $\{i,j\}$ is only allowed to be activated with one of the $\hat{p}_{max}$ power levels.

Given the discrete power nature, we should decide whether link $\{i,j\}$ can be turned on at a specific power level. Under the SINR model, this decision depends on the joint activity status of all other links in the network and there are $\hat{p}_{max}^{|E|-1}$ possibilities. It is clear that a classical protocol-based method [16] would not work well here since appropriately accounting for the effect of the SINR model would require an exponential complexity [17]. That is to say, although the problem may be solved by certain mathematical software tools when the network involved is quite small, it will

Algorithm 1 : Fast Column Generation Algorithm

```
1:  Set $\Gamma_o$ and $\mathcal{R}_o$ with an initial set of links; $n \leftarrow 0$;
2:  Solve [Master] in (13) to obtain the initial result of $E_S = \sum_{k=1}^{s} P_{tm}^{k} E_{T_M^k}$ (in Def. 3) as the upper bound $E_{upper}$;
3:  Calculate $E_{T_M^k} = \sum_{\forall\{i,j\} \in T_M^k} E_{b_{\{i,j\}}}$ (in Def. 3) for each $T_M^k$ in $\Gamma_o$, and collect these $E_{T_M^k}$'s in $\mathcal{E}_o$;
4:  Find the mean value of $\mathcal{R}_o$, namely $\overline{\mathcal{R}_o}$;
5:  Obtain $\Gamma_n$ and $\mathcal{R}_n$ using Algorithm 2 with the constraint that $\nexists\{i,j\} \in T_M^k : LC_{\{i,j\}} \leq \overline{\mathcal{R}_o}$ and $\nexists T_M^{k'} \in \Gamma_o == T_M^k$, and the limit that at most the number of $\mathcal{L}_o$ of $T_M^k$'s to be found for each $C_o$; $n \leftarrow n+1$;
6:  Find $T_{M_{opt}}^k = \arg\max \left( \sum_{\{i,j\} \in E, R_{b_{\{i,j\}}} \in \mathcal{R}_n} \frac{w_{\{i,j\}} R_{b_{\{i,j\}}}}{P_{tl_{\{i,j\}}}} \right)$. This step also gives $R_{opt}^k$ and $E_{T_{M_{opt}}^k}$ for $T_{M_{opt}}^k$;
7:  Add the selected mode and the related metrics into their matrices: $\Gamma_o \bigcup= T_{M_{opt}}^k$, $\mathcal{R}_o \bigcup= R_{opt}^k$, $\mathcal{E}_o \bigcup= E_{T_{M_{opt}}^k}$;
8:  Use the updated matrices to solve [Master], resulting in a new $E_S$ as the lower bound $E_{lower}$;
9:  if $E_{upper} - E_{lower} \leq \delta$ (a predefined value) | the number of iteration $\geq \beta$ (a predefined value) then
10:     Goto step 14;
11: else
12:     $E_{upper} \leftarrow E_{lower}$; Goto step 3;
13: end if
14: Terminate the CG iteration, and the solution of [Master] would be optimal or satisfactory enough;
```

be very time-consuming or even not possible to obtain its solution for a network with a reasonable size. Therefore, for solving such a hard problem, we consider an alternative that can fast provide a feasible column or transmission mode for the sub-problem, namely fast column generation algorithm, as introduced as follows.

4.2 Fast Column Generation Algorithm

The fast column generation algorithm (FCGA) under consideration is an approach that has a lower time-complexity for obtaining an energy-efficient Γ satisfactory enough for the MESP problem. To this end, FCGA first constructs an initial scheduling matrix Γ_o (as well as a rate allocation matrix $\mathcal{R}_o$ corresponding to the scheduling matrix) with an initial set of links, wherein each single link k is activated only in a transmission mode T_M^k. Then, the algorithm proceeds to compute the energy consumption given in Def. 3 for each T_M^k in Γ_o, and collect these values in the energy matrix, $\mathcal{E}_o$. Given that, it solves the **[Master]** problem in (13) to obtain the initial result. Clearly, the initial result may be sub-optimal and could be improved. Thus, the re-optimization procedure is invoked to add new T_M^k to Γ_o. For doing so, FCGA uses the rate matrix $\mathcal{R}_o$ to provide its mean rate as a lower bound for a link to be selected. This constraint is particularly considered to avoid selecting the helpless links with very low rates in order not to slow down this algorithm. Then, FCGA uses the transmission mode generating algorithm (TMGA) given in **Algorithm 2** to generate a new transmission matrix Γ_n and rate matrix $\mathcal{R}_n$ in the nth iteration. However, as shown by step 5 in **Algorithm 1**, FCGA has extra constraints on Algorithm 2 so as to avoid the same T_M^k in Γ_o and previous Γ_n to be included again, and enforce that at most the number of $\mathcal{L}_o$ of transmission modes can be found for each component (with the same color) C_o to reduce its time-complexity. Given that, in each iteration, FCGA picks up the most energy efficient transmission modes $T_{M_{opt}}^k$ based on the sub-problem formulization given in (16) with a different set of C_o. These modes are then added to the scheduling matrix Γ_o, and the corresponding energies $E_{T_{M_{opt}}^k}$ are added to the energy matrix $\mathcal{E}_o$ used for the master problem. Similarly, the rates are also included in the matrix R_o to provide its mean rate for the link selection just introduced.

4.3 Transmission Mode Generating Algorithm

The transmission mode generating algorithm (TMGA) under consideration is developed to generate a new transmission mode with respect to the actual SINR received in a node. In principle, this algorithm is designed to choose the sets of concurrent links satisfying the link contention constraint. For this, TMGA should resolve the interference due to the fact that an incoming link into a node cannot be scheduled in the same time slot with any outgoing link from the same node, and vice versa. In addition, we should assure that the SINR constraint can be satisfied for all the concurrent links involved. To this end, we define $G_c(V_c, E_c)$ as the contention graph, where a vertex in V_c denotes a link in E, and an edge in E_c denotes the corresponding links causing such interference in the node graph G. Given that, for the first aim, TMGA uses a greedy coloring algorithm, e.g., [18], that sorts vertices in G_c by decreasing vertex degrees, and with this order, colors them one by one using a first-fit greedy approach. Then, the links of G with the same color in G_c compose a C_o without the interference.

For the second aim, TMGA considers a link $\{i,j\}$ in a given C_o as the first to be admitted with a $P_{t_{\{i,j\}}}$ of level $D \in [1, \mathcal{P}]$ randomly obtained from a discrete power set, as shown in **Algorithm 2**[1]. Then it randomly chooses the other links, $\{a,b\} \in C_o \backslash \{i,j\}$, to see if they could cooperatively construct a valid T_M^k, in which no link has the same source or destination as the others, and no link has its SINR value less than the threshold ζ. In the seeking process, V is a set of values to represent the possibility that each link in C_o could be chosen. $\mathcal{Y}$ is a set of values to avoid any two links having the same source or destination to compete with each other. $\mathcal{L}$ is a tunable parameter to control how many times the seeking process would be repeated. With the above, we can control its time-complexity to be satisfactory enough while

[1] The notations shown in algorithm 2 in fact accounts for both intra-cluster phase and inter-cluster phase. For example, the notation of $P_{t_{\{i,j\}}}$ is used to represent $P_{ts_{\{i,j\}}}$ or $P_{tl_{\{i,j\}}}$, when it is considered with the parameters specific to either of the two phases.

obtaining a good Γ_n in the nth iteration that can cover all links in C_o and can evenly distribute the number of times that each link is included in certain transmission modes T_M^k. Finally, we note that TMGA also records the corresponding data rates into the rate matrix $\mathcal{R}_n$, and both Γ_n and $\mathcal{R}_n$ are used in FCGA to find $T_{M_{opt}}^k$, as introduced previously.

Algorithm 2 : Transmission Mode Generating Algorithm

```
Construct $G_c$, and color the graph to produce a set of $C_o$'s;
for all $C_o$ do
  Initialize $V_{\{i,j\}} = 0, \forall\{i,j\} \in C_o$; $\Gamma_{C_o} = \emptyset$; $\mathcal{R}_{C_o} = \emptyset$;
  for $l = 1$ to $l = \mathcal{L}$ do
    for all $\{i,j\} \in C_o$ sorted with an increasing order do
      Initialize $\mathcal{Y}_{\{i,j\}} = 1, P_{I_{\{i,j\}}} = 0, \forall\{i,j\} \in C_o$, and $flag$ = TRUE;
      $\mathcal{Y}_{\{i,j\}}$ -=1; $V_{\{i,j\}} = \mathbf{R_v}$ (a random value);
      while $flag$ == TRUE do
        if $\mathcal{Y}_{\{i,j\}} \geq 0$ then
          for all $\{a,b\} \in C_o \backslash \{i,j\}$, where $a == i \,|\, b == j$ do
            $\mathcal{Y}_{\{a,b\}}$ -=1;
          end for
          $P_{t_{\{i,j\}}} = \mathbf{PowerSet}(\mathbf{D})$ (D $\in [1, \mathcal{P}]$);
          $P_{r_{\{i,j\}}} = \frac{P_{t_{\{i,j\}}} G_{t_{\{i,j\}}} G_{r_{\{i,j\}}}}{\gamma_{\{i,j\}}^{\alpha}} \left(\frac{\lambda}{4\pi}\right)$ (in (3));
        end if
        for all $\{a,b\} \in C_o \backslash \{i,j\}$ with $\alpha_{\{a,b\}} > 0$ do
          $P_{I_{\{i,j\}}}$ += $\frac{P_{t_{\{a,b\}}} G_{t_{\{a,j\}}} G_{r_{\{a,j\}}}}{\gamma_{\{a,j\}}^{\alpha}} \left(\frac{\lambda}{4\pi}\right)$;
        end for
        if $SINR_{\{i,j\}} = \frac{P_{r_{\{i,j\}}}}{\eta_j + P_{I_{\{i,j\}}}}$ (in (4)) $< \zeta_{\hat{r}}$ then
          Roll back $V$, $\mathcal{Y}$, $\alpha$, and stamp $\{i,j\}$ as failed;
        else
          Record $\{i,j\}$ into $T_M^k$;
          Record $LC_{\{i,j\}} = W \log_2(1 + SINR_{\{i,j\}})$ (in (5)) into $R^k$;
        end if
        Find the next $\{i,j\}$ with $V_{\{i,j\}} = \min\{V\}$ and $\mathcal{Y}_{\{i,j\}} \geq 1$, among $\forall\{i,j\}$ not yet examined and failed;
        if no $\{i,j\}$ can be found then
          $flag$ = FALSE;
        else
          $\mathcal{Y}_{\{i,j\}}$ -=1; $V_{\{i,j\}} = \mathbf{R_v}$ (a random value);
        end if
      end while
      if $\nexists T_M^{k'} \in \Gamma_i == T_M^k$ then
        $\Gamma_{C_o} \bigcup= T_M^k$; $\mathcal{R}_{C_o} \bigcup= R^k$;
      end if
    end for
  end for
end for
$\Gamma_n = \bigcup_{\forall C_o} \Gamma_{C_o}$; $\mathcal{R}_n = \bigcup_{\forall C_o} \mathcal{R}_{C_o}$;
```

5. Numerical Results

In this section, we report on numerical results for the cross-layer optimization given previously. As abstractly represented in Fig. 2, we conduct the simulation environment with a network of $N = 1200$ sensor stations being divided into $k_c = 9$ clusters or nodes and each of them having $M_i \approx 133$ stations randomly distributed over an area of $2u \times 2u$, where u denotes the wavelength of carrier to be considered. In the scenario, each cluster i selects $n_i = |C_i| = 100$ stations for communication and selects the station closest to the centre as its cluster head, CH_i. Specifically, we conduct the topology so that the distance between two horizontally (vertically) neighboring cluster centres is $200u$, resulting in a regular topology that can simply exhibit the experiment results in details. For reference, we show the circuit and system parameters for the experiment in Table 1, where the power consumption values of various circuit blocks are quoted from [7].

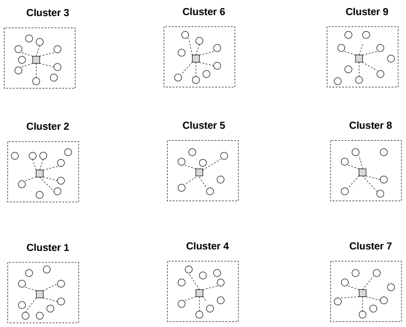


Fig. 2: Experiment topology: an abstract expression of the station/node graph.

Table 1: System parameters for the simulation experiment

Symbol	Quantity
P_{mix}	30.3 mW [7]
P_{syn}	50 mW [7]
P_{flit}	2.5 mW [7]
P_{flir}	2.5 mW [7]
P_{ADC}	9.85 mW [7]
P_{DAC}	15.48 mW [7]
P_{LNA}	20 mW [7]
P_{IFA}	3 mW [7]
W	250 KHz
α	2 : local, 3 : long-haul
β	1.9 [19]
$\zeta_{\hat{r}}$	$2^{R^{\hat{r}}/W} - 1$ from (5)
η	10^{-10}

Given the above, we set up an initial set of links $\{\{1,2\},\{2,3\},\{3,6\},\{6,9\},\{9,8\},\{8,7\},\{7,4\},\{4,1\}\}$ that can constitute an initial path for each session, and for each $\{i,j\}$ in this set, we set its P_{ts} and P_{tl} with a radio mode $TX_D, 1 \leq D \leq 8$, randomly chosen in Table 2 (quoted from [15]). Then, we concisely conduct two sessions, $\{s_1 = 1, d_1 = 9\}$ and $\{s_2 = 9, d_2 = 1\}$ in the regular topology, with

Table 2: Radio Modes and their Power Consumptions for P_t (P_{ts}, P_{tl}) (quoted from [15])

Radio Mode	Power Consumption
TX_1 (-25 dBm)	26.6 mW
TX_2 (-15 dBm)	29.8 mW
TX_3 (-10 dBm)	32.9 mW
TX_4 (-7 dBm)	36.0 mW
TX_5 (-5 dBm)	39.1 mW
TX_6 (-3 dBm)	42.1 mW
TX_7 (-1 dBm)	45.0 mW
TX_8 (0 dBm)	48.0 mW

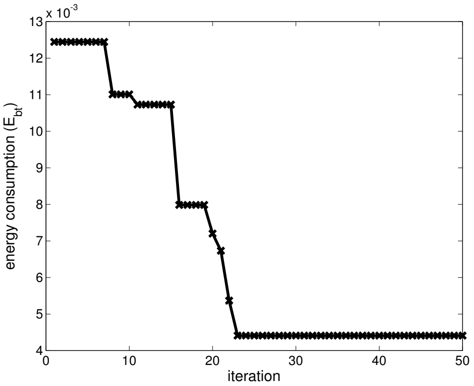


Fig. 3: Numerical results of the MESP optimization with FCGA.

their traffic loads to be the same, i.e., $TL_1 = TL_2 \approx 115.4$ Kbps, as the exemplified targets of transport layer to be achieved in the MESP problem.

With the simple setting, the time-complexity is in fact high enough. In fact, it had been indicated in [20] that the cross layer optimization problems are combinatorial in nature and computing exact solutions is, in general, NP-hard. Apparently, the same issue also exists in this work. In particular, when adopting a discrete power adaptation scheme as we did, one actually has a combinatorial problem that involves all possible links and all power levels under the SINR physical model, as indicated before. That is, even in this experiment of 9 clusters or nodes, there are already 72 possible links and 8 power levels to be considered. This is the reason why we aim to design a fast algorithm (FCGA) that can fulfill general design requirements for wireless networks with a reasonable, pragmatic size. Now, with the experimental network, the performance results in terms of E_b in each iteration are given in Fig. 3. From this figure, it can be readily seen that, at the very beginning, for the two sessions conducted the master problem will find two paths based on the initial links with their transmission modes arranged by a TDMA scheme. Then, our FCGA along with TMGA in each iteration produces new columns or transmission modes to increasingly improve the energy efficiency of this network. Finally, the optimization is stopped at the 50th iteration due to no improvement to be obtained according to our stop criterion predefined. Obviously, at the end of this experiment, we are exhibited with remarkable performance results showing that it can achieve nearly $1/3$ of energy consumption when compared with that in the initial.

6. Conclusion

In this work, we take into account the fact that for fully realizing the potential of MIMO technology, higher layer must be designed to be cognizant of the MIMO link capability. To this end, instead of simply translating the achievable gain for individual MIMO links into end-to-end gain in the network, we present a mathematical framework that can express the cross-layer gain on throughput as a function of network routing, link scheduling, and stream control in the presence of interference. With that, we propose a fast column generation algorithm and a transmission mode generating algorithm to produce TDMA-based scheduling matrices, and give a Linear Programming (LP) based optimization scheme to minimize the network energy consumption. The simulation experiment results show that the proposed optimization is capable on achieving our design aim with a low time-complexity.

Acknowledgements: This work was supported by the National Science Council, Republic of China, under grant NSC 99-2628-E-126-001.

References

[1] S. K. Jayaweera, and H. Vincent Poor. Capacity of multiple-antenna systems with both receiver and transmitter channel state information. *IEEE Trans. on Information Theory*, 49(10):2697–2709, Oct. 2003.

[2] R. S. Blum. MIMO capacity with interference. *IEEE JASC*, 21:793–801, June 2003.

[3] R. Narasimhan. Spatial multiplexing with transmit antenna and constellation selection for correlated MIMO fading channels. *IEEE Trans. on Signal Processing*, 51(11):2829–2838, Nov. 2003.

[4] Y. Lin, T. Javidi, R. L. Cruz, and L. B. Milstein. Distributed link scheduling, power control and routing for multi-hop wireless MIMO networks. *in Proc. IEEE Asilomar Conference on Signals, System and Computers*, pages 122–126, 2006.

[5] R. Bhatia, and L. Li. Throughput optimization of wireless mesh networks with MIMO links. *in Proc. IEEE INFORCOM'07*, pages 2326–2330, 2007.

[6] J. Liu, C. R. Lin, and K. Tung. Cross-Layer Design for End-to-End Throughput Maximization and Fairness in MIMO Multi-hop Wireless Networks. *EURASIP Journal on Wireless Communications and Networking*, 2009 (submitting).

[7] S. Cui, A. J. Goldsmith, and A. Bahai. Energy-Efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE Journal on Selected Areas in Communications*, 22(6):1089–1098, 2004.

[8] M. Dohler, E. Lefranc, and H. Aghvami. Space-time block codes for virtual antenna arrays. *in PIMRC, Lisbon, Portugal*, Sept. 2002.

[9] H. Ochiai, P. Mirtan, and H. V. Poor. Collaborative beamforming for distributed wireless ad hoc sensor networks. *IEEE Transactions on Signal Process*, 53(11):4110–4124, 2005.

[10] M. F. Ahmed and S. A. Vorobyov. Collaborative beamforming for wireless sensor networks with Gaussian distributed sensor nodes. *IEEE Transactions on Wireless Communications*, 8(2):638–643, Feb. 2009.

[11] C. A. Balanis. *Antenna Theory (3rd Ed.)*. John Wiley and Sons, Inc., 2005.

[12] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[13] R. Nelson and L. Kleinrock. Spatial-TDMA: a collision-free multihop channel access control. *IEEE Transactions on Communications*, 33:934–944, Sep. 1985.

[14] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali. *Linear programming and networks flows (3rd edition)*. John Wiley and Sons, 2005.

[15] M. Kohvakka, M. Kuorilehto, M. Hannikainen, and T. D. Hamalainen. Performance analysis of IEEE 802.15.4 and Zigbee for large-scale wireless sensor network application. *PE-WASUN'06*, pages 48–57, Dec. 2006.

[16] G. Sharma, C. Joo, and N. Shroff. Distributed scheduling schemes for throughput guarantees in wireless networks. *Allerton Conference*, 2006.

[17] O. Goussevskaia, Y. A. Oswald, and R. Wattenhofer. Complexity in geometric SINR. *Mobihoc*, pages 100–109, 2007.

[18] S. Ramanathan. A unified framework and algorithm for channel assignment in wireless networks. *Kluwer/ACM Journal of Wireless Networks*, 5(2):81–94, 1999.

[19] Shuguang Cui, A.J. Goldsmith, and A. Bahai. Modulation optimization under energy constraints. In *Communications, 2003. ICC '03. IEEE International Conference on*, volume 4, pages 2805 – 2811 vol.4, 11-15 2003.

[20] M. Johanssom and L. Xiao. Cross-Layer Optimization of Wireless Networks Using Nonlinear Column Generation. *IEEE transactions on Wireless Communication*, Feb. 2006.

Measurement and Comparative Analysis of UDP Traffic over Wireless Networks

Sumit Maheshwari, K. Vasu, C. S. Kumar*, Sudipta Mahapatra,
Department of Electronics and Electrical Communication Engineering
*Department of Mechanical Engineering
Indian Institute of Technology Kharagpur, Kharagpur, WB, India

Abstract - *With the increasing usage of mobile devices to ubiquitously access heterogeneous applications in wireless Internet, the measurement and analysis of Internet traffic has become a key research area. In this paper, we present the results and analysis of our measurements for CBR and VBR traffic over UDP in GPRS, UMTS and Wi-Fi networks carried over Vodafone-India and BSNL-India networks. We focus on Inter-Packet Arrival Time (IPRT) and Inter-Packet Transmission Delay (IPTD) and observe that the later has a significant impact on the round trip delay. It is also observed that choice of optimal IPTD for a particular application provides better QoS (Quality of Service) by avoiding congestion in the network. Numerical parameters for Weibull and Normal distribution are also presented in order to represent such traffic.*

Keywords- ***GPRS; UMTS, Wi-Fi, traffic measurement;***

1 Introduction

The transport layer provides a mechanism for the exchange of data between end systems. Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are two main transport protocols which provide connection-oriented and connectionless services respectively. TCP ensures reliable and ordered data delivery while also introducing processing overhead and bandwidth limitations due to congestion and flow control mechanisms. The lightweight UDP neither provides reliable delivery nor suffers from processing overhead and bandwidth limitations and hence is used in time-sensitive applications because dropping packets is preferable to waiting for delayed packets, which may not be an option in a real-time system like Voice over IP (VoIP), IPTV, video on demand and online gaming.

Although TCP is still the popular protocol in the Internet, with the increasing demand of real-time applications, which have specific Quality of Service (QoS) requirements like low delay, jitter and packet loss, UDP is gaining in popularity [1]. The self-similar nature of Internet traffic [2][10] allows researchers to measure and analyze characteristics of both flow level and packet level traffics which give a key to synthetically generate and use similar traffic for various applications, which is a time consuming process otherwise.

Wireless Internet traffic traces contain inherent information about user behaviour, interaction between users, wireless channel, applications and protocols. Thus, analysis of traces is more important than analytical modeling of wireless protocols alone. Though immensely rewarding, analyzing the statistical characteristics of wireless traffic is difficult partly due to the unfamiliar characteristics of wireless network traffic [3].

Wi-Fi and GPRS are two prime modes of accessing Internet via mobiles under the category of wireless Internet; others being 3G and WiMAX which are still gaining pace in terms of usage and accessibility in India [7]. Therefore, results of real-time measurement and analysis of traffic traces of GPRS and Wi-Fi networks can be utilized for better capacity management, congestion avoidance and testing before deployment of future networks.

Real-time applications prefer UDP over TCP. VoIP for example uses some standard codec like G.711 etc based upon the bandwidth available. By suitably varying the Inter-Packet Transmission Delay (IPTD) depending upon the codec used, congestion at the network can be significantly reduced even for data applications over UDP.

Packet switched networks have great advantages over circuit switched networks. One is that the bandwidth of the circuit is not limited to a small set of fixed allowed rates and the other is that it supports Variable Bit Rate (VBR) traffic [9]. VBR traffic makes efficient use of available bandwidth and thus its analysis becomes important. A comparative analysis of traffic over different networks like Vodafone GPRS, BSNL GPRS, BSNL UMTS and Wi-Fi gives an insight into the practical scenario and helps operators to implement better resource planning and congestion avoidance mechanisms.

Traffic measurements can be carried out at various levels like byte, packet, flow, session. Our measurements are carried out in the packet-level in view of the following benefits as compared to the other higher level methods: (a) most of the network problems (loss, delay, jitter etc.) occur at the packet level; (b) packet-level approach is independent of protocols being used; (c) traffic at the packet-level remains observable even after encryption made by different protocols.

The measurement and analysis done in this paper has the following salient features: (a) non-cooperative measurements are carried over practical networks like BSNL and Vodafone and are benchmarked by existing Wi-Fi network; (b) measurements are carried in the mobile device (essentially a mobile phone) using test-tool developed in J2ME specifically for that purpose. It deserves to note that Wireshark or any other similar measurement tool cannot be deployed in the mobile due to its limited capabilities; (c) India-specific networks are employed which gives scope for better traffic management for 3G networks which are soon to be deployed in India; (d) BSNL UMTS is also used during the experiments; (e) distribution parameters for analysis done are presented to be used by research communities to build suitable traffic models.

This paper is organized as follows. Section II introduces the testbed set-up for measurements. Section III provides the detailed description of the procedure. In Section IV we present and discuss the numerical results. Finally, Section V concludes the paper.

2 Wireless Testbed Set-up

The wireless IP QoS testbed is set-up in the CAD/CAM laboratory of Indian Institute of Technology Kharagpur, India as shown in fig. 1. It consists a few Access Points (APs), 802.11b/g, and a number of wireless (e.g. mobiles, laptops) and wired devices (e.g. computers) able to access Internet through Wi-Fi AP, LAN, GPRS or UMTS network. Wi-Fi is accessed using AP while all other networks are accessed using base stations. The packet goes through operator’s typical network, over which we have no control and thus we call these measurements as non-cooperative. The last mile i.e. from Access Point (AP) or Base Station (BS; for GPRS or UMTS) to mobile devices is Wireless. The mobile device supports Wi-Fi, GPRS and UMTS.

The server is Intel(R) core(TM)2 Duo CPU with 1.96GB RAM and 2.79GHz processor running Ubuntu 9.3 Operating System (OS) [4]. The client is Nokia N97 mobile with 128 MB RAM running Symbian OS v9.4, Series 60 rel. 5 and ARM 11 434 MHz processor which is able to access GPRS and Wi-Fi 802.11b/g using Universal Plug and Play (UPnP) technology. The mobile supports Mobile Information Device Profile (MIDP 2.1) and thus J2ME applications are used [5].

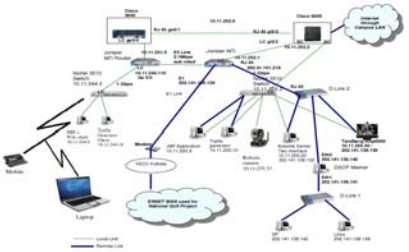


Fig. 1: Wireless IP QoS testbed

The simplified client-server architecture used for testbed used for measurements is shown in fig. 2. Base station depicts the GPRS or UMTS network. Server is set-up in the campus and can be accessed using Internet with one of the four networks i.e. Vodafone GPRS, BSNL GPRS, BSNL UMTS or Wi-Fi. All measurements and experiments are done using IPv4. Internal IP (local IP) is used while accessing server through Wi-Fi as it is an in-campus network. For all other networks, global IP is used. Java and J2ME based test applications run on the server and the client side respectively to collect the traces. Test-tool is described in the next section.

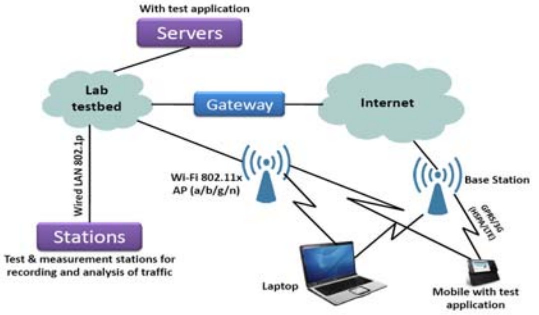

With test application
Servers
Lab testbed
Gateway
Internet
Wired LAN 802.1p
Wi-Fi 802.11x AP (a/b/g/n)
Base Station
GPRS/3G (HSPA/LTE)
Stations
Test & measurement stations for recording and analysis of traffic
Laptop
Mobile with test application

Fig. 2: Client-Server Architecture

3 Measurement Approach

To accurately measure the send and receive time of packets, we have developed Java based test tools, the architecture of which is shown in fig. 3. The main function of the tools is to time-stamp the packets at both client and server side, store it in the database and retrieve the trace file after completing the experiment.

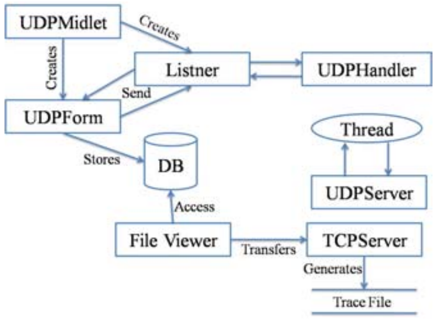


Fig. 3: Test tool architecture

The functions of various components of the tool are as follows:

UDPMIDlet: It runs in mobile client to provide the user interface.

UDPForm: This application is developed to store the actual trace file into the phone database (using RMS).

UDPEchoServer: This is the server application based on UDP to send the Echo reply to the client.

UDPHandler: The actual UDP mechanism is developed in this module.

FileViewerMidlet: It retrieves the trace file from the database and transfers to the TCPServerFile (described below) using socket mechanism.

TCPServerFile: It runs in the server to collect the trace file.

The client generates constant size packets when we input the first sample packet, and variable traffic by using random function which sends packet of variable size to the server using GPRS, UMTS or Wi-Fi network, which is selected at the beginning of each experiment. The packet size varies from 0 to 300 bytes. The maximum packet size limit is imposed by the limited capability of the mobile device. The send time for each packet is logged into the mobile (client). Server upon receiving packets keeps a copy and sends back the original packets to the mobile using the same network as used by the client. Thus, using UDP, which otherwise doesn't give acknowledgement, we are able to calculate round trip time of a packet which is used in many applications like probing etc. [11]. To avoid synchronization and other practical issues, it is assumed that the one way trip time is half of the round trip time (RTT) which is quite reasonable as the experiment is conducted several times for long run of series. Client logs the receive time for each packet and keeps the trace file in its memory. The procedure is repeated several times and then the complete trace file is transferred to a system for analysis. The number of consecutive repetitions of the experiment is limited by the fixed buffer size of the mobile device.

VoIP is a specific application of UDP protocol [12] and codec available in the literature suggest that fixed number of constant bit rate (CBR) traffic packets should be sent to the network every unit of time [13]. VBR traffic, if averaged over long period of time gives CBR traffic due to Long Range Dependence (LRD) [14]. Taking advantage of this fact and applying the features of VoIP codecs to the VBR traffic, we can fix the Inter-Packet Transmission Delay (IPTD) between the consecutive packets sent by the client to avoid congestion at the network as shown in fig 4 and fig 5.

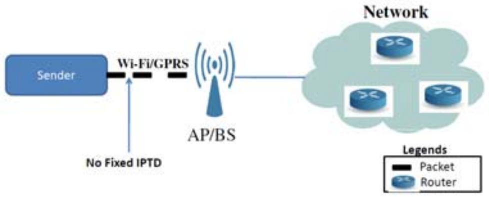


Fig. 4: No fixed Inter Packet Transmission Delay

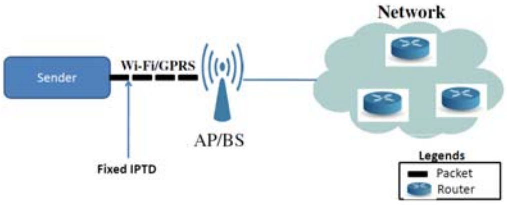


Fig. 5: Fixed Inter Packet Transmission Delay

Taking help of some standard codec we chose the value of IPTD as 0 and 25ms while the choice of the codec is arbitrary. IPTD=0 implies that there is no delay introduced between the packets. Experiments are conducted for Vodafone GPRS, BSNL GPRS, BSNL UMTS and Wi-Fi for both the cases. Round trip time (RTT) is considered as difference between the receive time and send time for the packet at the mobile side [8]. The processing and queuing delay is assumed to be negligibly small. Inter-Packet Arrival Time (IPRT) is the time difference between two successive packets arrived at the sender side. The word sender and mobile are synonymously used throughout the paper.

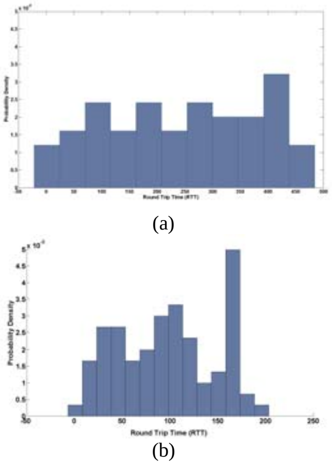
4 Numerical Results and Discussion

The data collected over all the networks i.e. GPRS, UMTS and Wi-Fi are statistically correlated. The parameters which influence the distribution are Round Trip Time (RTT) and Inter-Packet Receive Time (IPRT) for different values of IPTD. Data are found to follow some standard distributions whose parameters are obtained using MATLAB. The bin size of the distributions is decided by Sturges' formula or Square root choice for simplicity of the approach. It is observed that RTT and IPRT are higher for larger packet size for IPTD=0 case as compared to IPTD=25 for all the four networks as shown in Fig 7 and 8. The Vodafone GPRS and BSNL GPRS show almost the same performance for fixed packet case when there is no IPTD introduced. The RTT keeps on increasing for more packets entering into the network which may be due to buffering in the SGSN/GGSN. The traces collected by our test tool are verified with the Wireshark traces collected at the desired port in server and client side.

For constant packet size with IPTD=25ms, there are more fluctuations in BSNL network as compared to Vodafone GPRS network but the overall RTT is less in case of Vodafone. The introduction of 25ms IPTD avoids the buffering at the network and thus RTT seems to be constant after some time. Wi-Fi shows better performance due to its inherent advantage of having larger available bandwidth. Variable packet size case is similar to the fixed one when no IPTD is introduced between packets except that the packet loss is more when variable size packets are sent partly due to the fact that the network is unaware of the next incoming packet and thus inefficient resource allocation. BSNL network shows better performance as compared to Vodafone GPRS in the variable packet with no IPTD case. The situation for variable packets case with increasing RTT is controlled by introducing an IPTD of 25ms between packets. The performance of both, the BSNL and the Vodafone GPRS networks seem to improve in this case.

For constant packet size with no IPTD, even though the average IPRT is same for both BSNL and GPRS networks, Vodafone has low jitter than BSNL. Even after introducing IPTD of 25ms, there is no significant change in the jitter for both the networks. For variable packet size case, the jitter as well as the packet loss is high for both the networks which is controlled by introducing IPTD of 25ms. The choice of suitable value of IPTD for a particular application like conversational, streaming, interactive or background avoids congestion at the network side. By repeating the experiments for different values of IPTD, we can suitably get the IPTD range of values for different type of applications. It is worth noticing that introduction of 25ms reduced packet loss and jitter which are the important QoS parameters to be considered while running an application.

The conducted experiments use only data packets but since the network is unaware of the application running at the user end and it deals only with the packets, our measured data results can be suitably applied to other real-time applications if the delay is within the tolerable limits. The limited buffer size of the mobile imposed limit on the number of packets it can store in the trace file. Probability density functions are also found for all the data sets obtained and are fitted to the standard distribution using MATLAB. A single data set can be can be fitted by a number of distributions depending upon the values left while fitting. The choice of distribution is based upon the careful observation of data series. Fig. 6(a) shows the pdf plot for Vodafone network with variable packet size and no IPTD and Fig. 6(b) shows the same for BSNL network with constant packet size and IPTD=25ms. As we can observe that these distribution can easily be approximated to Normal or Weibull. Distribution is similarly fitted for all other network parameters.



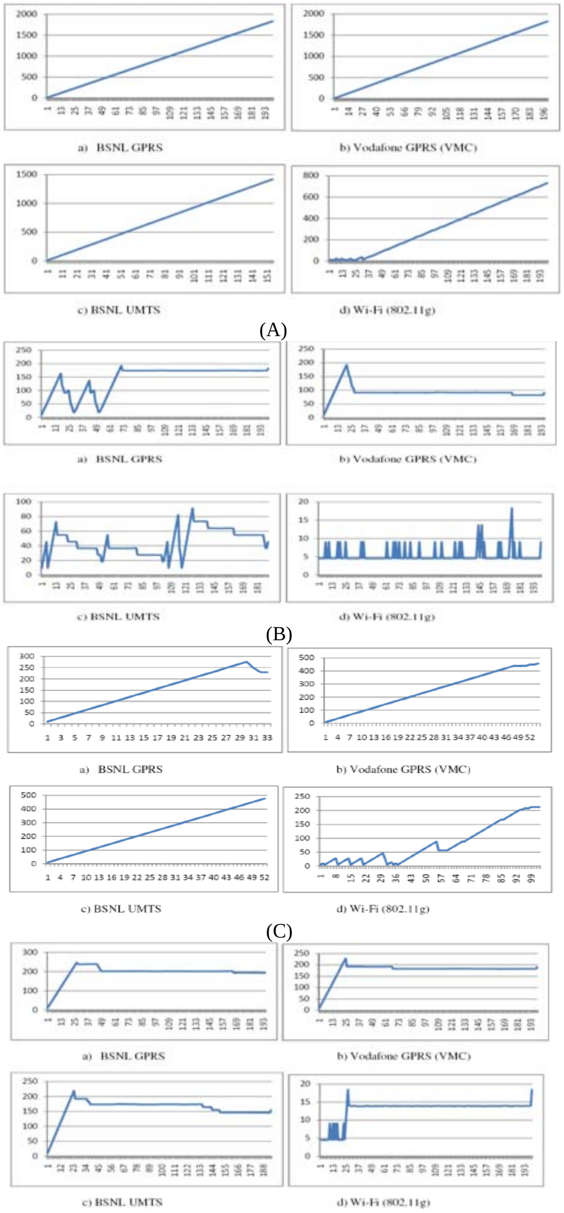
(a)

(b)

Fig. 6: PDF plot for (a) Vodafone Network for variable packet size; IPTD=0ms (b)BSNL Network for fixed packet size; IPTD=25ms; RTT is in ms.

RTT shows better performance for Wi-Fi as compared to GPRS and UMTS. The buffer management in the GPRS network is co-ordinated between the SGSN and the BSC. The rate of the data transmitted from the SGSN to the BSC in the downlink is controlled by the base station subsystem GPRS Protocol (BSSGP). The BSSGP buffers in the SGSN and in the BSC may be considered as one logical BSSGP buffer. After a maximum time determined by the BSSGP buffer setting, data is discarded from the SGSN or the BSC buffer, depending on where data is resided which results in more loss of data in GPRS than in Wi-Fi [6].

The parameters of various distributions for RTT and IPRT are presented in Table I and Table II, where the symbols have their usual meanings. RTT and IPRT give bestfit for Normal and Weibull distributions. IPTD values are in milliseconds (ms) while RTT and IPRT values are in seconds (s). Table III shows the percentage of UDP traffic in the overall traffic measured as seen in the Wireshark which runs at both, the client and the server before the start of every experiment. The percentage of low UDP traffic illustrates overhead in the network. BSNL network has very high overhead as compared to Vodafone network for all the four cases of fixed and variable packet size with and without IPTD. Table IV shows the percentage of UDP packet loss which is much higher for the variable packet case for both, BSNL and Vodafone network.



a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(A)

a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(B)

a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

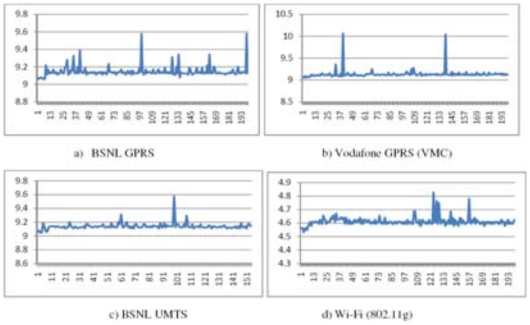
(C)

a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(D)

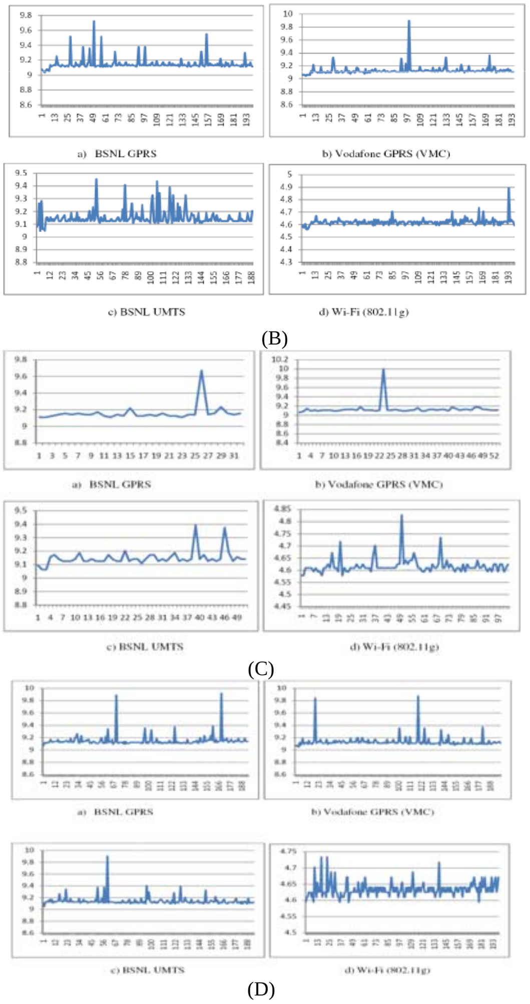
Fig. 7: Comparative Performance of RTT (A) Fixed Packet size with IPTD=0, (B) Fixed Packet Size with IPTD=25ms, (C) Variable Packet size with IPTD=0 and (D) Variable Packet size with IPTD=25ms; for all the four networks. X-axis: packet sequence number and Y-axis: RTT in seconds.



a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(A)



a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(B)

a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(C)

a) BSNL GPRS b) Vodafone GPRS (VMC)

c) BSNL UMTS d) Wi-Fi (802.11g)

(D)

Fig. 8: Comparative Performance of IPRT (A) Fixed Packet size with IPTD=0, (B) Fixed Packet Size with IPTD=25ms, (C) Variable Packet size with IPTD=0 and (D) Variable Packet size with IPTD=25ms; for all the four networks. X-axis: packet sequence number and Y-axis: IPRT in seconds.

TABLE 1 Distribution Parameters for RTT

N/W	Round Trip Time (RTT)		
	Distribution	IPTD=0	IPTD=25
Vodaf-one GPRS	Normal	μ=246.66 σ^2=139.26	μ=174.15 σ^2=32.21
	Weibull	λ=273.96 k=1.71	λ=182.46 k=9.19
BSNL GPRS	Normal	μ=149.09 σ^2=81.40	μ=192.16 σ^2=39.20
	Weibull	λ=166.09 k=1.82	λ=203.44 k=7.17
BSNL UMTS	Normal	μ=240.61 σ^2=138.67	μ=158.37 σ^2=31.11
	Weibull	λ=267.05 k=1.68	λ=167.98 k=7.21
Wi-Fi	Normal	μ=76.97 σ^2=68.92	μ=11.16 σ^2=2.46
	Weibull	λ=77.45 k=1.01	λ=11.91 k=7.27

TABLE 2 Distribution Parameters for IPRT

N/W	Inter Packet Receive Time (IPRT)		
	Distribution	IPTD=0	IPTD=25
Vodaf-one GPRS	Normal	μ=9.13 σ^2=0.12	μ=9.13 σ^2=0.08
	Weibull	λ=9.21 k=34.89	λ=9.19 k=46.99
BSNL GPRS	Normal	μ=9.16 σ^2=0.09	μ=9.15 σ^2=0.08
	Weibull	λ=9.22 k=51.65	λ=9.21 k=45.00
BSNL UMTS	Normal	μ=9.15 σ^2=0.05	μ=9.14 σ^2=0.07
	Weibull	λ=9.18 k=104.21	λ=9.19 k=50.62
Wi-Fi	Normal	μ=4.62 σ^2=0.03	μ=4.63 σ^2=0.02
	Weibull	λ=4.63 k=77.97	λ=4.65 k=150.53

TABLE 3 Percentage of UDP Traffic in Total Traffic

Percentage of UDP traffic in total traffic (as observed in Wireshark)				
	BSNL GPRS	Vodafone GPRS	BSNL UMTS	Wi-Fi
Fixed Packet No IPTD	47.35%	56.69%	78.78%	40.52%
Fixed Packet 25ms IPTD	76.62%	94.36%	87.66%	41.25%
Variable Packet No IPTD	94.34%	99.58%	92.98%	46.14%
Variable Packet 25ms IPTD	90.38%	92.77%	97.5%	40.51%

TABLE 4 Percentage Packet Loss

Percentage Packet Loss				
	BSNL GPRS	Vodafone GPRS	BSNL UMTS	Wi-Fi
Fixed Packet No IPTD	0%	0%	22.5%	0%
Fixed Packet 25ms IPTD	0%	0%	5%	0%
Variable Packet No IPTD	83%	73%	74%	48.5%
Variable Packet 25ms IPTD	2.5%	0%	2.5%	0%

5 Conclusion

Although TCP is widely accepted as the transport protocol for most of the Internet applications, UDP is showing growth because real-time applications need fast delivery. The network congestion can be avoided by making best use of UDP by introducing transmission delay between the packets. For the available bandwidth and hence the data rate, the suitable value of IPTD can be chosen which has a significance impact on RTT and IPRT. This paper shows the results of measurements carried out in a wireless testbed for UDP traffic over GPRS, UTMS and Wi-Fi networks by using test tools developed specifically for the measurement purpose. Non-cooperative, comparative analysis of the packet loss, jitter and delay is performed for Vodafone and BSNL networks and the analysis shows that UDP traffic can be imitated by using Normal and Weibull distributions. Distribution parameters are also presented to be used by research community.

Acknowledgment

This work was carried out under the Vodafone Essar Research Project, IIT Kharagpur, India.

References

[1] M. Zhang, M. Dusi, W. John, C. Chen, “Analysis of UDP traffic usage on internet backbone links” in saint'09, Ninth Annual International Symposium on Applications and the Internet, 2009

[2] M. E. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic: evidence and possible causes” in Proc. 1996 ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst., May 1996.

[3] I. W.C. Lee, A. O. Fapojuwo, “Analysis and modeling of a campus wireless network TCP/IP traffic” in Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 53, Issue 15, Pages: 2674-2687, October 2009

[4] Ubuntu Home Page, “http://www.ubuntu.com/”

[5] Nokia N97: Full Phone Specification, “http://www.gsmarena.com/nokia_n97-2615.php”

[6] A. Wennstro¨m, A. Brunstrom and J. Rendo´n, “Impact of GPRS buffering on TCP performance”, ELECTRONICS LETTERS, 30th September 2004, Vol. 40, No. 20

[7] Lehr, W. and McKnight, L., “Wireless Internet Access: 3G vs. WiFi ?”, Telecommunication Policy, pp. 351-370, 2002.

[8] Jean-Chrysostome Bolot, “End-to-End Packet Delay and Loss Behavior in the Internet”, SIGGCOMM’93, Ithaca, N.Y., USA, 1993

[9] Garrett, M., Willinger, W., “Analysis, modeling and generation of self-similar VBR video traffic”, Proceedings of ACM sigcomm, London (1994)

[10] K. Park and W. Willinger, “Self-Similar Network Traffic and Performance Evaluation”, Wiley Interscience, 2000.

[11] A. Pasztor and D. Veitch, “Active Probing using Packet Quartets,” Proc. Internet Measurement Wksp., 2002.

[12] Larissa O. Ostrowsky, Nelson L. S. da Fonseca and Cesar A. V. Melo, “A Traffic Model for UDP Flows”, ICC 2007 proceedings

[13] Codecs- voip-info.org, http://www.voip-info.org/wiki/view/Codecs”, Last accessed on Sep’ 2010.

[14] Jin Cao, William S. Cleveland, Dong Lin, and Don X. Sun, “Internet Traffic Tends To Poisson and Independent as the Load Increases”, Bell Labs Technical Report, 2001.

WiNTeB (Wireless National Test Bed) Description, Status and Opportunities

M. Cummings, Ph.D.
mcummi10@kennesaw.edu
Research Professor of Computer and Information Sciences, Kennesaw State University, Atlanta, Ga. USA

Abstract – *WiNTeB (Wireless National Test Bed) is a project that will allow research at scale (numbers and geographic reach) on cellular and hybrid networks through cooperative relationships with US national cellular network operators. The project has received a small amount of funding from NSF. NSF, expressing interest in the concept, has requested a very detailed proposal, which is in preparation. The proposal will include funding for early projects running on the test bed which creates some interesting opportunities. This presentation will include a discussion of:*
Background,
Objectives,
Description of Test Bed Services,
Status,
Opportunities.

Keywords: Wireless, National Test bed, research, Infrastructure

1 Introduction

This paper starts with the historical background including a discussion of the needs that inspired the WiNTeB [1]. It then describes its evolution through support from the leadership of GENI and NSF. It then describes how with this support a Workshop was held in the DC area bringing together leaders in academia, industry and government further refining the concept and moving to the planning stage. Finally, the paper discusses both near term opportunities (including funding) and longer term opportunities created by the project.

2 Historical Background

Several years ago, the author was working on new technology for cellular networks. Working through the conventional venture backed start-up process, he was stymied by the lack of a test bed where he could prove his technology. A friend of the author James Kempf, at Ericsson USA Labs, was frustrated in his interactions with academic researchers who had access to wired and WiFi test beds, but not cellular. The two came together and started talking about the possibility of creating a wireless test bed. This led to early written sketches.

These sketches focused on the fact that researchers in US universities interested in working on wireless networks have few choices available to them. Most are forced to use WiFi. While many interesting experiments can be done with WiFi, the ability to do at-scale experiments on the dominant wireless technology (cellular), in real world situations, is limited.

Although cellular network operators are reaching out to the application development community for Apps and this is leading to a more open environment, the carriers are driven by a need for product that can be deployed in a relatively short period of time. Therefore, it is hard for them to work with researchers and early stage technology developers who may be doing more basic work or whose results will come to market in a more substantial time frame.

2.1 Support from GENI and NSF

The leadership of the GENI [2] Project (NSF funded wired test bed) found out about the early work by Cummings and Kempf and started to help fleshing out the concept. This led to the creation of a talking document was that used to introduce the concept to NSF leadership at a GENI regular meeting. This led through a proposal process to NSF funding a May 2010 Workshop in the DC area to validate the need for such a test bed and to crystallize its definition.

2.2 WiNTeB Workshop

On May 5 & 6, 2010 the WiNTeB Workshop [1] was held in Arlington sponsored by NSF and organized by Mark Cummings of KSU. Response to the Call For Participation was very strong. Early planning had been for 24 participants. That grew to 35 and finally was expanded to 60. Even so, a large number of potential attendees had to be turned away. Workshop participants included representatives of all major classes of stakeholders:

Explanation of the WiNTeB Concept - **Mark Cummings, KSU**

Desired Outcomes for the Workshop - **Victor Frost, NSF**

What Can Be Learned From the GENI Experience - **Aaron Falk, GENI GPO, BBN**

Wireless Equipment Company View - **James Kempf, Ericsson Labs USA**

Application Researcher View - **Deborah Estrin, CENS, UCLA**

Infrastructure Researcher View - **Sachin Katti, POMI, Stanford**

Broadband Perspective - **Walter Johnston, FCC EOT**

FCC Technology Perspective - **Jon Peha, FCC Chief Technologist**

Operators Perspective - **Andrew Apple, AT&T Mobility**

WiNTeb Recommendations From an Internet Pioneer **Vint Cerf, Google**

Lightning Talk presenters were:
Hongwei Zhang, **Wayne State University**;
Rudra Dutta, **North Carolina State University**;
Driss Benhaddou, **University of Houston**;
Kuang-Ching Wang, **Clemson University**;
B. S. Manoj, **Univ. of California, San Diego**;
Rajeev Koodli, **CISCO Systems**;
Dijiang Huang, **Arizona State University**;
Per Johannson, **Univ. of California San Diego**;
Preston Marshall, **Univ. of Southern California, Information Sciences Institute**;
Larry Foore, **National Aeronautics and Space Administration**;
Rangam Subramanian, **Idaho National Laboratory**;
Aleta Ricciardi, **SRI International**.

Other participants covered a wide range. Academic participants ranged from small regional colleges to MIT. Academic researchers covered both networking and social / environmental areas. Industry participants included AT&T Mobility, Clear Wire, and Cisco. Government organizations included the FCC, Defense Spectrum Office and NASA.

The Workshop resulted in validation and refinement of the WiNTeB concept by defining three visions:

- Vision I - Applications Research
- Vision II - Research On Existing Networks
- Vision III - Research on New Modes

Vision I focuses on how WiNTeB can enable researchers in Mobile Health, Environmental Science, Social Science, Political Science, etc. to use existing cellular networks. In this vision, WiNTeB acts as an aggregator and "impedance matcher" for and between researchers and cellular network operators to ease access, match billing plans to researcher needs while providing cost effective services to researchers and protect production networks and the companies that operate them. It also envisions research scenarios which employ heterogeneous networks combining multiple technologies and multiple operators.

Vision II focuses on WiNTeB enabling research surrounding improving and evolving current cellular networks. Here WiNTeB provides a deep trusted relationship with network operators surrounded by technical and procedural safeguards to allow researchers access to network internals.

Vision III focuses on WiNTeB enabling research on new modes with special consideration of interference issues and security. To do this, the FCC would authorize experimental spectrum near enough to existing commercial bands that available commercial equipment could be tuned to it, and far enough away that research activities in the experimental band could not have a detrimental effect on production networks. In some cases, network operator partnerships would assist in the fielding of equipment.

By the end of the Workshop, a consensus developed that although these appear to be three separate visions with separate and possibly competing audiences, they are actually interactive and mutually supportive. This realization led to the conclusion that each needs the other two and that all three should be pursued together. Vint Cerf's description of DTN played an important role in the group's coming to this realization. Vint pointed out that the history of the development of DTN (Delay Tolerant Networking) showed how Application research plays a key role in Network research. He pointed out that DTN was developed as a direct result of Application research in that Exo Planet Researchers found that they could not get the data being collected by the Mars probes back to Earth with current wireless technology and protocols. This led to the development of DTN, which can be roughly described as a store and forward version of TCP/IP. It was created and loaded into craft in orbit around Mars, and the Earth to make it possible to get the data and pass it on. The development of the first version of DTN opened up an extremely wide area of experimentation and implementation in terrestrial networks.

As this consensus developed attention started to turn to the intellectual merit of WiNTeB and the other benefits it would generate. Conclusions included that WiNTeB's intellectual merit flows from what is learned about wireless networking and applications, in setting up the test bed, and the research projects conducted on it. These projects span Health, Social, Political, Environmental, Networking, and RF sciences. For example, researchers will be able to advance health and environmental sciences through field monitoring, test theories in Smart Grids, increase understanding of problems and solutions in multimedia on wireless, quality of service measurement on wireless, femtocell networking, wireless security and new modes in advanced wireless networking.

The broader impacts that flow from WiNTeB will include enabling new educational, new research and new integrated research and education opportunities. By its very nature, WiNTeB will democratize wireless research by opening the doors to significant wireless research to smaller universities, institutions and start-up companies. It will also enable the larger universities which lead wireless research today, to make their work more meaningful. It will enable researchers to provide regulators with the data they need to make the best possible public policy and regulatory decisions. It will enable research results which help NOAA, NASA, DOE and other government agencies to better fulfill their missions. Network operator, and equipment, companies will benefit from WiNTeB in two ways: improved staff availability and improved early research results pipeline feeding their innovation efforts. Finally, having more reliable, more secure, more robust networks with advanced services will have a profoundly positive impact on society. In addition, WiNTeB will result in Improvements in US industrial competitiveness of wireless equipment, software, semiconductor, and network operator companies.

2.3 Activities Following the Workshop

Following the Workshop a team came together to prepare a proposal for an NSF Community Research Infrastructure grant to provide the financing necessary to create the test bed and get it to the point where it could be self sustaining. The team consisted of people from KSU, UC San Diego, USC ISI with several advisors from industry. The author was the PI for the team. The team first worked closely with national cellular network operators to develop a practical plan for the cooperative implementation of WiNTeB. Based on this plan the team decided to prepare a single proposal covering all three visions. Following submission and review of the proposal, NSF formally expressed strong interest in the proposal but asked for more detail. Because of the proposal page limitations, with the three Visions combined, this had not been possible.

3 Current Status

Responding to NSF's request for more detail and yet complying with proposal page limit constraints, the team decided to reorganize its effort into three separate but interrelated teams developing three interrelated detailed plans / proposals, each focused on a clearly defined area: The three areas are:

- Field Research Services
- Network Research Services
- New Modes Services

The Field Research Team consists of Per Johansen from UC San Diego, Dr. Kevin Patrick from UC San Diego, Debra Estrin from UCLA, Russ Clark from Georgia Tech., Victor Clincy from KSU, Victor Marshal from KSU, with Mark Cummings from KSU as the PI.

The Network Research Team consists of Victor Frost from the University of Kansas, Per Johansen from UC San Diego, Russ Clark from Georgia Tech., Victor Clincy from KSU, Victor Marshal from KSU, with Mark Cummings from KSU as the PI.

The New Modes Team consists of Amitab Mishra from Johns Hopkins, Mark Cummings from KSU, with Preston Marshall from USC ISI as the PI.

The industry advisors will continue to support the three teams. Each of these three teams will deliver a proposal specific to their area of focus to NSF in October 2011.

The Field Research Service focuses on providing researchers using wireless sensors / actuators for research and testing in the US, a cost effective business and technical interface that fully meets their needs. These sensors /actuators can be simple cell phones, smart phones, plug in modems that support other devices, wireless pads, or more specialized equipment. It does so by having special relationships in place with national cellular operators. The need for the Field Research Service has been highlighted in a number of recent mHealth meetings. Three examples are the mHealth Summit organized by NIH [3] and held in Washington DC in the Fall of 2010; the Health Cyber infrastructure Workshop sponsored by NIH & NSF [4] held At UC San Diego at the beginning of 2011, and the Wireless Life Sciences Alliance (WLSA) Convergence Summit [5] held in San Diego in the Spring of 2011. What is emerging is a consensus that wireless can play a key role in making the US's health care system efficient and effective, but that it requires evidence-based proof of effectiveness of specific measures in changing outcomes. Early work, on a very small scale, has shown promise, but the required proof can only be achieved with studies done on a large scale, both in numbers and geographic reach.

The Network Research Service provides researchers two types of services. The first is a data mining service. This allows researchers to work with data captured by national cellular operators. This data may be used to support medical and social science research or to support networking research. An example of the type of medical research that will be supported includes epidemiological studies. An example of the type of networking research that will be supported includes network performance and security research.

The second type of Network Research Service supports research and testing of new hardware and software components for existing cellular networks. It provides a staged process for introducing these first in controlled environments on the operators premises and then, as the operators become comfortable with the controls and network protections, moving to installations on WiNTeB and then researcher premises.

Both the Field Research and the Network Research proposals will have extremely detailed plans and budgets including draft contracts with national cellular operators.

The New Modes Research Service supports research in new RF technologies and new wireless networks. It relies on the FCC to allocate experimental spectrum close enough to existing cellular spectrum to allow existing commercial equipment to be tuned over to it, but far enough away that experimental operation will not interfere with commercial operations. Thus this service requires action by the FCC. FCC staff attended the WiNTeB Workshop [1] and since then the WiNTeB team has met several times with those and other FCC staff members. This has resulted in the release of an NPRM [6] on experimental spectrum, which promises to support WiNTeB. However, since the specific frequencies are not yet known and detailed rules must wait on finalization of the new rules, it is not possible to provide the level of detail requested by NSF for this service at this time. Therefore, the New Modes proposal will be a request for a study grant leading to a formal CRI proposal the following year.

4 Opportunities

WiNTeB has been advised that it should include infrastructure funding for specific projects which can make early use of the Field Research and Network Research Services. The requirement for these is that they have all the resources in place except those needed from WiNTeB.

For example, a project which has the resources in place to analyze the data and produce valuable results if smart phones are distributed to a number of users in a number of locations and airtime is provided, would be a good candidate for this opportunity.

Similarly, a project which has the resources in place to analyze the data produced in mining data in a cellular network, but which has no mechanisms (technical, business, and financial) in place to acquire the data would also be a good candidate.

Also, a project which has the new hardware or software and resources in place to analyze the results produced in inserting the hardware or software into a cellular network, but which has no mechanisms (technical, business, and financial) in place to do that insertion, would also be a good candidate.

The possibilities for taking advantage of this near term opportunity are limited only by the imagination of the researcher. The only requirement is that the researcher has to have in hand the other resources necessary to complement the wireless infrastructure resources provided by WiNTeB.

Longer term, WiNTeB services should be factored into all relevant researchers' thinking and planning. It is the intention of WiNTeB to support a wide and expanding range of creative research and testing projects.

The only limitation is getting the word out to all the various communities who can make use of WiNTeB. That is the primary purpose of this paper and the author requests the assistance of all readers in helping to disseminate the information about WiNTeB.

Similarly, it would be helpful to have as many letters of support as possible to accompany the proposals which will be submitted in October.

For further information or to send letters of support, please contact Mark Cummings at mcummi10@kennesaw.edu.

5 Conclusion

This paper has traced the development of the WiNTeB concept from early conception to the threshold of maturity along the way highlighting the need for and uses of such a national test bed. It then discussed the potential near term opportunity for those with the capability to take advantage of early funded test bed infrastructure. Finally, it discussed the need to help with information dissemination, both by getting the word out to the full range of user communities and by providing the proposal teams with letters of support.

6 References

[1] http://www.kennesaw.edu/ogc/WiNTeb/agenda.html

[2] http://www.geni.net

[3] http://www.mhealthsummit.org

[4] http://healthcyberinfrastructure.org

[5] http://www.wirelesslifesciences.org

[6] Promoting Expanded Opportunities for Radio Experimentation and Market Trials under Part 5 of the Commission's Rules and Streamlining Other Related Rules
ET Docket No. 10-236

2006 Biennial Review of Telecommunications)
Regulations - Part 2 Administered by the)
Office Of Engineering and Technology (OET)
ET Docket No. 06-105

Detecting Masqueraders in 802.11 Wireless Networks

Liran Ma
Department of Computer Science
Texas Christian University
Fort Worth, TX 76109, USA.
Email: l.ma@tcu.edu

Abstract— *Due to the broadcast nature of the wireless medium,* 802.11 *wireless networks are especially vulnerable to masquerading attacks, where an adversary forges the identity of another victim host. Masquerading allows the adversary to gain unauthorized access to network resources or services that are designated for legitimate hosts. In this paper, we develop a novel scheme for detecting a masquerader under various scenarios inside an* 802.11 *wireless network. Our proposed scheme employs unique "fingerprints" about a wireless host to identify a masquerader. These fingerprints are extracted from the networking activities of a wireless host and applied on a naïve Bayesian classifier. Our comprehensive empirical study demonstrates that the proposed scheme is able to provide an accurate and consistent detection. To the best of our knowledge, it is the first scheme that can detect a sophisticated masquerader that is capable of frequently switching the forged identity and displays little greedy or malicious behavior. In addition, our proposed scheme does not require any modifications to the existing wireless standards.*

1. Introduction

One of the key features of the wireless medium is its broadcast nature. Unfortunately, such broadcast nature makes an 802.11 wireless network particularly vulnerable to masqueraders who forges the identities of legitimate hosts. Specifically, each 802.11 host is identified by a globally unique 12 byte media access control (MAC) address. There is a field in the MAC frame that holds the sender's MAC address, which is reported by the sender of the frame. When a MAC frame is broadcasted, so does its identity (i.e., the MAC address). However, there is no mechanism for validating the authenticity of the self-reported identity in the 802.11 standard specifications. As a result, an adversary may "masquerade" the identities of other hosts and request various network services (such as access to the Internet) on their behalf.

A masquerader not only consumes network resources without proper authorization, but also introduces significant security threats. A masquerader has all legitimate security credentials and therefore can launch various insider attacks, or even completely disrupt the normal network operations. It is indeed an "attack multiplier" and therefore is extremely destructive to the network. Thus, it becomes imperative to provide an effective and efficient solution for masquerader detection to limit damages. Much efforts have been devoted towards the problem of masquerader detection in the literature. Traditionally, masquerading can potentially be thwarted by using cryptographic tools (such as RADIUS and EAP protocols). However, such cryptographic based authentication approaches may introduce overheads (such as certificate distribution and management) that are prohibitive for commodity wireless networks. Subsequently, the research community has also sought non-cryptographic approaches to identify masqueraders. Example of these approaches include comparing the MAC sequence numbers [1], [2] and the received signal strength values [3]–[5]. These approaches can detect a simple masquerader. However, a sophisticated masquerader can frequently change its spoofed identity to escape from being revealed. Existing approaches cannot be applied directly to detect a sophisticated masquerader. Detecting a sophisticated masquerader is a nontrivial problem due to the challenges resulted from the frequently changing spoofed identities. Another key challenge is that a sophisticated masquerader may show little suspicious activities such as greedy or malicious behavior.

To answer these challenges, we propose a novel masquerader detection scheme for IEEE 802.11 wireless networks. The proposed scheme employs unique "fingerprints" about a wireless host to identify a masquerader. These fingerprints are extracted from the networking activities of a wireless host and applied on a naïve *Bayesian* classifier. The proposed scheme possesses four desirable features. Firstly, to our best knowledge, it is the first scheme that can detect a sophisticated masquerader. Secondly, our comprehensive empirical study demonstrates that the proposed scheme is able to accurately and consistently detect a sophisticated masquerader. Thirdly, it works in conjunction with current wireless security protocols, and it does not require modifications to the underlying 802.11 standards. Finally, it can be implemented in the firmware of existing access point hardware or a computer that is able to capture wireless traffic.

The rest of the paper is organized as follows. Section 2 discusses related work. The design of our proposed scheme is elaborated in Section 3. Section 4 and Section 5 present the evaluation settings and results. Finally, our conclusion appears in Section 6.

2. Related Work

There is some work in the literature on detecting a masquerader inside a wireless network. One initial approach is to examine the order of the MAC frame sequence numbers. Each MAC frame header has a field that contains a 12 bit sequence number, which is incremented for each subsequent frame. A masquerader may have a different sequence number ordering compared to that of the legitimate host. Thus, the abrupt changes in sequence number ordering are exploited to detect a masquerader in [1], [2]. Nonetheless, there are various reasons (such as a reboot) that a wireless host may reset the frame sequence counter. When a reset happens, MAC frame sequence number based techniques can produce false positives.

Another popular approach relies on scrutinizing the physical location of a transmitting host. The observed received signal strength indicator (RSSI) values can be used to infer the physical location of a transmitter. A tuple of RSSI values reported by different access points are employed to generate “signalprints” for a transmitter in [3]. A mismatch between signalprints reveals the existence of a masquerader. In another work [4], RSSI values are applied on the k-means clustering to detect a masquerader. If a cluster centroid separation exceeds 6 db, it can be concluded that there is a masquerader. However, physical location based solutions are not effective enough in mobile scenarios. When a host is moving, the measured RSSI values may oscillate heavily. As a result, it can introduce a significant number of false positives.

To further reduce false positives, a layered architecture that combines a series of detectors (such as the MAC frame sequence number and physical locations) is proposed in [5]. This approach can achieve a higher detection accuracy compared to the aforementioned solutions. Yet, this type of approach will fail if the masquerader and the victim do not present in the network at the same time because there is no target to compare with. A sophisticated masquerader can intelligently change its spoofed identity to be that of an offline host. Thus, none of these existing schemes can detect a sophisticated masquerader.

Recently, advancements in wireless network interface card (NIC) fingerprinting offer new perspectives towards the problem of masquerader detection. For example, the minute imperfections of wireless transmitter hardware are generated at manufacture. These imperfections are unique to a specific transmitter. Therefore, a technique called PARADIS is proposed to identify the source NIC of an 802.11 frame through passive radio-frequency analysis on these imperfections in [6]. Examples of other NIC fingerprinting based techniques of this category include using the clock skews [7] and the probe request frequency [8]. Nevertheless, there still lacks a satisfactory and practical solution that is competent enough to tackle a sophisticated masquerader.

There are other works that focus on greedy (or selfish) user detection [9]–[11]. A greedy user refers to a user that disproportionately occupies the network resources. For example, an AP-based system called DOMINO is proposed for the identification of greedy users that excessively occupy network bandwidth [9]. In [10], [11], attention is paid to specifically detect the MAC layer misbehavior of selfish wireless stations. Yet, due to the fundamental differences between a masquerader and a greedy user, these schemes cannot be applied directly to the problem of masquerader detection.

3. Our Masquerader Detection Scheme

A masquerader may bear the identity of a legitimate wireless host. However, the masquerader may not display the same externally observable features as the legitimate host. These feature discrepancies can result from network activity differences between the masquerader and the legitimate host. Our proposed scheme extracts information from wireless frames to identify the masquerader. We assume that wireless frames can be collected continuously from the entire area of interest. Such collection can be easily done by a commodity wireless interface that is able to tune into the promiscuous mode.

In the follow subsections, we will explain: i) how to select features that are discriminating; ii) how to construct a naïve *Bayesian* classifier based on the selected features that can accurately and efficiently detect a masquerader.

3.1 Feature Selection

Feature selection plays an important role in classification. The ability to identify powerful discriminating features is critical to classification accuracy. Moreover, selecting non-redundant features can reduce computational overhead, and thus improve efficiency. Inspired by previous work [7], [8], [12], [13], we choose three features. These three features are obtained from destination <IP address, port number> tuples, 802.11 MAC layer frames, and application level broadcasts, referred as *IPs*, *MACs*, and *Broadcasts*, respectively.

The destination <IP address, port number> tuple has been widely used for user classification [12], [13] because it includes unique information about a user’s network visiting habits. Yet, this information may not be always available due to link-layer encryption (e.g., WEP or WPA).

The MAC layer frames contain a wide range information of a user’s wireless NIC driver and its customized configurations. Due to the ambiguity of the 802.11 standard, different implementations of the same protocol specification in NIC drivers behave differently. The configurable parameters include power management algorithms, collision control algorithms (e.g., RTS/CTS), supported data transmission rates, transmission power, timing of probing, and etc. In addition, how the driver is manipulated also makes some differences in the fingerprints of a NIC driver. For instance, the *Wireless Zero Configuration* of Windows XP manages the configuration of the settings for a wireless device by default. Besides, there exist stand-alone programs (vendor provided, e.g., Intel PROSet/Wireless Software) that perform those functions. These stand-alone programs are provided by the manufacturers of wireless devices and often support more networking functionalities. The fundamental reason for these parameter differences lies in the inherent ambiguity of human specifications, and manufacturers' implementation flexibility to meet different constraints. Using these parameters, a NIC, and thus its host, can be identified [8]. Note that these parameters are always sent in the clear even when IP layer data are encrypted.

Broadcasts are used by many user level application software to announce their existence. The example of these broadcast packets include Microsoft Printing Service advertisements, Samba broadcasts, the P2P services (e.g., BT downloading), and the Apple iTune advertisements. Most of these broadcasts can be uniquely recognized by the combination of their sizes and port numbers, which is fixed according to the corresponding protocols. Furthermore, if the port number becomes invisible because of applying encryption, the broadcasted packet size alone still has discriminating power among users. Thus, the collection of a user's broadcast packets can be employed to pinpoint the user as well [13]. It is worth mentioning that a user cannot change the features of broadcasts easily because the broadcasts are carried out without user intervention by default.

These three features have strong distinguishing power because they contain unique information about a wireless host. In addition, features are non-redundant in that they represent different aspects of a wireless host. To be specific, the MACs feature narrates the behavior of a host's NIC driver, while the other two features focus on the user level activities. The difference between IPs and Broadcasts is that the former emphasizes on the Internet visiting habits, and the latter centers around the application software usage (e.g., office editing or music downloading). Therefore, the combination of these features can be exploited together to identify a masquerader.

3.2 Classifier Construction

The three aforementioned features need to be applied on a classifier in order to identify a masquerader. It has been shown that a simple *naïve Bayesian classifier* is accurate and effective for such type of classification problems [12], [13]. The naïve Bayesian classifier assumes that its underlying features are independent. As it is explained in section 3.1, the three features are independent in that they represent different aspects of a wireless host. Hence, these three features can be put together to construct a naïve Bayesian classifier.

We can build a host specific classifier using data samples pulled out from the its network traffic (i.e., a training data set). To be specific, for each host A, we build a naïve Bayesian classifier C_A based on the three features that are present in its training data set. Given a piece of validation data sample V, C_A returns "true" if it believes V is from A and "false" otherwise. These decisions are made based on the *posterior* probability $Pr[V$ is from $A|V$ has $(c_1, c_2, c_3)]$, where c_1, c_2, c_3 are IPs, MACs, and Broadcasts features, respectively. The posterior probability, $Pr[V$ from $A|V$ has $(c_1, c_2, c_3)]$, is calculated using the following equation:

$$\frac{Pr[V \text{ has } (c_1, c_2, c_3)|V \text{ from } A] \cdot Pr[V \text{ from } A]}{Pr[V \text{ has } (c_1, c_2, c_3)]},$$

where the prior probability $Pr[V$ is from $A]$, the likelihood $Pr[V$ has $(c_1, c_2, c_3)|V$ is from $A]$, and the marginal probability $Pr[V$ has $(c_1, c_2, c_3)]$ are obtained from the training data set. Therefore, for a pre-determined threshold T, C_A returns "true" if and only if $Pr[V$ is from $A|V$ has $c] \geq T$ and "false" otherwise. The threshold T is a design parameter that needs to be decided based on the network scenarios.

Next, we evaluate the performance of our proposed masquerader detection scheme.

4. Evaluation Configuration

In this section, we will illustrate the configuration of the traces used for the evaluation and the evaluation metrics.

4.1 Trace Configuration

There are two network traces used in our study. The first trace is gathered at the 2004 SIGCOMM conference [14]. The second one is collected in the Philips Hall at the Department of Computer Science (CS) at the George Washington University (GWU).

4.1.1 SIGCOMM conference trace

The widely used SIGCOMM conference trace [14] was collected using standard wireless cards in monitor mode with tcpdump-like tools for a span of 5 days. For privacy concerns, IP and MAC addresses are anonymized consistently throughout the entire trace (i.e., there is a unique one-to-one mapping between addresses and anonymous "marks"). Thus, each user is uniquely identified by its anonymous mark. There are over 350 users presented in the trace. This trace is refereed as *SIGCOMM trace* in this paper.

4.1.2 Phillips Hall trace

This trace collection took place within the CS department at the Academic Center building, a large 7-story structure of the George Washington University. The CS department occupies half of the 7th floor (i.e., the Phillips Hall) of the Academic Center building. The traces were collected using standard wireless cards in monitor mode with Wireshark for a span of 14 days. Our wireless cards can receive data from the CS department wireless network, and neighboring wireless networks (such as GWU Wireless, research group networks, and etc). The users of these networks include students, staff and faculty members, residents of school dorms, and even visitors. In total, there are over 700 users presented in the trace. This trace is refereed as *Phillips Hall trace* for future use.

4.1.3 Trace Partition

According to the requirements of the naïve Bayesian classifier, each trace is divided into two portions: the training data set and the validation data set. Additionally, each portion is further divided into pieces that are of one-hour duration. In the SIGCOMM trace, a training data set of 24 one-hour duration pieces (i.e., one-day-duration) is enough to build classifiers for a large percentage of hosts. Therefore, the remainder of the trace (about 4 days) is used as the validation data set. In the case of the Phillips Hall trace, a training data set of 120 one-hour duration pieces (i.e., five-day-span) is good enough to build classifiers for a similar percentage of hosts as that of the SIGCOMM trace. The reason that the Phillips Hall trace needs a longer duration of training data set in that hosts' activities in a campus setting are less intensive compared to a conference setting.

4.2 Evaluation Metrics

The two major metrics used in our evaluation study are as follows: i) The **true positive rate** (TPR) refers to the percentage of the validation data pieces that a legitimate user does not generate but we correctly classify as from a masquerader; ii) The **false positive rate** (FPR) refers to the percentage of the validation data pieces that a legitimate user generates but we incorrectly classify as from a masquerader; iii) TPR variations over time.

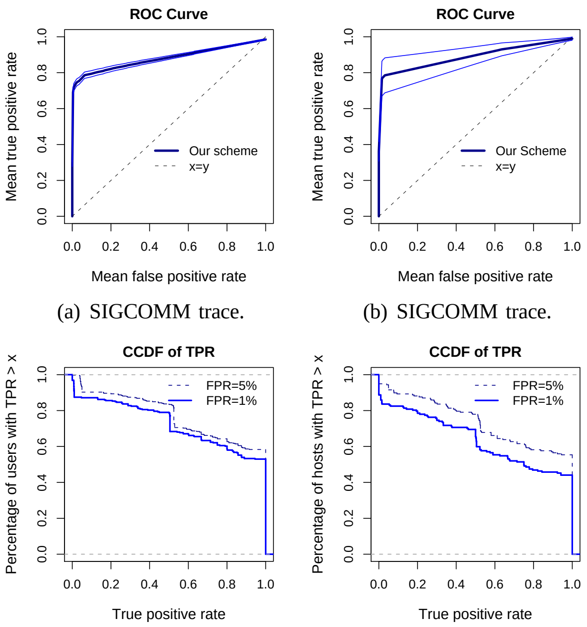
The TPR and FPR represent the *sensitivity* and *specificity* of identifying a masquerader, respectively. The relation between the mean TPR and the mean FPR is known as the receiver operating feature (**ROC**) curve. The TPR variations over time measures whether the proposed scheme can offer consistent performance. We detail our evaluation results below.

5. Evaluation Results

In this section, we first demonstrate the relationship between the detection sensitivity and specificity (i.e., the ROC curve) of the proposed scheme for both traces. Next, we examine whether the proposed scheme can deliver a consistent performance over different time periods.

5.1 TPR vs. FPR

The ROC curve is plotted in Fig. 1(a) and Fig. 1(b) for SIGCOMM trace and Phillips Hall trace, respectively. Standard deviation of the ROC curves is also drawn in these two figures. For reference, the dotted line $x = y$ represents the performance of random guessing. It can be seen that our proposed scheme achieves a good sensitivity and specificity for both traces. For example, in both traces, the mean TPR exceeds 80% when the mean FPR is chosen to be 1%. Such a high level of detection accuracy is fundamentally due to the excellent discriminating power of the three chosen features applied on the naïve Bayesian classifier. It also can be seen that there are higher variations presented in the Phillips Hall trace because it has a more complex user body (over 700 users in total) with rich user dynamics. It demonstrates that our proposed is able to handle very complicated scenarios.



(a) SIGCOMM trace.

(b) SIGCOMM trace.

(c) Phillips Hall trace.

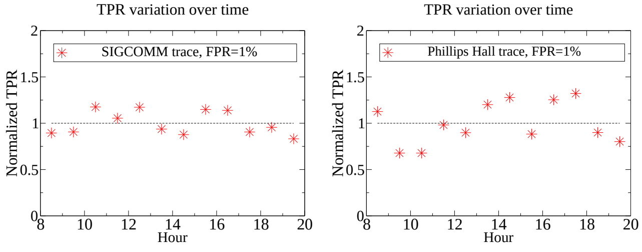
(d) Phillips Hall trace.

Fig. 1: ROC study.

Additionally, the complementary cumulative distribution function (CCDF) of TPR when the FPR is set to be 5% and 1% is shown in Fig 1(c) and Fig 1(d). For an instance, when FPR is 1%, we see that over 60% of masqueraders can be identified at a TPR that is about 80% for the SIGCOMM trace. In the case of Phillips Hall trace, over 50% of masqueraders can be identified at a TPR that is about 80%. This difference is partly due to the rich diversity that lies in the attendees of the SIGCOMM conference. A large portion of the attendees come from different universities, and therefore become distinguishable when accessing their school web and email servers.

5.2 TPR Variations

In this subsection, we examine the TPR variations over time under the two traces. Fig. 2(a) and Fig. 2(b) display the TPR variations over 12 hours span (i.e., from 8am to 8pm) in a day under the SIGCOMM and Phillips Hall trace, respectively. We only consider the 12 daytime hours of a day since there are obviously likely to be more user activities than that of early morning and late night. The mean FPR is fixed to be 1% for both figures. The mean TPR of each hour is normalized by the overall mean TPR for the entire 12 hour period.



(a) SIGCOMM trace. (b) Phillips Hall trace.

Fig. 2: TPR variations over time.

We observe that the normalized TPR values of both traces are close to 1, which demonstrates that our proposed techniques are capable of delivering a consistent performance over time. In addition, the mean TPR under the Phillips Hall trace exhibits higher variations compared to that of the SIGCOMM trace because the former has richer user dynamics in nature.

6. Conclusion

In this paper, we develop a novel masquerader detection scheme for protecting 802.11 wireless networks under varying operating conditions. To the best of our knowledge, it is the first scheme to offer the functionality of sophisticated masquerader detection. The proposed scheme provides all-around protection through an elegant coupling of fingerprints based machine learning techniques. An attractive feature of our proposed scheme is that it requires neither specialized hardware nor modification to existing security standards. Further, it can be implemented in the firmware of existing access point hardware or a computer that is able to capture wireless traffic. Our evaluation study results show that the proposed infrastructure is effective in masquerader detection. As a part of our future work, we plan to develop new features and apply different classifiers so as to further increase masquerader detection sensitivity and specificity, and reduce computation overhead.

References

[1] F. Guo and T. cker Chiueh, “Sequence number-based mac address spoof detection,” in *Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID’05)*, 2005, pp. 309–329.

[2] Q. Li and T. Wade, “Relationship -based detection of spoofing -related anomalous traffic in ad hoc networks,” in *Proceedings of the 3rd IEEE Communications Society on Sensor and Adhoc Communications and Networks (SECON’06)*, 2006, pp. 50–59.

[3] D. B. Faria and D. R. Cheriton, “Detecting identity-based attacks in wireless networks using signalprints,” in *Proceedings of the 5th ACM workshop on Wireless security*, ser. WiSe ’06. New York, NY, USA: ACM, 2006, pp. 43–52.

[4] Y. Sheng, K. Tan, G. Chen, D. Kotz, and A. Campbell, “Detecting 802.11 MAC layer spoofing using received signal strength,” in *Proceedings of the 27th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM’08)*, Apr. 2008, pp. 1768–1776.

[5] G. Chandrashekaran, J. A. Francisco, V. Ganapathy, M. Gruteser, and W. Trappe, “Detecting identity spoofs in 802.11e wireless networks,” in *GC’09: Proceedings of the IEEE Globecom 2009 Communications and Information Security Symposium*, November/December 2009, pp. 1–6.

[6] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, “Wireless device identification with radiometric signatures,” in *Proceedings of the 14th ACM international conference on mobile computing and networking*, ser. MobiCom ’08. New York, NY, USA: ACM, 2008, pp. 116–127.

[7] T. Kohno, A. Broido, and K. C. Claffy, “Remote physical device fingerprinting,” in *Proceedings of the 2005 IEEE Symposium on Security and Privacy*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 211–225.

[8] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Van Randwyk, and D. Sicker, “Passive data link layer 802.11 wireless device driver fingerprinting,” in *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15*, 2006.

[9] M. Raya, J.-P. Hubaux, and I. Aad, “DOMINO: a system to detect greedy behavior in IEEE 802.11 hotspots,” in *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, ser. MobiSys ’04. New York, NY, USA: ACM, 2004, pp. 84–97.

[10] P. Kyasanur and N. H. Vaidya, “Selfish MAC layer misbehavior in wireless networks,” *IEEE Transactions on Mobile Computing*, vol. 4, pp. 502–516, September 2005.

[11] Y. Rong, S. K. Lee, and H.-A. Choi, “Detecting stations cheating on backoff rules in 802.11 networks using sequential analysis,” in *INFOCOM ’06: Proceedings of the 25th IEEE International Conference on Computer Communications*, 2006.

[12] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, ser. SIGMETRICS ’05, 2005, pp. 50–60.

[13] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, “802.11 user fingerprinting,” in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, ser. MobiCom ’07. New York, NY, USA: ACM, 2007, pp. 99–110.

[14] M. Rodrig, C. Reis, R. Mahajan, D. Wetherall, J. Zahorjan, and E. Lazowska, “CRAWDAD data set uw/sigcomm2004 (v. 2006-10-17),” 2006.

Survey: Security Measures and Associated Drawbacks of Wireless Body Area Network (WBAN)

Iehab A. AL-Rassan , Naveed Khan

Department of computer Science, College of Computer & Information Sciences
King Saud University, P.O.Box 51178 Riyadh 11543,Saudi Arabia

irassan@ksu.edu.sa : naveed@ksu.edu.sa

Abstract—*Body sensor network is a critical life-saving infrastructure, which could maintain complete privacy of an individual. Body Area Network (BAN) are widely used in health care, military, sports and first responders. Recently, BAN security and its energy consumption has been an interesting research topic to monitor human body, which enable doctors to predict, diagnose and react to situations effectively and in time. By securing BAN, we are actually securing human life. Numerous intruders and attacks may challenge the system and can lacerate the data that can be used for destructive purpose. This paper review BAN security and its energy conservation. This review classifies BAN security techniques and their brief comparison along with other critical issues which needs to be studied.*

Keywords – BAN (Body Area Network), ECG/EKG (Electrocardiography), SG (Secure Group)

1. Introduction

BAN is an emerging technology, which is widely used for research in literature, military, healthcare, sports and first responder, etc [1, 2,3]. It consists of wearable intercommunicating sensors, as shown in Fig 1**.** The sensors are wirelessly connected onto human body for monitoring body movement and measuring physiological parameters such as body temperature, cardiac motion and so on [4] They collect data and send it to the base station through wireless multi-hop network [1] for analysis, storage and processing. The data are then sending in a secure manner to remote medical servers through internet or other communication media. Security is important because the data may contain sensitive information obtained from physiological values hence providing security to the inter-sensor communication which is more essential [5]. In fact, the requirement for protecting health data is legal as from the health insurance policy and accountability act (HIPAA) [6], which authorize that all individual personal information should be protected and safe from unauthorized access. The main purpose of BAN is to secure human life because it deals with human body and its physiological values to securely distribute cryptographic key in the BAN [7] which may enable us to handle emergency situation(specific event of heart attack) in correct time [8].

This paper review the current BAN technologies which implement various types of security and energy consumption measures, evaluate the possible attacks and provide solutions to achieve the possible goals in order to provide the desired security and energy conservation (as minimum as possible) because wireless devises remain alive on their battery life. Furthermore, it summarizes BAN security techniques in each class to provide a brief comparison among these techniques. The remainder of the paper is organized as follows: section II discusses BAN security techniques , track and analyzed four classes of BAN Security approaches such as electrocardiogram (EKG) based, an improved EKG based, Secure Group (SG) Communication based and Physiological Signal based were discussed respectively. Section III provides a brief comparison of BAN security approaches where as section IV identifies the problems in each BAN technique and proposes solution for it. The survey concludes in Section V

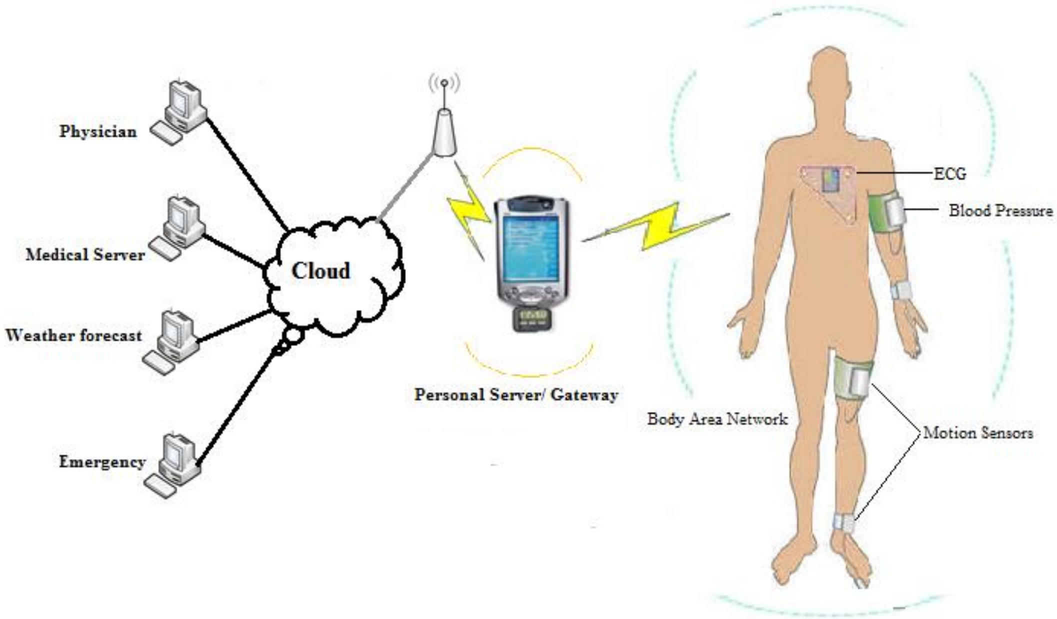


Fig.1 -Architecture of a typical Wireless Body Area Network

2. BAN Security Measures

1.1 EKG based key agreement scheme in Body Area Networks

The EKG based key agreement is briefly described by Venkatasubramanian et al [9]. The author proposes a scheme, which produces identical keys from the ECG/EKG signal. The scheme works by first generating (processing phase) the features from the ECG input using fast Fourier transform (FFT). The generated features are then gathered in the form of blocks and concatenated horizontally to form feature vector. Finally the quantization of the generated is performed. In quantization the signal is just converted to digital form. In the key generation phase the blocks are exchanged for the sake of generating a common key. The blocks are hashed before sending or exchanging using SHA-256 hashing algorithm. The exchange phase is called the commitment phase in which the hashes received are arranged in matrix where the local hashes are further arranged in to U- matrix and those received from the sensors other then commitment phase are arranged in V- matrix. From the hamming distance of U & V a matrix W is computed such as that every element of W is a hamming distance. At the end of the processing phase both the sensors have the keys, which are identical on both sides. In the third phase which is the de-commitment phase the integrity of the received blocks are checked by using message authentication codes (MAC).

1.2 An improved EKG based key agreement scheme for body area sensor networks

In this work the authors proposed [10, 11] an improved version of the EKG based key agreement scheme. The first problem is solved by replacing the FFT by DWT because the time complexity of FFT is O(nlgn) while that of DWT is O(n). So in this technique, the feature generation is performed by using the DWT, the generated features are gathered in the form of blocks and then horizontally concatenated to form feature vector (as in the original work) [4] and then the quantization of the generated is performed. In quantization the signal is just converted to digital form [4]. In the key generation phase before the blocks are exchanged, the blocks were first hashed by using SHA-256 and then watermark was included in the hashed blocks by using the finger prints as a seed to the random machine. The random numbers generated by the random machine were used as the locations for the watermarks to be embedded on the sender side and the same position are used to remove the watermark from the blocks. On the receiving end first of all the watermark is removed by using the position generated by fingerprints seeded random machine and the pure hashes of the blocks are gathered. These hashes received and local hashes are arranged in a matrixes U and V respectively. A matrix W was computed from the hamming distance of U & V in such a manner that every element of W was a hamming distance. At the end of the processing phase both the sensors have the keys, which were identical on both sides. In the third phase, which was the de-commitment phase, the integrity of the received blocks was checked by using message authentication codes (MAC).

1.3 Secure Group (SG) Communication for BAN

The main focus of this technique was to make communication secure in inter-ban and intra-ban. It includes group server and multiple ban's. The group server is responsible for issuing group key, revoking, renewing and management [12,13]. A central server with sufficient resources receives data from multiple bans to make the necessary computational tasks e.g. encryption, decryption, and data processing. The scalable group key management was based on keystone. In [14] keystone, the identity of client authentication was uploaded to one or more registrar to improve its performance. For multicast delivery keystone used user datagram (UDP), internet protocol (IP) to make key updates more efficient and reliable using forward error correction (FEC) to reduce loss of messages. SG-BAN use keystone based group key because of its scalability, extensibility and robustness. The group server consists of registrar, key server and control manager. The registrar initiates itself with a registrar key Kr and client access control list from key server. ACL contains information about access and permission about clients. When client join the network first it will be register with the registrar who verify its client MAC address by checking ACL. Now to make the group communication secure a session key is granted to client for encryption/decryption of communication [15]. It also provide facility for revoking and renewing of key to ensure maximum security like if a client leave the network and fail to respond it group key will be discarded and if he join again his key will be renewed by the control manager. In intra BAN communication each device is assigned a key Kd, which is 4 digit of the group key in BAN plus device MAC address.

1.4. Physiological Signal based key agreement and energy Conservation

This technique plays a very helpful role in BAN because it mainly concerns with two approaches; one to minimize the energy consumed by wireless instrument and other to provide security for wireless devices while they exchange information using cryptographic technique. It uses photoplethysmogram (PPG) signal based features to enable two sensors to agree on a common key. For sharing a common key PPG is synchronized in a loosely manner and a range of frequency domain features are generated and FFT is perform on the PPG signals and then perceive the peak in the FFT [16] coefficients. Finally features are derived from peaks by producing tuples each of which capture a peak-value and its corresponding peak index and convert to form a feature vector. After generating these features the communication sensor on sender side generate a random symmetric key using 128 bits, which then hides using the feature vector obtained from PPG signal and to make it applicable a fuzzy vault cryptographic primitives [17, 18] is used to hide the key. Few steps are taken to accomplish this task.

a) Sender generates vth order polynomial [5]

b) Computes the polynomial at each point in feature vector

c) Set of legitimate coordinates of x, y coordinates by adding a large number of bogus, random coordinates called chaff point [5]

This set of legitimate points and chaff points are called elements of vault. This vault is transfer to the receiver sensor through wireless medium. The receiver side upon receiving vault identify the v+1 elements from vault whose x coordinate values, are identical to its own feature values and find polynomial values by langrangian interpolation., once generated the correct polynomial sends back an acknowledgment. The advantages of this technique are; it is impossible to hacking and brute force it and the second issue described here is energy consumption of PKA, which is divided into two parts; computational energy cost and communication energy. The former is consumed during the execution of PKA while the latter one consumed energy while transmitting and receiving the vault. Traditionally in sensor networks communication costs overwhelming the computational cost but due to considerable processing requirement of PKA in terms of FFT computation, feature generation, polynomial generation, evolution and langrangian interpolation. It is believed that the computational cost is a considerable portion of the total energy. So there is a trade-off between security and energy. So the energy problem is solved through the energy scavenging techniques [19,20]. Normal PKA consumed 57mW. The scavenge energy can be obtained from (1) body Heat : The neck braces scavenge energy from the latent heat of vaporization occurring due to the vaporization of the perspiration of the individual. The estimated power gain is 0.2 W to 0.32 W with this approach, which is sufficient for PKA. (2) Respiration: One of the prominent movements is due to respiration. It is estimated that about 420 mW of power can be extracted from a stretchable dielectric elastic band worn around the chest of an individual. But can be extracted only from heavy breathing. This amount of power is sufficient for PKA. (3) Ambulation: Demonstrated systems have been successful in recovering 1.5 W of power from the human arm motion and more than 1.6 W of power from the ambulatory motion of human beings (using piezo-electric soles in shoes), which is sufficient for PKA.[2] Photovolatic Cells.: Photovoltaic cells can produce 100 mW/cm2 of power when under sunlight, which is again sufficient for PKA.

3. Comparison of BAN security and energy Conservation Techniques

This section provides a brief comparison of BAN techniques. In Table 1, the BAN security and energy conservation approaches based on key features including; ability to authenticate, secure communication using encryption, minimum energy consumption, accuracy and safety etc, were compared.

Table1. Comparison of different key features of BAN

Techniques	Ability to Authenticate	Attack	Secure Communication	Energy Consumption	Accuracy	Safety
EKG Based	X	√	X	√	X	X
An Improved EKG based	√	√	√	√	X	X
Secure Group (SG) Communication based	√	√	√	√	√	X
Physiological Signal based	√	X	√	√	√	X

As shown in this table, EKG based technique can only do Integrity where as other techniques are able to provide secure communication as well as less energy consumption. However the other approaches provides a few key features regardless of other important factors which can also effect human body in Body area sensor networks. Among all security and energy consumption approaches, the only approach that allows real time and more effective communication is physiological signal Based because it uses Energy scavenging techniques [19]. However the, active countermeasures run the risk of false measure. The most recent BAN secure approaches based on cryptographic as well as physiological signal based approach in [16] provide promising tradeoff between security and energy. But the safety issue of human body tissue is still unclear, which needs to be studied by the researchers. Apart from minimum energy consumptions and safety of human tissue with very low false rates, these approaches can provide secure communication.

4. Discussions

Different types of techniques are used to overcome the security and energy conservation problems in wireless BAN. In the analysis phase of EKG based key agreement scheme two major problems were found; (1) the use of FFT based feature generation process, the drawback of which is that its time complexity is O(nlogn) which is too expansive as the sensors have the problem of having limited resources in both memory and processing and (2) in the commitment phase where the blocks were exchanged by just applying a hash function and the values generated by using the hash function are 64 bits can be easily attacked by brute force. We can solve this problem if Elleptive Curve cryptography is used as well as replace FFT with DWT whose complexity is O(logn) .Similarly second technique also encounters some problems i.e. this technique uses the fingerprints as a seed, where the fingerprints are constant means and these values are not time variant like EKG, and if these fingerprints are captured the technique will fail. We can solve this problem by using other biometrics like iris print, face, voice etc. as a seed to the random machine instead of finger prints. Moreover the third technique has the problem of low transmission efficiency, because the overhead of secure communication and mutual authentication is too high. This problem can be solved by implementing a more efficient key scheme such as quantum cryptography to reduce the overhead as well as provide better security. In last technique the security problem has been solved up to possible extent by using 128 bit symmetric key as well as provides solution to fulfill the required energy consumption by using scavenging techniques. The problem arises here is of safety because of heavy computation heat produced, which is harmful for human tissues as to propose a safe way to minimize this problem as much as possible. This technique needs to be studied more to find a suitable solution the extensive heat generated during computation.

5. Conclusion

Body sensor has a lot of applications in life-saving infrastructure with ensured privacy. Security and energy conservation in BAN is very challenging task. In this survey paper four techniques EKG, enhanced EKG, physiological signal based and SG-Group communication were reviewed, which tried to tackle the problems associated with BAN. The proposed techniques have also encounters some problems. i.e. in EKG, FFT has the time complexity o(nlogn), is too expensive and only provides integrity. Enhanced EKG based technique has the problem of fingerprints, which can be hacked easily where SG-Group and physiological signal based has the problems of transmission overhead and safety, respectively. The solutions for all the proposed techniques were described in detail. We believe that this study will prove a potential step forward in developing more secure and less energy consumptive BAN.

References

[1] K.Venkatasubramanian, G. Deng, T. Mukherjee, J. Quintero, V. Annamalai, and S. K. S. Gupta. Ayushman.: A Wireless Sensor Network Based Health Monitoring Infrastructure and Testbed. In Distributed Computing in Sensor Systems, pages 406{407, July 2005.

[2] K. Hung and Y.T. Zhang. Implementation of a wapbased telemedicine system for patient monitoring. IEEE Transactions on Information Technology in Biomedicine,7(2):101–107, 2003.

[3] R.S.H. Istepanian, E. Jovanov, and Y.T. Zhang. Guest editorial introduction to the special section on m-health:Beyond seamless mobility and global wireless healthcare connectivity. IEEE Transactions on Information Technology in Biomedicine, 8(4):405–414, 2004.

[4] E. Jovanov, A. Milenkovic, C. Otto, and P.C. Groen. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. Journal of NeuroEngineering and Rehabilitation, 2005.

[5] .Venkatasubramanian, A. Banerjee, and S. K. S. Gupta.: Plethysmogram-based Secure Inter-Sensor Communication in Body Area Networks. Pages 1–7, November 2008. In Proc. of IEEE Military Communications Conference.

[6] HIPAA-Report 2003, “Summary of HIPAA Health Insurance Probability and Accountability Act,” US Department of Health and Human Service, May 2003.

[7] D. Singel, B. Latr, B. Braem, M. Peeters, M. D. Soete, P. D. Cleyn, B. Preneel, I. Moerman, and C. Blondia.: A secure cross-layer protocol for multi-hop wireless body area networks. In 7th International Conference on AD-HOC Networks and Wireless, September 2008.

[8] S. K. S. Gupta, T. Mukherjee, and K. Venkatasubramanian, “Criticality Aware Access

Control Model for Pervasive Applications," March 2006,pp. 251–257, In Proc. of 4th IEEE Conference on Pervasive Computing Pervasive Computing and Communications.

[9] K. Venkatasubramanian, A. Banerjee, and S. K. S. Gupta. EKG-based Key Agreement in Body Sensor Networks. April 2008. In Proc. of the 2nd Workshop on Mission Critical Networks.

[10] Aftab Ali, Farrukh A. Khan, "An Improved EKG-based Key Agreement Scheme for Body Area Networks", WNS Workshop, International Conference on Information Security and Assurance (ISA 2010), Miyazaki, Japan.

[11] E. M. Yeatman. Rotating and gyroscopic mems energy scavenging. Wearable and Implantable Body Sensor Networks, International Workshop on, pages 42–45, 2006.

[12] E. Jovanov, A. Milenkovic, C. Otto, and P.C. Groen. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. Journal of NeuroEngineering and Rehabilitation, 2(6), 2005

[13] A. Dunkels, F. Osterlind, N. Tsiftes, and Z. He. Demo-software-based sensor node energy estimation. pages 409–410, Nov 2007. In Proc. 5th ACM Conference on Embedded Networked Sensor Systems.

[14] Yu Cai and Jindong Tan "Secure Group Communication in Body Area Networks" Proceedings of the 2008 IEEE International Conference on Information and Automation June 20 - 23, 2008, Zhangjiajie, China, International Conference on Information and Automationv , Proceedings of the 2008 IEEE ,June 20 -23, 2008, Zhangjiajie, China

[15] S. Zhu, S. Setia, and S. Jajodia. LEAP+: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks. ACM Trans. on Sensor Networks (TOSN), 2(4):500–528, Nov 2006.s

[16] K. K. Venkatasubramanian, A. Banerjee, S. K. S. Gupta "Green and Sustainable Cyber-Physical Security Solutions for Body Area Networks" IMPACT Lab (http://impact.asu.edu)
School of Computing and Informatics ,Arizona State University,Tempe, Arizona 85287, BSN2009 ,{kkv,abanerj3,sandeep.gupta}@asu.edu

[17] A. Juels and M. Sudan. A Fuzzy Vault Scheme. page 408, 2002. In Proc. of IEEE Intl. Symp. on Inf. Theory.

[18]A. Banerjee, K. Venkatasubramanian, and S. K. S. Gupta. Challenges of Implementing Cyber-Physical Security Solutions in Body Area Networks.April 2009. In Proc. of 4th International Conference on Body Area Networks (Accepted for Publication).

[19] J. A. Paradiso and T. Starner." Energy scavenging for mobile and wireless electronics". Pervasive Computing, IEEE, 4(1):18–27, Jan.-March 2005.

[20] E. Yeatman, D. O'Hare, C. Dobson, and E. Bitziou. Approaches to selfpowered biochemical sensors for in-vivo applications. In Proc. of 3rd International Conference on Body Area Networks, pages 1–2, 2008.

Empirical Channel Model for 2.4GHz IEEE 802.11 WLAN[*]

Stanley L. Cebula III, Aftab Ahmad, Jonathan M. Graham, Cheryl V. Hinds, Luay A. Wahsheh, Aurelia T. Williams, and Sandra J. DeLoatch
Information Assurance Research, Education, and Development Institute (IA-REDI), Norfolk State University (NSU), Norfolk VA USA

Keywords: WiFi, IPsec, WLANs, Information Assurance, Channel Models

Abstract - *In order to design robust systems for security and forensics of Wireless LANs (WLANs), real-time channel measurements are imperative. We introduce a WLAN Forensics (WiFo) system that uses a grid of WiFi sensors to generate a real-time channel model as well as locate each user within one of the grid areas surrounded by four sensors (or two sensors and the access point). With the help of measurements spread over an extended period of time, we show that these real channel profiles do not follow any recommended model. We conclude that only real-time, empirical channel models can benefit a WiFo system, where users are required to be identified with respect to their location and the signal needs to be contained for environments such as classified ones.*

1 Introduction

One way to look at the extent of security provided in the IEEE 802.11 standard is by comparing it with a more trusted protocol suite, such as IPsec. As reported in [1], WiFi is not quite a match for the flexibility and robustness of IPsec. Specifically, there is no security on the physical layer of WiFi. With wired LANs, one can physically trace each packet's source and destination machine. This is the definition of privacy in a LAN environment. However, there is no way to physically trace a packet on a WiFi network to see the source or destination machine. Also, wired LANs contain physical connections that can be controlled (via physical cable). There is no way provided to control or view the WiFi signal in the 802.11-2007 standard. In classified environments, signal spilling can occur when a WiFi signal is transmitted further than intended. This makes it possible for attacks to occur outside of the physical building where the WiFi network is operated. A tool that allows system administrators to view each machine connected to the WiFi network on a map would provide the ability to identify attackers, thus increasing security. Such a tool will create a map of the signal strength of the WiFi network. With the help of a power control loop, it can be used to restrict the WiFi signal to desired physical boundaries. In order to map the signal strength of a WiFi network, a real-time channel model is required to predict the signal strength at any given distance. Such a system is shown in Figure 1. The system consists of a grid of WiFi sensors that provide signal strength feedback to the access point (AP) to determine a LIVE channel profile. Such a system can also be used to locate users within the grid, thus providing the same level of privacy as a wired LAN would provide. In this paper, we report on research results of one aspect of this system. Instead of using WiFi sensors we based our channel model on actual measurements in a lab that was moderately populated with students and faculty using regular computing equipment. The measurements were made for several days on fixed locations around an Actiontec GT704WG access point. Besides providing our own empirical channel model, we will also compare it with popular existing models.

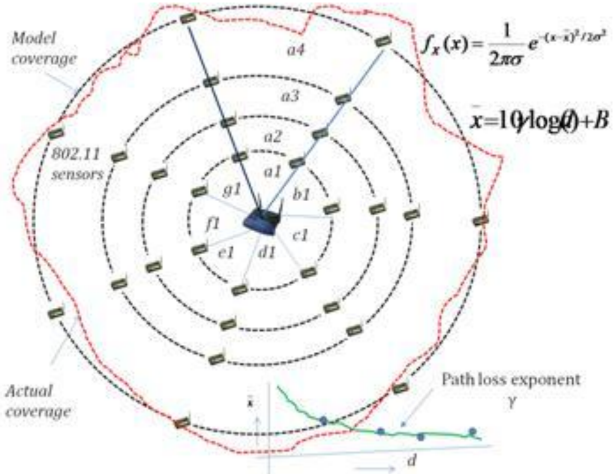


Figure 1. WiFo System Architecture

1.1 WiFo system model

We introduce a new term, WiFo system, in this paper. WiFo is an acronym for WiFi Forensics. The system architecture is shown in Figure 1. The WiFo system consists of an access point surrounded by a grid of low-power, inexpensive (low-memory, low-processing) WiFi sensors. There are two main algorithms that run in the AP: one for plotting and consequently controlling the signal power coverage and the other for location determination. These algorithms can take two types of inputs from each sensor:

(i) the signal level of the sensor itself, and

(ii) a list of user stations (STA) sorted by the RSSI as seen by the sensor.

The quantity in (i) is employed in obtaining a LIVE signal coverage map by measuring the signal path loss exponent and modeling the signal power, either using an empirical probability density function (pdf) of the signal or using a well-known distribution, such as *lognormal*, as shown in the Figure 1. The purpose of this paper is to report one aspect of this system, the empirical channel model.

The remainder of this paper is organized as follows: obstacles that affect signal strength are discussed in Section 2. Existing channel models are outlined in Section 3. We discuss how we found the path loss and developed our channel model in our environment in Section 4. Lastly, we conclude the paper in Section 5.

1.2 Obstacles

In any type of WiFi signal transmission, the output signal from the transmitting STA or AP will differ from the signal that is received. There are many factors that affect the signal while it is in transit. These include attenuation, free space loss, fading, reflection, diffraction, scattering, refraction, and noise. Attenuation occurs when the strength of a signal falls off with distance [2]. Basically, the further the signal travels, the weaker the signal will get. This can be represented logarithmically [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The rest of this section has made heavy use of [2]. Free space loss is a form of attenuation that means the signal disperses with distance. In other words, the further the signal travels, the more the signal spreads out in other directions. The spread of the signal makes the signal weaker. When variation of the signal power occurs due to changes in the transmission medium or path, fading occurs. Basically, any interruption in the transmission medium (atmospheric changes) or path (objects) can affect the strength of the signal. Reflection exists when the signal bounces off large objects causing the signal to change. These changes can increase or decrease the signal strength. This usually happens when the signal reflects off walls, floors, or ceilings. Diffraction is produced when the signal runs into a large object. The secondary waves resulting from the obstructing surface are present throughout the space and behind the large object negatively affecting the strength of the transmitted signal. This can occur when the signal runs into a wall partition or cubicle. Scattering exists when the transmitted signal passes through many small objects that cause the signal to go in many different directions. Scattered waves are produced by rough surfaces, small objects, or by other irregularities in the channel. Refraction is defined as a change in direction of a transmitted signal resulting from changes in velocity. This usually occurs when only part of the line of sight transmitted signal reaches the destination. Noise can be characterized as various distortions imposed by the transmission medium or additional unwanted signals. Noise is usually caused by interference or reception of unwanted signals from other electronic devices. Due to the large number of obstacles that affect the strength of a transmitted WiFi signal, the channel models used to represent the environment must be very specific to each environment.

2 Some Existing Pathloss Models

The channel models that are discussed in this section include the Okumura-Hata model, Log-distance Path Loss model, and JTC Indoor Path Loss model. We will briefly describe each model and determine its suitability for the environment under consideration.

2.1 The Okumura-Hata model

The Okumura-Hata model is a combination of the Okumura model and empirical models developed by Masaharu Hata [3]. The Okumura-Hata model is represented below:

$$\begin{aligned} &L_{50}(urban)(dB) \\ &= 69.55 + 26.16\log(f_c) - 13.82\log(h_{te}) - a(h_{re}) \\ &+ (44.9 - 0.55\log(h_{te}))\log{(d)} \end{aligned}$$

where L_{50} is the 50th percentile median path loss, f_c is the centre frequency in megahertz, h_{te} is the base station antenna height in meters, h_{re} is the receiver station antenna height in meters, $a(h_{re})$ is a vehicular station antenna height-gain correction factor that depends on the environment, and d is the link distance in kilometers [3, 4, 5].

The Okumura-Hata model is extremely accurate, because it is based on measurements in a specific environment. However, while the Okumura-Hata model is popular and accurate, it is mainly used in outdoor, urban environments [3, 4, 5]. The Okumura-Hata model would work well if we were determining path loss in network located outdoors. We should not use the Okumura-Hata model in the development of our signal strength monitoring system, because we are conducting measurements inside an office environment.

2.2 Log-distance pathloss model

The Log-distance Path Loss model is a very popular logarithmic model that is based on a linear dependence between the path loss in decibels and the logarithm of the distance between the transmitter and receiver [2, 4, 6, 7, 8]. This model predicts path loss inside a building or in densely populated areas. There also exist many studies that use a variation of the Log-distance Path Loss model [3, 5, 8, 9, 10, 9, 12]. The Log-distance Path Loss model is represented below:

$$PL(d)(dB) = PL(d_0)(dB) + 10n\log(d/d_0) + X_\sigma$$

where $PL(d_0)(dB)$ is the measured path loss in decibels one meter from the transmitted signal, n is a path loss exponent dependant on the surroundings and building type, d is the distance between the transmitter and receiver in meters, d_0 is typically one meter, and X_σ is a normal (Gaussian) random variable in decibels that has zero mean and standard deviation of σ decibels [2, 4, 6, 7]. This model also takes into consideration different obstacles in the transmitter to receiver path (also known as log normal shadowing). Table 1 lists the path loss exponents based on different environments [2, 5, 12].

Environment	**Path Loss Exponent, n**
Free Space	2
Urban area cellular radio	2.7 to 3.5
Shadowed urban cellular radio	3 to 5
In building line-of sight	1.6 to 1.8
Obstructed in building	4 to 6
Obstructed in factories	2 to 3

Table 1. Log-distance Path Loss Exponent

According to many studies which used the Log-distance Path Loss model or a variation of this model, the Log-distance Path Loss model is accurate and simple to use [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The Log-distance Path Loss model also will work in our environment and could be used in the development of our signal strength monitoring system.

2.3 JTC indoor pathloss model

The JTC (Joint Technical Committee) Indoor Path Loss model is the official path loss model for office environments presented by the International Organization for Standardization (ISO). This model has traits from the Okumura-Hata model (based on specific measurements of different factors) and the Log-distance Path Loss model (based on the relationship between the logarithm of the distance between the transmitter and receiver to the path loss in decibels). The JTC Indoor Path Loss model is represented below:

$$L_{Total} = A + B\log_{10}(d) + L_f(n) + X_\sigma$$

where A is an environment dependent fixed loss factor in decibels, B is the distance dependent loss coefficient, d is the distance between the transmitter and receiver in meters, L_f is a floor/wall penetration loss factor in decibels, n is the number of floors/walls between the transmitter and receiver, and X_σ is a normal (Gaussian) random variable in decibels that has zero mean and standard deviation of σ decibels (log normal shadowing) [3, 5, 13]. Table 2 contains the corresponding variables dependent on the type of environment [3, 5, 13].

Environment	**Residential**	**Office**	**Commercial**
$\boldsymbol{A(dB)}$	38	38	38
$\boldsymbol{B}$	28	30	22
$\boldsymbol{L_f(n)(dB)}$	4n	15 + 4 (n-1)	6 + 3 (n-1)
Log Normal Shadowing Std. Dev. $\boldsymbol{(dB)}$	8	10	10

Table 2. JTC Indoor Path Loss Model Variables

The JTC Indoor Path Loss model will work in our environment. According to [3], the JTC Indoor Path Loss model may be more accurate than the Log-distance Path Loss model due to the addition of the $L_f(n)$ function. The JTC Indoor Path Loss model could be used in the development of our signal strength monitoring system.

The Okumura-Hata, Log-distance Path Loss, and JTC Indoor Path Loss models are all accurate and reliable in different environments. Based on our environment of an indoor office setting, the Log-distance Path Loss and JTC Indoor Path Loss models could be used in the development of our signal strength monitoring system.

3 Developing a Channel Model

In order to develop our own channel model, we took many measurements in our custom environment and find the predicted value line based on our data. This predicted value line acts as a path-loss model in our environment. The area that we used for testing and developing our own channel model is a portion of the Information Assurance Research, Education, and Development Institute (IA-REDI) located on the sixth floor of the Marie V. McDemmond Center for Applied Research (MCAR) at Norfolk State University. This area is a computer lab (approximately twenty feet by sixty-five feet) next to an office, three conference rooms, and one long hallway. We used commercially available hardware and software. The access point we used was an Actiontec GT704WG router on default settings. The wireless card we used to connect to the access point was an Intel® PRO/Wireless 3945ABG Network Connection built-in a Dell Latitude D830 laptop running Windows XP on default settings. The program we used to measure the signal strength is called inSSIDer (freeware) [14]. We measured the signal strength at eighteen locations on every hour between 9:00 a.m. and 5:00 p.m. (EST) for one week. Figure 2 displays the experiment area, measurement locations, and access point (AP). The access point is the circle in the upper right hand corner. The letter 'x' represents the measurement locations, and the arcs represent measured distances (ten feet) from the access point. Table 3 contains the measurement locations and their distance from the access point.

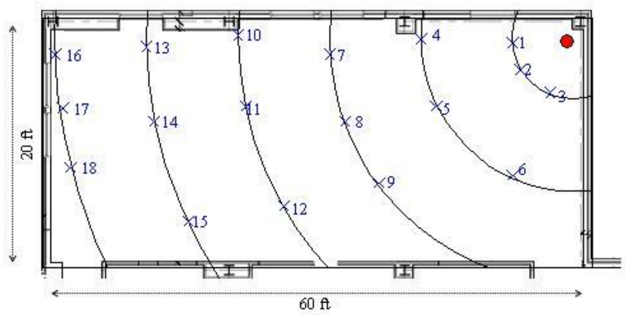


Figure 2. Environment Overview

Location	Distance from AP (ft)	Distance from AP (m)
1	10	3.408
2	10	3.408
3	10	3.408
4	20	6.816
5	20	6.816
6	20	6.816
7	30	10.224
8	30	10.224
9	30	10.224
10	40	13.632
11	40	13.632
12	40	13.632
13	50	17.04
14	50	17.04
15	50	17.04
16	60	20.448
17	60	20.448
18	60	20.448

Table 3. Distance of Locations from AP

The access point was sitting on a desk approximately three feet from the ground. When measuring the signal strength, the laptop was held approximately five feet from the ground with the screen facing away from the access point. The program inSSIDer was used to collect data, and a total of 810 measurements were taken over a one-week time period.

Displayed in Figure 3 is the cumulative distribution function (CDF) of the measured signal strength versus distance. Figure 3 shows the data collected (displayed in Figure 4) can be classified as normal.

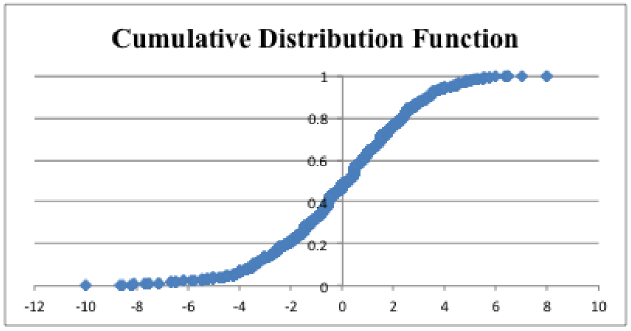


Figure 3. Cumulative Distribution Function of Collected Data

Table 4 represents the average signal strength compared to distance from the access point. Figure 4 displays this relationship along with a predicted value line.

Distance from AP (ft)	Average RSSI (dB)	Variance
10	-28.82	7.55
20	-33.89	6.93
30	-37.31	7.4
40	-40.77	7.51
50	-42.44	5.14
60	-47.61	12.42

Table 4. Average Signal Strength Compared to Distance

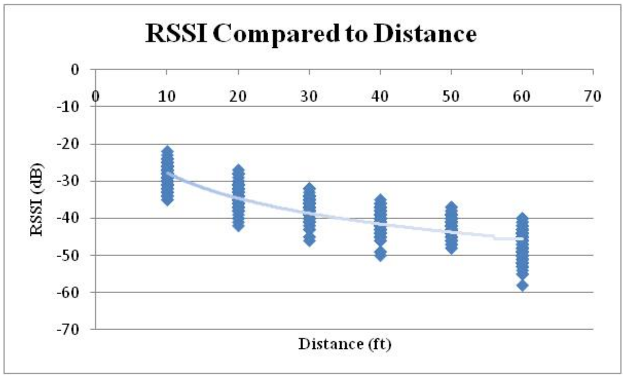


Figure 4. Signal Strength Compared to Distance

The equation for the predicted value line is:

$$RSSI = (-2.1) \cdot 10 \log(d) - 7$$

where $RSSI$ is the predicted signal strength and d is the distance from the access point in feet. This equation could serve as a channel model for this specific environment.

4 Conclusion

The research presented in this paper identified three existing channel models and one new channel model to consider for implementation in this signal strength monitoring system. The new empirical model gives a path-loss exponent of about 2.1, which is closer to the free-space path-loss exponent. In most experiments by other researchers, this value is much higher than the one we obtained. Since custom models can be more accurate in custom environments, a permanent, universal model may not be employed in designing signal mapping systems for 2.4 GHz IEEE 802.11 networks. Upon further review, we also conclude that one should not use the Okumura-Hata model, because we are conducting measurements inside an office environment. The Log-distance Path Loss model and JTC Indoor Path Loss model also will work in our environment, because they are tailored to be universal models for indoor use. However, upon completing the development of our own model (predicted value line from the measurements), a real-time model would be the most accurate to use in the same environment – as proposed in Figure 1 and Figure 4.

References

[1] Cebula, S. L., Ahmad, A., Wahsheh, L. A., Graham, J.M., Williams, A. T., and DeLoatch, S.J., "How Secure is WiFi MAC Layer in Comparison with IPsec for Classified Environments?". In Proceedings of the 14th Communications and Networking Simulation Symposium, April 2011.

[2] Tummala, D. "Indoor Propagation Modeling at 2.4GHz for IEEE 802.11 Networks". M.S. Thesis, University of North Texas, 2005.

[3] Pahlavan, K. and Levesque, A. H. "Wireless Information Networks". Wiley-Interscience. New York, NY, 1995. 73-112.

[4] Vig, J. "ISM Band Indoor Wireless Channel Amplitude Characteristics: Path Loss vs. Distance and Amplitude vs. Frequency". M.S. Thesis, Ohio University, 2004.

[5] Tipper, D. "Wireless Communication Fundamentals". University of Pittsburgh lecture. 2005. 40-42.

[6] Faria, D. B. "Modeling Signal Attenuation in IEEE 802.11 Wireless LANs". Stanford University, July 2005.

[7] Akl, R., Tummala, D., and Li, X. "Indoor Propagation Modeling at 2.4 GHz for IEEE 802.11 Networks". In Proceedings of the 6th IASTED International Multi-Conference on Wireless and Optical Communications, Banf, AB, Canada, 2006.

[8] Borrelli, A. et al. "Channel Models for IEEE 802.11b Indoor System Design". In Proceedings of IEEE Conference on Communications, vol. 6, 2004. 3701-3705.

[9] Andrade, C. B. and Hoeful, R. P. F. "IEEE 802.11 WLANs: A Comparison on Indoor Coverage Models". In Proceedings of the 23rd Canadian Conference on Electrical and Computer Engineering, 2010.

[10] Capulli, F, et al. "Path Loss Models for IEEE 802.11a Wireless Local Area Networks". In Proceedings of the 3rd International Symposium on Wireless Communications Systems, 2005.

[11] Liechty, L. "Path Loss Measurements and Model Analysis of a 2.4 GHz Wireless Network in an Outdoor Environment". M.S. Thesis, Georgia Institute of Technology, 2007.

[12] Phaiboon, S. "An Empirically Based Path Loss Model for Indoor Wireless Channels in Laboratory Building". In Proceedings of the IEEE TENCON'02, 2002.

[13] Joint Technical Committee of Committee T1 R1P1.4 and TIA TR46.3.3/TR45.4.4 on Wireless Access, "Draft Final Report on RF Characterization," Paper No. JTC(AIR)/94.01.17-238R4, Jan. 17, 1994.

[14] http://www.metageek.net/products/inssider

[*] *Acknowledgement:* "This material is based upon work supported by the Department of Energy National Nuclear Security Administration under Award number DE-FG52-09NA29516/A000." *Disclaimer:* "This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof."

Throughput Analysis using Smartphone in WiFi Network

Woojin Sohn, Saleem Aslam, and Kyung-Geun Lee
Department of Information and Communication Engineering, Sejong University
98 Gunja-Dong, Gwangjin-Gu, Seoul, Korea
wjsohn@nrl.sejong.ac.kr, saleem83@nrl.sejong.ac.kr, kglee@sejong.ac.kr

Abstract - *We measure the network performance and study the most critical elements which can influence the throughput. This paper analyzes the throughput variation using Android based smartphone under the assumptions that the rooms have the same size and their area is limited small. We evaluate the performance of network on the basis of different factors like the distance between APs, number of APs, deployment of APs and channel assignment of APs. Experimental results show that interference due to channel overlapping is the most critical factor for network degradation. We also observe that when more than two APs operate in limited small area then their mutual interference becomes the crucial factor for performance degradation of network.*

Keywords: Smartphone, Access Point, Network Performance, Throughput.

I. Introduction

Recently the need of WiFi network increases due to the massive dissemination of smartphones and popularization. Mobile carriers take advantage of WiFi network as their marketing strategy. Accordingly, the use of smartphone in WLAN is increasing in proportion to the increase of Access Point (AP). However, APs installed at dense public areas such as concatenated shops and apartments can degrade the performance of WLAN due to interference of adjacent APs. Smartphone drastically changes the ways of information use and communication mechanism. Recently the increasing number of smartphones is attributable to 3G mobile radio network like WCDMA, Wibro and the fast infrastructural completion of wireless network and data fare policy. Accordingly, the use of AP has rapidly increased to expand the coverage of wireless communication network. The users of WiFi network are growing exponentially not only in public areas but also in private offices and homes. Individuals as well as companies are more inclined in the use of APs due to their low cost and easy installation. However, APs generally use the license-free ISM band. The range and number of channels at 2.4GHz or 5GHz ISM band are limited. Therefore in the limited area, the dense utilization of wireless network using smartphone and the increasing number of APs per unit area degrade the performance of network inevitably. It is difficult to assure QoS in real time applications such as streaming service. Also, real time service such as u-NMS [1] is absolutely important in case of managing network using smartphone and some serious problems could occur due to this performance degradation of network. Previous research for performance improvement of WLAN was carried out either to discover ideal deployment of APs or to study an algorithm for selecting AP for communication by simulation. Also many analytical studies proceeded using simulation tools instead of measuring performance in real environment [2][3][4]. The previous APs are layer 2 device, but the newer APs can supports not only at layer 2 but also handle layer 3. Thus, we carry out our analysis using smartphone in WiFi network in order to analyze performance of network at layer 3.

In this paper, we consider a smartphone as a communicating device and we guess and analyze the possible factors by predicting and implementing possible situation under the specific environment such as apartment or small limited area. The AP that we have used for our experiments uses 2.4GHz band and supports 802.11b/g in WLAN environment. Experiments proceed with three cases and in two ways i.e., the case of the same and different channel.

II. Environment for Experiment

APs usually cover service area up to 100m. However, in the small and limited area such as home or office; it can cause communication problems among the devices due to peripheral obstacles or interference from nearby APs. Experiments proceed after removing all other devices using 2.4GHz frequency band. We assume that APs used for experiments have same model with support of 802.11g standard. Ethernet is connected directly to AP and it can support speed of 100Mbps.

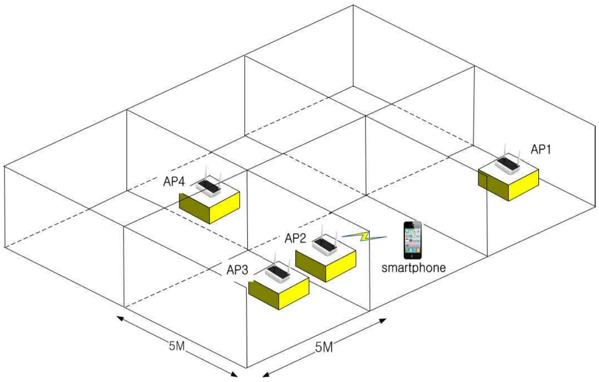


Fig 1. Real measurement environment

Figure 1 shows the experimental setup in which each square block represents a room. The size of one side of room is 5 meters and there are four APs in operation. The smartphone can communicate only with the AP2, which is in the same room. Experiment proceeds in three cases under variations in network conditions like the number of APs, location of APs, etc. For each case we further consider two situations. In first situation we perform experiments using channels of the same frequency while in second situation we assume that the channels have different frequency. Galaxy S based on Android 2.2 is used for the test. We make cache buffer of Galaxy S empty and terminate all other applications for accuracy. We examine the equal sized UDP packet, which is more simple protocol as compared to TCP packet due to its processing retransmission and time out, etc.

III. Results and Analysis

CASE 1. Throughput variation on the basis of distance between smartphone and AP:

Figure 2 represents the throughput degradation of smartphone as it moves from A to B. There is no obstacle between two APs and the distance between A and B is 10 meters.

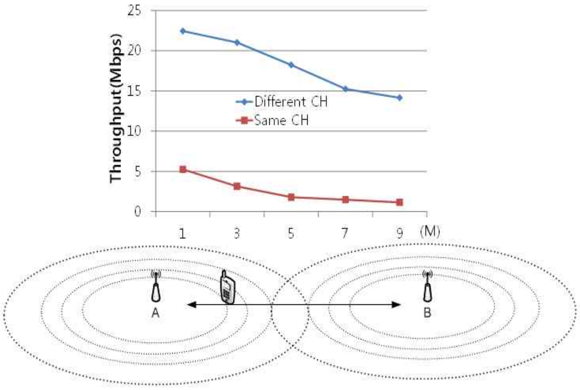


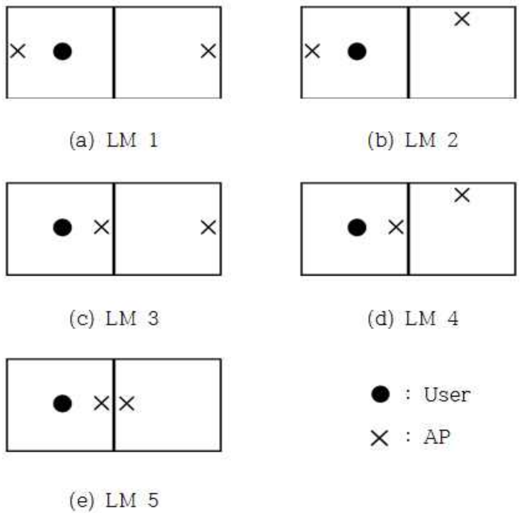
Fig 2. Throughput variation with smartphone movement from A to B

We measure the throughput at the distance of 1, 3, 5, 7, and 9 meters from A. Experiments are carried out in two situations, by using the same frequency channels and using different frequency channels. To make it close to real environment, different channels that we assigned to APs in our experiments are somewhat overlapped.

In the situation, when we assigned different channels to APs, the throughput decreased about 40% as APs moves from 1m point to 9m point. However, the performance decreased significantly about 80% when we perform same experiment using the same frequency channels to APs. The performance of network using different channels in APs is always better than the case of using same frequency channels in APs. According to the results, we can conclude that the channel frequency assigned to AP is an important factor for the network performance as compared to distance between AP and the smartphone.

CASE 2. Throughput variation due to deployment of APs:

In this case we deploy the APs in five (5) possible manners depending upon their location. The adjacent rooms have equal size and we turnoff all other devices which are using the same frequency band. We assume that there is only one AP in each room to make it near to practical scenario. Also, we use α as propagation coefficient, which represents property of wall to transmit data between rooms. We assume α=1 in order to simplify the complexity analysis. Figure 3 indicates deployment of APs in five different manners from (a) to (e) with variation in the distance between APs. “X” mark shows installed AP and “●” mark represents a smartphone user. The smartphone user can communicates with only AP2 which is located in the same space. To make it simple, smartphone is in the middle of the room.



(a) LM 1

(b) LM 2

(c) LM 3

(d) LM 4

(e) LM 5

Fig 3. Location Map of APs and smartphone user

Figure 4 shows the performance of smartphone during communication with AP2 as shown in Figure 3. The x-axis represents five cases (a) to (e) and y-axis indicates throughput against each case.

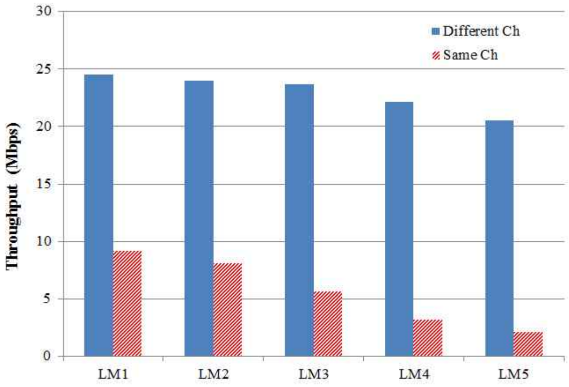


Fig 4. Throughput of smartphone depending on location and channels of APs

The case (a) indicates the situation when two APs are far away from each other while case (e) represents the situation when two APs are in close vicinity. The plain bar graph represents the situation when APs are using different channels and the dashed bar graph illustrate situation when APs are utilizing same channels. There is a drop of 12.5% and 75% in the throughput when APs are using different and same channels beside the location variation of APs. So, the deployment of APs can be an important factor for smartphone performance in case the APs are using same channels. According to the location of AP, the performance when APs are using different channels is always higher than the case when APs are using same channels in all five cases. We consider two factors, interference due to channel assigned to APs and communication distance between AP and smartphone.

As the result, we put more weight to interference resulted from channel on smartphone performance. Comparing throughput variation of CASE 1 and CASE 2, we induce that signal attenuation due to a wall is also significant. We use α here to represent propagation coefficient by walls, α depends on materials or thickness of walls. The larger the value of α better the transmittance and α is under condition for 0< α ≦1.

CASE 3. Throughput variation with the increase of APs:

In this case an AP is additionally supplemented in order to perceive the throughput variation with addition of APs. Each room has same size and only one AP is installed in each room. For cases (A) and (B) shown in Figure 5, we measure the throughput by using the same and different channels in APs. We compare the throughput of the best case with the throughput of the worst in CACE 2 shown in Figure 4. By adding new AP, we compare the throughput against the best (A) and the worst (B) cases shown in Figure 5. We express change of situation as followed that [number of users, number of APs, added AP number].

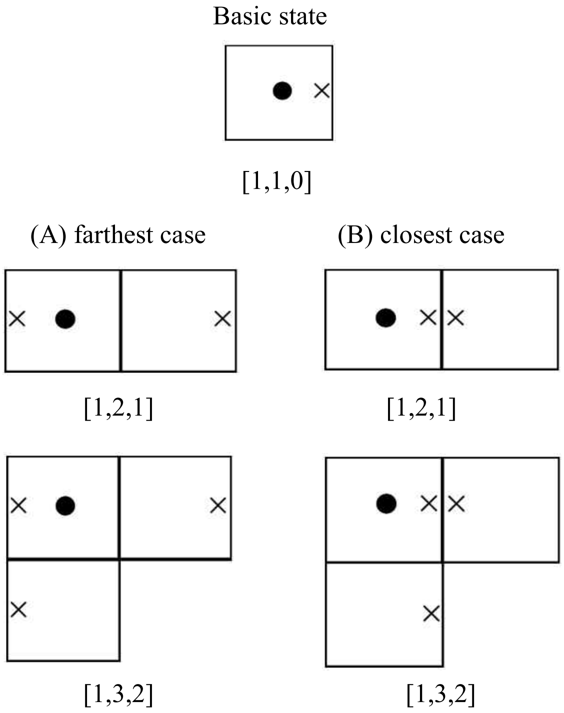


Fig 5. Experimental setup: phased AP addition.

In Figure 6 the x-axis represents number of APs and y-axis indicates the throughput. The dashed bars represent the throughput in case (A) while plain bars indicate throughput for case (B).

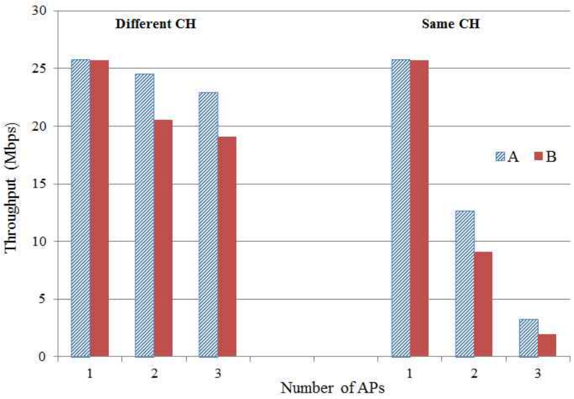


Fig 6. Throughput of smartphone depending on the number of APs

The throughput of smartphone is 26.5Mbps when only one AP is working which is about 50% of theoretical maximum speed that is given in IEEE 802.11g standard [5]. The addition of new AP declines the performance of network for both cases when APs are working with the same or different channels. It is clear from the results that network performance decreases with the increase in number of APs. In case of using different channels in AP as shown in Figure 6 the throughput is reduced as a whole and the degree of performance degradation is remarkable in case B instead of case A of Figure 5. It means that even though two APs are assigned different channels, the throughput can be affected by the deployment of APs. From the above results, we can conclude that performance degradation of smartphone could be reduced if we deploy the APs properly. However if the same frequency channel is assigned, the performance degradation is more than 60% by the addition of just single AP. It means proper AP deployment has no large effect in this case. Therefore, we conclude that a most intensive factor influencing the performance of smartphone is the channel assigned to AP instead of proper deployment of AP.

IV. Conclusion

Lately, the demand of WiFi network increases because of the massive dissemination of smartphones and popularization. However, increasing number of APs in small and limited area is recognized as a new problem which is not well addressed in previous research. Therefore, we analyzed the network performance by measuring throughput of smartphone in variation with factors like distance of adjacent APs, the number of APs, deployment of APs and channels assignment.

Experimental results show that the deployment of APs affects the performance when APs are using same channels. It is about 60% with the addition of just single AP therefore proper deployment of APs is the solution for this issue. However, it is obvious from the results that the throughput reduced linearly with the use of different channels while it decreased exponentially in case of using same channels. Thus the channel assignment to APs is the most critical factor for the performance degradation of network and it could be optimized using different frequency channels in APs. Moreover, we also found that signal attenuation due to a wall is also a significant factor for network performance.

We are planning to study on propagation coefficient (α) and throughput based on operating system such as Android, iOS, blackberry, Symbian etc. We are also planning to study on an algorithm for channel distribution server, which can control AP remotely at TCP layer, in order to assure minimum throughput.

Acknowledgment

This research was supported by the Ministry of Knowledge Economy (MKE), Korea, under the national HRD support program for convergence information technology supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2010-C6150-1001-0013).

References

[1] C.H. Kim, S.B Lee, K.G. Lee, "u-NMS Using Smartphone for Wireless Integrated Network Management", *Worldcomp'10 ICWN*, Las Vegas, NV, vol. 2, pp. 617-618, July, 2010.

[2] J. R. Gallardo, D. Makrakis, H. T. Mouftah, "Performance Analysis of the EDCA Medium Access Mechanism over the Control Channel of an IEEE 802.11p WAVE Vehicular Network", *IEEE ICC 2009*, Dresden, pp. 1-6, June 2009.

[3] I. Broustis, K. Papagiannaki, S. V. Krishnamurthy, M. Faloutsos, V. P. Mhatre, "Measurement-Driven Guidelines for 802.11 WLAN Design", *IEEE/ACM Trans. Networking*., vol. 18, pp. 722-735, June 2010.

[4] F. Zarinni, S. R. Das, "Adaptive Spectrum Distribution in WLANs", *IEEE GLOBECOM 2010*, Miami, FL, pp. 1-6, December 2010.

[5] IEEE Std. 802.11g-2003; Supplement to part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Amendment 4: Further Higher Data Rate Extension in the 2.4GHz Band, June 2003.

Intelligent Smart Building and Its Wireless Network Environments

Chaehwan Kim, Jungchan Ahn, Kyung-Geun Lee
Department of Information and Communication Engineering, Sejong University
98 Gunja-Dong, Gwangjin-Gu, Seoul, Korea
astro0821@nrl. sejong.ac.kr, zxkal@nrl.sejong.ac.kr, kglee@sejong.ac.kr

Abstract – *intelligent smart building (ISB) is defined as a building which allows interaction of user with wireless sensors and building systems by forming a wireless sensor network to share or transfer real time data. This paper proposes the novel idea of a specialized server called wireless communi -cation server (WCS) within the ISB to provide seamless service for mobile node (MN). The WCS performs MAC authentication, MAC forwarding, cache and buffering func -tions. It is apparent from the experimental results that WCS has reduced 52% of initial delay and 62% of handover delay time.*

Keywords: Intelligent Smart Building, Wireless Sensor Network, Wireless Communication Server

1 Introduction

The intelligent building (IB) employs an automatic wireless system to control sensor network (temperature, CO_2 and humidity sensor, etc.) for the ideal environment of the building [1][2]. This paper proposes an ISB as a user intimate building which is more advanced than the traditional IB in terms of sensing and communication. The MN in ISB represents an entity like an employee, a manager, or a customer who can use wireless network (3G, WiFi or RFID) for communication. The ISB collects data from users on the basis of their characteristics and performs transmission after analysis on gathered data. The target for the proposed communication system (WCS) is a building within which MN can request the information about the state, location and video of the building. The components of the ISB are *smart gathering server* (SGS), s*mart transmission serve*r (STS) and *smart authentication server* (SAS). Whenever MN requests info, each server gathers and analyzes the requested info and transfers it in real-time by verifying the location of MN.

However, a WiFi network has longer handover delay time as compared to a 3G network. Handover latency and packet loss are serious concerns for the ISB as it needs to support a real-time service based on the position and status information of MN. Proxy mobile IP protocol provides solutions to these problems within homogenous networks. However, current available WiFi AP's can execute handover protocols at layer 3 and layer 2 simultaneously, therefore it is not feasible for PMIPv6 to judge the handover of a MN at the LMA (due to lack of support about handover at L3). Moreover, coexistence of PMIPv6 with WiFi demands reinstallation of AP's within an ISB which increases the overall cost. This paper proposes a novel idea of *wireless communication system* (WCS) that employs the MAC address of MN for WiFi network installed devices in an ISB. The WCS provides mobility support by using the MAC address of MN which can be easily incorporated in current established network.

2 Architecture of Intelligent Smart Building

The ISB is a building that allows the interaction of users with wireless sensors and building systems by forming a wireless sensor network to share or transfer real time data.

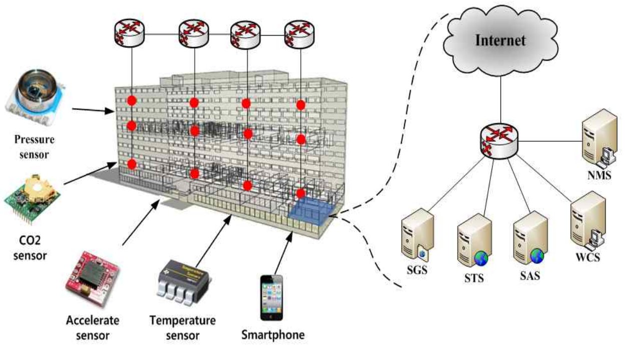


Figure 1. Basic architecture of ISB

Figure 1 shows the architecture of the ISB. The proposed ISB consists of a small authentication server (SAS) that authenticates the MN, a gathering server (SGS) which collects the position information of MN, a smart transmission server (STS) that performs analysis and transmission of collected information and a wireless communication server (WCS) which provides mobility support to the MN. The MN in the ISB constantly requests state, location and image info of the building with the mobility support at the same time. The mobility of MN causes handover latency and packet loss therefore an efficient wireless system is required to support real time applications.

This paper proposes WCS to support real time implementation and explains its function execution process. The proposed system prepares a MAC table on the basis of MNs attributes and realizes it by referring to the protocol stack process where the frame in the IEEE 802.11g standard is converted into an IEEE 802.3 wired communication network frame [3]. The proposed MAC table consists of the MAC address, IP address and triggering event of MN. The

This research was supported by The Ministry of Knowledge Economy (MKE), Korea, under the national HRD support program for convergence information technology supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2010-C6150-1001-0013).

triggering event occurs whenever layer 2 issues handover triggering ACK. The WCS detects the handover process by observing this triggering ACK. The functions of the WCS can be explained as follows:

1) **MAC authentication and access log caching:** If MN tries initial access, then the WCS caches the access log information and records the MAC address of the MN in the MAC table. The access log file in the cache can be transmitted from the WCS to the MN in order to minimize the processing delay.
2) **MAC forwarding:** The MAC forwarding function converts IP address into MAC address during handover process of MN. This conversion process provides mobility support to MN and allows it to maintain a session during handover process.
3) **Packet buffering:** When a MN moves to another floor or selects another AP in the ISB, a handover is executed.

3 Experimental Results and Analysis

In this paper, we propose the ISB that includes SGS, STS, SAS and WCS. The prime objective of the WCS is to provide mobility support to MN within the ISB. Experiments are carried out for initial access time and latency time during a handover process. We also compare and analyze the WCS protocol with the previously proposed MIPv6, FMIPv6, PMIPv6 protocols. The structure of testbed is composed of two Access Routers(ARs), one backbone router, one WCS, one STS and one SGS. In this scenario we estimate the handover latency and initial access time when a MN moves across different floors in the ISB.

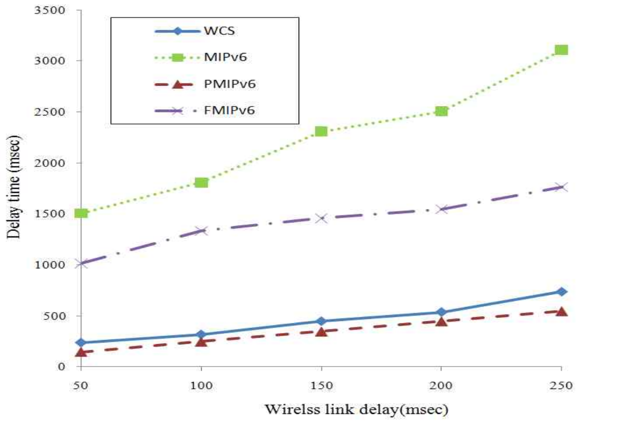


Figure 2. Initial access delay between MN and CN

Figure 2 shows that initial delay time of WCS has reduced up to 74% as compared to the MIPv6 and FMIPv6. The cache function decreases the processing delay time compared to MIPv6 and FMIPv6. Moreover, the WCS prevents the traffic congestion by controlling the redundant data.

On the other hand, the delay time of the WCS shows similar values like PMIPv6 because the LMA of PMIPv6 performs a similar caching functions to the WCS. However, PMIPv6 is not applicable within WiFi networks because it increases the reinstallation of new APs.

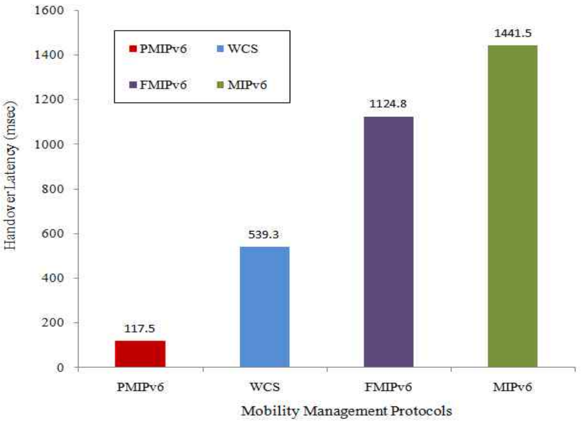


Figure 3. Handover latency between MN and CN

Figure 3 shows handover latency for MN in the ISB. The bar-graph shows that the average delay time in PMIPv6 is 117.5 msec, which supports the most flexible handover among the experimental values, but again the problem of deployment cost is a significant barrier in its implementation. Comparing with the other protocols, the WCS reduces delay time up to 62% in comparison with MIPv6 and up to 52% in comparison with FMIPv6.

4 Conclusion

This paper defines the intelligent smart building and wireless sensor network environment and proposes the WCS in order to provide mobility support inside an ISB. The ISB is a building which allows the interaction of users with wireless sensors and building systems.

Mobility support is an important consideration owing to the mobile behavior of MN within the ISB in real time environment. Thus, WCS reduces latency during handover process and provides seamless service. It can be clearly seen through experimental results that the WCS has significantly reduced the handover delay time by 62% and initial access time by 52 % at best as compared with MIPv6.

Future work will include development of data transmission algorithms within SGS and STS for reliable communication. Moreover, by expanding the testbed to inter-domain the experiments on throughput and delay time will be carried out.

References

[1] H. Wicaksono, S. Rogalski, E. Kusnady, “Knowledge-based intelligent energy management using building automation system”, in *Proc. IEEE IPECON 2010*, Singapore, pp. 1140-1145, October 2010.

[2] J. Byun, S. Park, “Development of a self-adapting intelligent system for building energy saving and context-aware smart service”*, IEEE Trans. Consumer Electronics*., vol. 57, pp. 90-98, February 2011.

[3] IEEE Standard. 802.11g, “Wireless LAN Medium Access Control and Physical Later Specifications: Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements”, 2005.

SESSION

ALGORITHMS AND PROTOCOLS

Chair(s)

TBA

The Ins and Outs of Distance-Based WSN Localization Schemes

Kerri Stone, Tracy Camp
Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, Colorado, USA

Abstract— *Localization is a fundamental problem in wireless sensor networks. In many applications, sensor location information is critical for data processing and understanding. While the global positioning system (GPS) can be used to determine mote locations, the high cost often prohibits the ubiquitous use of GPS for location estimates. The cost of GPS has motivated researchers to develop localization protocols that determine mote locations using cheap hardware and wireless intra-network measurements. In selecting a localization algorithm for use in a WSN deployment, it is critical to understand the performance of each algorithm under different network topologies. We contribute a comprehensive review and realistic performance comparison of five distance-based WSN localization algorithms.*

One of the five algorithms that we compare was first published in 2010, and is included because the authors report accurate localization results: TSL. Two other chosen algorithms (MDS-MAP and dwMDS) are older protocols, but have also been shown in the literature to perform well. Finally, two other algorithms that we compare are well-known and commonly used for localization algorithm comparison: DV-Distance and Robust Quadrilateral. In the current literature a head-to-head comparison of these five algorithms does not exist. We present a realistic performance assessment of these five distance-based localization protocols and provide insight into the inner-workings of each protocol.

Keywords: WSN localization, distance-based localization, localization algorithm, realistic localization algorithm evaluation.

1. Introduction

Wireless sensor network (WSN) localization is the process of calculating the geometric location of motes and is critical for many WSN applications. For example, one WSN application that requires precise mote localization is geophysical monitoring. Wireless geophysical monitoring often includes wave or electrical methods to study the subsurface of the earth. Both wave and electrical geophysical monitoring methods collect data across a field of sensors and use an inversion processing technique to image subsurface properties and phenomenon. The post-processing of geophysical data requires precise location information for each sensor. Any error in sensor location introduces uncertainties and noise into the geophysical signal processing and eventual subsurface characterization.

We could equip each mote with a global positioning system (GPS) chip for localization purposes; however, due to cost constraints (both financial constraints and energy constraints), the ubiquitous use of GPS for WSN localization is infeasible. Because of increased hardware cost and increased energy consumption, WSNs requiring location aware sensing usually only equip a small number of motes with a GPS chip. These motes with GPS chips are called *anchor* motes.

Localization algorithms for static networks are plentiful; however, knowledge on algorithm performance in *realistic* environments with *various network topologies* is lacking. For example, most distance-based algorithm performance analysis is conducted in simulation using what is known as the *noisy disk* ranging model. The *noisy disk* ranging model assumes a spherical transmission pattern with random Gaussian noise that is not representative of real-world ranging scenarios [1]; thus, simulation analysis using the *noisy disk* ranging model often results in unrealistic analyses. Moreover, localization algorithm performance is typically evaluated using random network topologies. Many real-world network topologies, however, are not random [2]. For example, geophysical sensors are typically aligned in grid patterns. Thus, motes in a geophysical WSN would also be positioned in a grid. To target the requirements of specific WSN applications, it is crucial to understand algorithm performance under various network topologies. In this paper, we provide a comprehensive review of distance-based WSN localization techniques. The main contribution of our work is to provide a *realistic* performance assessment of five popular and/or high-performing distance-based WSN localization algorithms under *varying network configurations*.

2. Background

Distance-based localization algorithms have two steps: 1) estimate inter-mote distances and 2) calculate locations based on inter-mote distances. Additionally, distance-based localization techniques have characteristics that define algorithm operation and performance. The following sections detail distance measurement techniques, location computation techniques, and distance-based algorithm characteristics.

2.1 Distance Measurement Techniques

Distance based localization algorithms can use several different measurement techniques, and the chosen technique defines localization hardware requirements. The five different measurement techniques used in distance-based localization algorithms are:

1) ***time of arrival*** **(TOA)**,
2) ***round trip time of arrival*** **(RTTOA)**,

3) ***time difference of arrival*** **(TDOA)**,
4) ***received signal strength*** **(RSS)**, and
5) ***connectivity***.

Of the five measurement techniques, received signal strength (RSS) and connectivity are the most popular. This popularity is due to their simplicity and generality (i.e., no special hardware is required). For our algorithm comparisons, we use RSS and connectivity distance measurement techniques, and we summarize these techniques next. For a complete review of the five distance measurement techniques, the reader is directed to Liu et al. [3].

The ***received signal strength (RSS)*** measurement technique estimates distance based on received signal strength and known signal attenuation properties. Localization algorithms that use RSS to estimate distances model the wireless signal strength as a monotonically decreasing function with increased distance between the sender and the receiver. This relationship between distance, d, and path loss at a receiver, P_r, is commonly modeled with the log-normal equation:

$$P_r(d) = P_o(d_o) - 10n_plog_{10}(\frac{d}{d_o}) + X_\alpha, \qquad (1)$$

where $P_o(d_o)$ is a reference path-loss in decibel milliwatts (dBm) measured at a distance, d_o, from the transmitter, n_p is a path loss exponent that represents how the environment affects the RSS value, and X_α is a zero mean Gaussian distributed random variable that accounts for shadowing affects, i.e., $X_\alpha \sim N(0, \sigma^2)$. Variables n_p and σ are environment dependent and vary from application to application [4]. Some researchers refer to localization protocols that use RSS measurement techniques as "range-based" or "fine-grained" localization because these techniques calculate inter-mote distances based on known signal propagation properties.

The ***connectivity*** distance measurement technique uses packet hop count to estimate distance. Connectivity based localization algorithms use *anchor* motes that broadcast their locations. All motes that hear a given anchor-mote location broadcast are assumed to be one-hop away from that *anchor*. The distance between an anchor-mote and every mote that receives the *anchor's* transmission is calculated based on the expected one-hop communication propagation length. In general, the connectivity technique sacrifices location accuracy for simplicity. Some researchers refer to localization protocols that use connectivity distance measurements as "range-free" or "coarse-grained" localization because mote location estimates are based purely on network connectivity information.

2.2 Location Computation Techniques

Distance-based localization algorithms use various mote-location computation techniques; the techniques used by the distance-based algorithms we evaluate are:

1) ***multidimensional scaling*** **(MDS)**,
2) ***linear programming***,
3) ***statistical estimation***, and
4) ***trilateration***.

Each of the five algorithms that we compare use one of these four location computation techniques. We review these four location computation techniques next. Details on other location computation techniques used by WSN localization algorithms (e.g., machine learning and stochastic optimization) exists in the work by Mao and Fidan [4].

Multidimensional scaling (MDS) can use any of the distance measurement techniques discussed in Section 2.1, including connectivity. MDS algorithms compute the shortest paths between pairs of motes (based on euclidean distances) and store the information in a distance matrix. MDS then uses the distance matrix to assign a location to each mote in N-dimensional space. To assign a location, MDS uses cosines and liner algebra to reconstruct relative locations of motes based on inter-mote distances. When MDS is used in localization algorithms, N is either two or three (depending on network dimensionality).

Many localization algorithms use ***linear programming (LP)*** techniques for mote location estimation. To use linear programming to estimate mote location, localization is formed as a convex optimization problem. LP localization uses a mathematical model to determine the best outcome for mote locations. Usually LP is used to calculate the minimum error in location estimates for each mote. Similar to MDS, LP location estimation techniques can use any of the distance measurement methods discussed in Section 2.1.

Localization algorithms use ***statistical estimation*** for mote locations through a maximum likelihood (ML) estimator. When ML is used for mote localization, parameters for the underlying statistical model are selected (i.e., mote locations), given inter-mote distance measurements, that maximize the likelihood distribution. The ML estimator can be modeled as follows:

$$\hat{X} = \text{argmax}_X \hat{f}(Z|X) \qquad (2)$$

where Z is a vector of inter-mote distance measurements and X denotes potential coordinate vectors of non-anchor motes; thus, $\hat{f}(Z|X)$ is the conditional probability of $f(Z)$ when the non-anchor motes are located at X. Unlike MDS and LP, statistical estimation techniques cannot operate with connectivity information; in other words statistical estimation must have inter-mote distance measurements for operation.

Trilateration is a geometric estimation technique that uses distance measurements and the geometry of shapes (e.g., spheres [5, 6], triangles [7], or Bezier curves [8]) to estimate relative locations of points. Trilateration is often used in WSN localization techniques to calculate the location of a mote using inter-mote distances calculated using any of the distance estimation techniques discussed in Section 2.1. For trilateration to succeed in computing relative point locations, the known locations should be non-collinear (in two-dimensional space) and non-coplanar (in three-dimensional space). Section 2.1.

2.3 Algorithm Characteristics

We categorize distance-based localization algorithms as *centralized*, *distributed*, or *distributed-centralized*. A centralized localization algorithm uses a base station to calculate mote location, which means each mote must transmit measurement information to the base station for processing. Since wireless transmissions consume large amounts of energy, centralized algorithms typically consume more energy than distributed and distributed-centralized algorithms; however, as shown in our results, centralized localization algorithms generally produce better location estimates than distributed and distributed-centralized algorithms. In a distributed algorithm, each mote calculates its location in-network using local measurement information. Distributed algorithms decrease the energy consumption required for the localization process, but sacrifice location estimation accuracy. Distributed-centralized algorithms run a centralized algorithm on overlapping clusters within the network. This cluster distribution alleviates the need to transmit all measurement information to a central location, which decreases the localization energy consumption when compared to centralized algorithms. Due to inter-cluster communication, distributed-centralized algorithms consume more energy than distributed algorithms.

Distance-based localization algorithms can also be classified as *anchor-based* or *cooperative*. An anchor-based algorithm uses motes with known locations (i.e., anchor motes) to infer the location of other motes within the network. In a cooperative algorithm, none of the motes know their location prior to the execution of the localization algorithm; instead, motes use inter-mote distance measurements to create location estimates for every mote.

Several distance-based localization algorithms are written for the two-dimensional case (i.e., localization in the x and y direction). We note that real-world localization deployments are not limited to two-dimensional topologies and require an algorithm that works in three-dimensions (i.e., x, y, and z). In Section 2.4, we categorize algorithms as three-dimensional when they are capable of localizing motes in the x, y, and z direction.

Not all distance-based localization algorithms are capable of localizing 100% of the network. The percentage of the network localized is referred to as coverage. When a localization algorithm does not guarantee 100% coverage, the network topology and node degree (number of one hop neighbors) dictates how much of the network the protocol can localize. In Section 2.4, we note which algorithms provide 100% coverage.

2.4 Localization Algorithms

WSN localization is widely researched. As such, there are too many distance-based localization algorithms to offer a complete review on every published method in this paper. Instead, we chose five distance-based algorithms to evaluate due to either their popularity (i.e., DV-Distance and Robust Quadrilateral) or their reported localization accuracy (i.e., MDS-MAP, TSL, and dwMDS). In addition to choosing algorithms based on popularity or accuracy, we chose two *centralized* (i.e., MDS-MAP and TSL), one *distributed* (i.e., DV-Distance), and two *distributed-centralized* (i.e., dwMDS and Robust Quadrilateral) algorithms. In this section we review the five distance-based algorithms; see Table 1 for a high-level summary of these algorithms. In the following sections, material discussed in Sections 2.1 – 2.3 is used to help summarize and categorize these five localization algorithms.

2.4.1 Centralized

a) MDS-MAP: MDS-MAP [9] uses MDS to produce mote coordinates that are the best fit to all measured inter-mote distances; as mentioned, any measurement technique discussed in Section 2.1 can be used with MDS. The coordinates produced by MDS lie at an arbitrary rotation and translation. MDS-MAP normalizes the coordinates of the location estimates to take into account *anchor* motes with known locations. This normalization rotates and translates the estimated mote locations to correlate with known *anchor* locations.

In the original MDS-MAP publication, Shang et al. [9] reported extensive Matlab simulation results comparing MDS-MAP to the traditional MDS algorithm. The simulation comparisons used various network topologies and differing levels of ranging error and the authors found that MDS-MAP improves MDS. MDS-MAP is able to produce better location estimates than traditional MDS because MDS-MAP rotates and translates the location estimates to align the network with known *anchor* mote locations. Shang et al. [9] concluded that MDS-MAP is an effective localization protocol, particularly when there are few *anchor* motes and the network topology is relatively uniform.

b) TSL: Temporal Stability Localization (TSL) [10] uses one-hop communication ranges and distance measurements to create a system of constraints to statistically refine each mote's possible location. While other algorithms (e.g., Sextant [8] and Centroid [5]) use only positive (i.e., mote can be at) constraints, TSL uses both positive and negative (i.e., mote cannot be at) constraints to calculate mote locations. TSL first uses the estimated distance measurements to create positive constraints based on distance equality and bound constraints. Distance equality constraints are derived based on RSS measurements: the distance between mote i and mote j is equal whether that distance is calculated at mote i or mote j. Distance bound constraints are derived from connectivity information: if mote i receives a packet from mote j, then mote i must be within communication range of mote j. Similarly, a negative constraint exists if mote i does not receive a packet from mote j. To find a mote's location based on these positive and negative constraints, TSL uses a gradient based local search algorithm.

While TSL was developed for mobile networks, Rallapalli et al. [10] compared TSL to several static localization algorithms and found that TSL produces less location estimation error than Centroid [5], Sextant [8], and traditional MDS. For this comparison, 10% ranging error was added to pairwise distances between motes before the localization algorithms were run. To summarize their simulation results, Rallapalli et al. [10] concluded that TSL effectively accounts for the temporal stability observed in mobile networks (i.e., the direction and speed of mobile nodes are often the same at adjacent points in time). This observation enables the TSL algorithm to be robust to ranging errors.

2.4.2 Distributed

a) DV-Distance: DV-Distance [11] approximates the distance between non-anchor mote i and *anchor* mote j as the shortest path, sp_{ij}, between mote i and j. DV-Distance uses sp_{ij} to constrain the location (i.e., (x_i, y_i)) of non-anchor mote i in terms of *anchor-mote* j:

$$0 = (x_i - x_j)^2 + (y_i - y_j)^2 - sp_{ij}^2. \qquad (3)$$

A system of three such equations, in terms of at least three *anchor* motes (i.e., $j = 1, 2, 3, ...n$), are linearized and solved using least squares for the coordinates of mote i.

Niculescu and Nath [11] compared DV-Distance to DV-Hop, which was introduced in the same publication and uses hop counts instead of distance estimates. The authors found that, in general, DV-Distance produces more accurate distance estimates than DV-Hop, if the ranging method is subject to < 50% error. The authors also concluded that DV-Distance offers 100% coverage if the *anchor* to mote density is >= 10%.

2.4.3 Distributed-Centralized

a) dwMDS: Distributed weighted MDS (dwMDS) [12] uses the MDS protocol on clusters within the network and pre-processes the distance estimates before running MDS. dwMDS selects a radius, r, which is smaller than the communication range of the motes in the network, and uses it to prune large communication links. If a link's distance is estimated as larger than the selected r, the connection is dropped from the MDS distance matrix before the algorithm executes.

Costa et al. [12] used empirically collected ranging data (TOA and RSS) to compare dwMDS to traditional MDS and found that dwMDS offers better localization accuracy. Costa et al. [12] concluded that dwMDS is well suited for use in wireless sensor networks because the algorithm provides accurate location estimates and does not consume a lot of power. In other words, dwMDS has lower intra-network communication overhead than MDS because it is a distributed-centralized algorithm instead of a centralized algorithm.

b) Robust Quadrilateral: Robust Quadrilateral [13] uses inter-mote ranging information to identify all of the robust quadrilaterals that exist between neighbors, and then creates a subgraph of overlapping robust quadrilaterals within clusters. A triangle of three motes i, j, and k is defined as robust if the smallest angle in the triangle formed by the three motes, α, satisfies the requirement: $dcos^2(\alpha) > \theta$, where d is the shortest edge within the triangle and θ is a pre-defined threshold. The algorithm enforces robust quadrilaterals by limiting triangle angles between all motes within a quadrilateral as larger than the minimum threshold, θ. A robust quadrilateral exists if and only if all triangles formed by motes in the quadrilateral are robust. Moore et al. [13] state that this criteria for robustness eliminates many of the ranging-related issues that introduce location approximation errors into other distance-based localization algorithms.

Moore et al. [13] used simulation to analyze the algorithm's performance. The authors found that, with increasing levels of ranging inaccuracies, the localization accuracy of Robust Quadrilateral decreases. To achieve 100% coverage with Robust Quadrilateral, the network must have node degree >= 10. Furthermore, the authors found that, as ranging inaccuracies increase, Robust Quadrilateral performs best with high node degree.

Table 1: Localization algorithm summaries. A "X*" denotes that MDS-MAP and TSL can be an anchor-based algorithm if *anchors* are available.

	Algorithm	distance measurement technique	location comp. technique	anchor-based	cooperative	3D	100% coverage
Centralized	MDS-MAP	TOA, RTTOA,TDOA, RSS, or connectivity.	MDS	X*	X	X	X
	TSL	TOA, RTTOA, TDOA, or RSS; and connectivity.	statistical	X*	X	X	X
Distributed	DV-Distance	RSS	linear prog.	X			
Distributed-Centralized	dwMDS	TOA, RTTOA, TDOA, or RSS.	MDS		X	X	
	Robust Quads	TOA, RTTOA, TDOA, or RSS.	trilateration		X	X	

3. Algorithm Evaluation

Many of the published localization algorithms include a simulation comparison of several algorithms. From publication to publication, however, it is hard to compare algorithm performance because no standards for comparison exist; algorithms are reviewed in different simulation environments, with different sets of ranging data, under different assumptions. Additionally, algorithms are typically analyzed with a randomly generated topology, which is not representative of real-world WSN deployments [2]. We evaluate the localization accuracy and coverage of five localization algorithms in simulation under various network configurations.

For our evaluation, each algorithm is implemented in Matlab using the Silhouette [14] localization framework. To compare algorithm performance, we used RSS ranging models produced with *empirically collected* RF data (from the CC1000 radio module); in addition, we used the Statistical Emulation [14] method to generate network topologies that accurately model real-world ranging noise. To compare algorithm performance over different network configurations, we ran 12 experiments with 100 trials each (i.e., 100 network topologies). The network configuration used for each experiment is summarized in Table 2.

Table 2: Network configurations for our 12 experiments.

Experiment	Topology	Network Size	Anchor Motes
1	uniform	25	4
2	uniform	25	9
3	random	25	4
4	random	25	9
5	uniform	49	9
6	uniform	49	16
7	random	49	9
8	random	49	16
9	uniform	100	9
10	uniform	100	16
11	random	100	9
12	random	100	16

Our evaluation results are summarized in Figures 1(a) – 1(f). Algorithm localization error (in cm) for each experiment is depicted using a bar plot (corresponding with the left hand side y-axis) for each algorithm/anchor mote pair; the corresponding coverage (in %) is depicted within the same graph using a bar plot (corresponding to the right hand side y-axis) drawn with dashed lines. We use the results from our 12 experiments to make four comparisons:

a) **Protocol**: performance under similar network configurations (i.e., number of motes, number of anchors, and topology).
b) **Anchors**: performance with different number of *anchors* under the same network size and topology.
c) **Network Size**: performance with different network sizes using the same network topology and similar anchor to mote ratios.
d) **Network Topology**: performance with different network topologies (i.e., uniform and random) and the same network size and number of *anchors*.

3.1 Protocol

Of the five algorithms we compared, the centralized algorithms (i.e., MDS-MAP and TSL) generally produce the lowest localization error. Table 3 summarizes simulation algorithm performance. In Table 3, algorithms are ordered by localization error for each experiment, where algorithm #1 has the lowest localization error for the specified experiment. An “X” in the Significant Improvement column denotes that the algorithms’ performance is statistically significant when compared to the next best performing algorithm (i.e., from algorithm ranked #1 to algorithm ranked #2); we calculated statistical significance between the two distributions using a t-Test with $\alpha = 0.05$. From the 12 experiments analyzed in Table 3 we can make two conclusions: 1) the centralized algorithms perform the best independent of anchor count and network topology and 2) of the two centralized algorithms, MDS-MAP outperforms TSL in 8 of the 12 cases.

Table 3: Algorithms ordered by localization error in each experiment. Robust Quadrilateral is not included in the ordering until the algorithm was able to achieve >= 50% coverage.

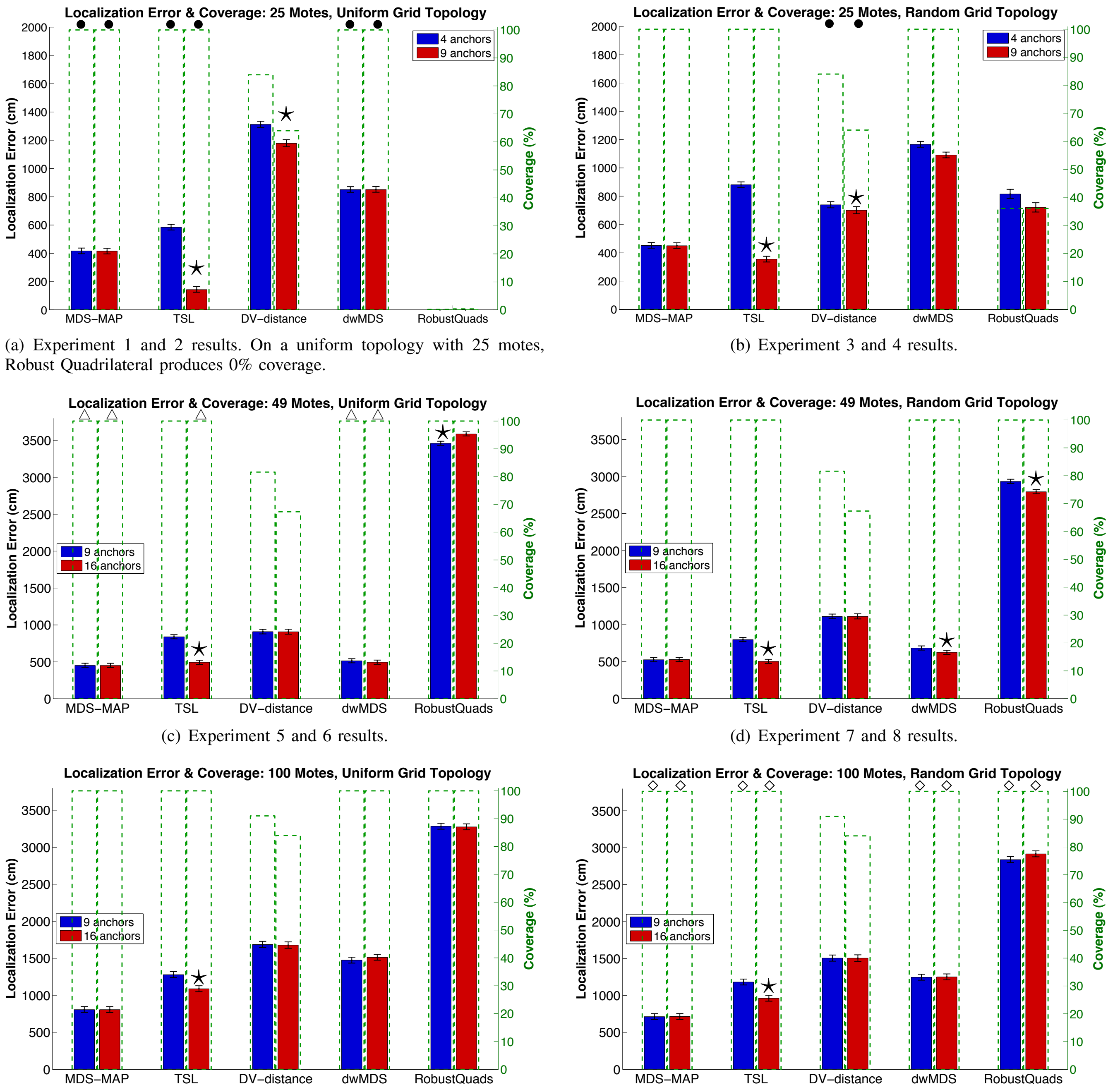
Exp.	**Algorithm Ranking**	**Significant Improvement**	**Exp.**	**Algorithm Ranking**	**Significant Improvement**
1	1. MDS-MAP	X	**2**	1. TSL	X
	2. TSL	X		2. MDS-MAP	X
	3. dwMDS	X		3. dwMDS	
	4. DV-Distance			4. DV-Distance	
3	1. TSL	X	**4**	1. TSL	
	2. MDS-MAP	X		2. MDS-MAP	X
	3. dwMDS	X		3. DV-Distance	X
	4. DV-Distance			4. dwMDS	
5	1. MDS-MAP	X	**6**	1. MDS-MAP	X
	2. dwMDS	X		2. dwMDS	X
	3. TSL	X		3. TSL	X
	4. DV-Distance	X		4. DV-Distance	X
	5. Robust Quads			5. Robust Quads	
7	1. MDS-MAP	X	**8**	1. TSL	X
	2. TSL	X		2. MDS-MAP	X
	3. dwMDS	X		3. dwMDS	X
	4. DV-Distance	X		4. DV-Distance	X
	5. Robust Quads			5. RobustQuads	
9	1. MDS-MAP	X	**10**	1. MDS-MAP	X
	2. TSL	X		2. TSL	X
	3. dwMDS	X		3. dwMDS	X
	4. DV-Distance	X		4. DV-Distance	
	5. Robust Quads			5. Robust Quads	
11	1. MDS-MAP	X	**12**	1. MDS-MAP	X
	2. TSL	X		2. TSL	X
	3. dwMDS	X		3. dwMDS	X
	4. DV-Distance	X		4. DV-Distance	X
	5. Robust Quads			5. Robust Quads	

3.2 Anchors

Each graph in Figure 1 displays protocol performance with different number of *anchor* motes (e.g., TSL with 4 *anchors* vs. TSL with 9 *anchors*). In each figure, statistically significant improvements for each algorithm are depicted with a $\star$; again, we calculated statistically significant improvements in localization error using a t-Test with $\alpha = 0.05$. When one considers all network configurations and topologies tested, we found TSL is the only algorithm that has significant improvements in localization accuracy when the number of anchor motes is increased. TSL’s accuracy consistently improves due to its process of statistically refining motes’ locations based on positive and negative ranging constraints. With a higher number of anchor motes, more of the constraints are based on known locations, which results in improved algorithm accuracy.

3.3 Network Size

To analyze the impact varying network size has on algorithm performance, we ran experiments with the same



(a) Experiment 1 and 2 results. On a uniform topology with 25 motes, Robust Quadrilateral produces 0% coverage.

(b) Experiment 3 and 4 results.

(c) Experiment 5 and 6 results.

(d) Experiment 7 and 8 results.

(e) Experiment 9 and 10 results.

(f) Experiment 11 and 12 results.

Figure 1: Localization error in cm (with 95% confidence intervals) and corresponding median coverage in % for each protocol in all 12 experiments in Table 2. In each figure, localization error is reported for two anchor mote configurations per protocol. Coverage is represented by the dashed bar plot and corresponds with the y-axis on the right hand side of the figures. Please note that figures (a) and (b) have a different y-axis scale for localization error than figures (c), (d), (e), and (f). The $\star$'s represent statistically significant improvements (via a t-Test with $\alpha = 0.05$) for each algorithm with the different number of anchor motes. The $\bullet$'s, $\triangle$'s, and $\diamond$'s at the top of each figure denote statistically significant improvements (again, via a t-Test with $\alpha = 0.05$) in the algorithm's performance under different topologies (between figures (a) and (b), (c) and (d), and (e) and (f)).

network topology and similar anchor to mote ratios; in other words, we compare experiments 1, 5, and 10 (for uniform topologies), and we compare experiments 3, 7, and 12 (for random topologies). Table 4 summarizes algorithm performance with varying network sizes and similar mote to anchor ratios. Of the five algorithms analyzed, Robust Quadrilateral is the only algorithm that performs better on large networks. (Again, we only analyze performance of this algorithm when coverage >= 50%.) Robust Quadrilateral performs poorly (i.e., has low coverage) on small networks due to its need to form a *robust quadrilateral* to localize each mote. With small networks, the ability to form *robust quadrilaterals* is limited, which results in decreased coverage and decreased localization accuracy. MDS-MAP always performs best on small networks. Similarly, DV-Distance, TSL, and dwMDS perform poorly on large networks.

Table 4: Algorithm performance under different network sizes with similar anchor to mote ratios. In the Performance Order column, the network sizes are ordered from best to worst performing under each algorithm category.

Alg	Topo.	Performance Order	Alg	Topo.	Performance Order
MDS-MAP	Unif.	1) 25mote 4anchor 2) 49mote 9anchor 3) 100mote 16anchor	TSL	Unif.	1) 25mote 4anchor 2) 49mote 9anchor 3) 100mote 16anchor
	Rand.	1) 25mote 4anchor 2) 49mote 9anchor 3) 100mote 16anchor		Rand.	1) 49mote 9anchor 2) 25mote 4anchor 3) 100mote 16anchor
DV-Distance	Unif.	1) 49mote 9anchor 2) 25mote 4anchor 3) 100mote 16anchor	dwMDS	Unif.	1) 49mote 9anchor 2) 25mote 4anchor 3) 100mote 16anchor
	Rand.	1) 25mote 4anchor 2) 49mote 9anchor 3) 100mote 16anchor		Rand.	1) 49mote 9anchor 2) 25mote 4anchor 3) 100mote 16anchor
Robust Quads	Unif.	1) 100mote 16anchor 2) 49mote 9anchor			
	Rand.	1) 100mote 16anchor 2) 49mote 9anchor			

3.4 Network Topology

To statistically compare our 12 experiments with different topologies (i.e., uniform or random) and the same network configuration (i.e., same mote and *anchor* number), we used a t-Test with $\alpha = 0.05$. Statistically significant improvements in algorithm performance between uniform and grid topologies in network sizes of 25, 49, and 100 are indicated with $\bullet$'s, $\triangle$'s, and $\diamond$'s, respectively. The symbols graphically mark the better performing protocol/configuration between Figures 1(a) and 1(b), Figures 1(c) and 1(d), and Figures 1(e) and 1(f). For example, in Figure 1(a), there is a $\bullet$ above MDS-MAP in both the 4 anchor and 9 anchor configurations. This graphical marking delineates that MDS-MAP has significant accuracy improvements in uniform topologies with 25 motes over random topologies with 25 motes.

MDS-MAP, TSL, and dwMDS generally perform better on uniform topologies than on random topologies when the network size is small or medium. In large networks, most protocols perform better on random topologies; in a random topology, motes do not necessarily span the entire grid (i.e., some motes will be close), which produces less ranging error and more accurate location estimates in most algorithms.

4. Conclusions and Future Work

We have provided a comprehensive review of distance-based localization technologies and techniques. Distance-based localization algorithms can be categorized as: 1) centralized, distributed, or distributed-centralized and 2) anchor-based or cooperative. We provided a detailed review of five distance-based localization algorithms and presented a realistic simulation comparison of these five algorithms using the connectivity and RSS ranging methods under various network configurations. From our comparisons we make the following three conclusions. Our first conclusion is that localization algorithm performance is dependant on network configuration. Thus, in selecting a localization algorithm for a WSN deployment it is important to understand the target topology of the application. Second, our simulations show that MDS-MAP and TSL generally outperform the other tested algorithms. And lastly, MDS-MAP outperforms TSL in most cases. In the future we will test TSL and MDS-MAP algorithm performance on a real-world medium sized geophysical WSN aligned in a uniform topology.

5. Acknowledgements

This work is supported in part by NSF Grant DGE-0801692.

References

[1] K. Whitehouse, C. Karlof, A. Woo, F. Jiang, and D. Culler, "The effects of ranging noise on multihop localization: an empirical study," in *Proceedings of Information Processing in Sensor Networks, IPSN '05*, Los Angeles, California, USA, 2005.

[2] M. Welsh, "Sensor networks for the sciences," *Communications of the ACM*, vol. 53, pp. 36–39, November 2010.

[3] Y. Liu, Z. Yang, X. Wang, and L. Jian, "Location, Localization, and Localizability," *Journal of Computer Science and Technology*, pp. 274–297, 2010.

[4] G. Mao and B. Fidan, *Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target Tracking*. IGI Global, 2009, ch. Introduction to Wireless Sensor Network Localization, pp. 1–32.

[5] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," *IEEE Personal Communications*, vol. 7, no. 5, pp. 28 –34, Oct. 2000.

[6] J. Blumenthal, R. Grossmann, F. Golatowski, and D. Timmermann, "Weighted Centroid localization in Zigbee-based sensor networks," in *Symposium on Intelligent Signal Processing, WISP '07*, 2007.

[7] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher, "Range-free localization schemes for large scale sensor networks," in *Proceedings of Mobile Computing and Networking, MobiCom '03*, San Diego, CA, USA, 2003.

[8] S. Guha, R. Murty, and E. G. Sirer, "Sextant: a unified node and event localization framework using non-convex constraints," in *Proceedings of Mobile Ad Hoc Networking and Computing, MobiHoc '05*, Urbana-Champaign, IL, USA, 2005.

[9] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz, "Localization from mere connectivity," in *Proceedings of Mobile Ad Hoc Networking and Computing, MobiHoc '03*, Annapolis, Maryland, USA, 2003.

[10] S. Rallapalli, L. Qiu, Y. Zhang, and Y. Chen, "Exploiting temporal stability and low-rank structure for localization in mobile networks," in *Proceedings of Mobile Computing and Networking, MobiCom '10*, Chicago, Illinois, USA, 2010.

[11] D. Niculescu and B. Nath, "Ad hoc positioning system (APS)," in *Proceedings of Global Telecommunications Conference, GLOBECOM '01*, 2001.

[12] J. A. Costa, N. Patwari, and A. O. Hero, "Distributed weighted-multidimensional scaling for node localization in sensor networks," *ACM Transactions on Sensor Networks*, vol. 2, pp. 39–64, 2005.

[13] D. Moore, J. Leonard, D. Rus, and S. Teller, "Robust distributed network localization with noisy range measurements," in *Proceedings of Embedded Networked Sensor Systems, SenSys '04*, Baltimore, MD, USA, 2004.

[14] K. Whitehouse and D. Culler, "A robustness analysis of multi-hop ranging-based localization approximations," in *Proceedings of Information Processing in Sensor Networks, IPSN '06*, Nashville, Tennessee, USA, 2006.

On the Energy Detection of Unknown Deterministic Signals over Generalized Fading Channel

O. Olabiyi, S. Alam, O. Odejide, A. Annamalai,

Center of Excellence for Communication Systems Technology Research
Department of Electrical and Computer Engineering
Prairie View A&M University, TX 77446

Abstract— **In this article, we develop new analytical techniques for the performance analysis of energy detection with selection diversity combining (SDC) and square-law selection (SLS) diversity schemes over generalized fading channels. We present two novel approaches in this work. The derived frameworks are based on the canonical series representation of Marcum Q-function with derivatives of moment generating function (MGF) of signal-to-noise ratio (SNR) of fading channel and single integral of cumulative distribution function (CDF) of SNR for SLS and SDC diversity schemes respectively. Using these frameworks, we found out that the performance of both SLS and SDC diversity schemes improves with smaller number of samples, increasing fading index, mean SNR and number of diversity branches. Also, comparing both diversity schemes shows that the choice of better detection diversity scheme varies with mean SNR. At low SNR, SDC performs better than the SLC scheme, but as the mean SNR increases, SLS gives better performance while high mean SNR values favours again the SDC scheme. More so, lower number of samples favours the SLS scheme than the SDC scheme. Although, the framework is applicable across fading channels, selected numerical results are presented on the Rice fading channel since there are limited results in literature in this area.**

Keywords— **Energy detection, Cognitive radio, Fading channels, Selection diversity combining, Square-law-selection**

1 Introduction

The emerging technology of cognitive radio has created a paradigm shift in the design of wireless system where radio can now adapt their operating behaviour to take advantage of unused spectrum. One of the main requirements of this system is to ensure that the incumbent (i.e. primary or licensed user) is not interrupted by the activity of these cognitive radios. Therefore the new radio must be capable of determining the presence or absence of an incumbent before spectrum usage. Among the various known spectrum sensing techniques, blind sensing using energy detectors is perhaps the simplest (i.e., low complexity) and more importantly, the "secondary users" do not require any unauthorized details or a priori knowledge of the "primary user" transmissions (i.e., the energy detector of a secondary user treats the received primary user transmission as an unknown deterministic signal). While energy detection of the received signal waveform over an observation time window may be more versatile than the cyclo-stationary feature detection approach, its performance (reliability) is severely limited by multipath fading. Therefore, performance characterization of energy detectors with diversity reception over different wireless fading environments is of practical interest.

Urkowitz, [1] first studied the detection of an unknown deterministic signal over a flat band-limited Gaussian noise channel using an energy detector. The results in [1] are further extended to Rayleigh, Rice and Nakagami fading channels in [2]. Average probabilities of detection (P_d) and false alarm probability (P_f) over Rayleigh, Rice and Nakagami-m fading channels are presented. The probability of detection was derived in closed-form for Rayleigh fading channel; however the Nakagami fading channels results to single integral solution while the Rice fading channel's result is an infinite summation. Nakagami-*m* and Rice fading channels are considered in [3] and [4]. The reported result for single Nakagami-*m* channel is limited to integer values of the shape parameter (*m*) whereas the result for Rice fading channel is restricted to unity time-bandwidth product. The performance of energy detector in diversity system such as Maximal ratio combining (MRC), square-law selection (SLS) and switch and stay (SSC) diversity detectors over i.i.d Rayleigh fading channels was considered in [3] and [4]. In [5]-[7], the Authors analyse the performance of the energy detector with selection diversity combining (SDC) in Nakagami-m channel, MRC in Nakagami-m and Rice channel and equal gain combining (EGC) in Nakagami-m channel respectively with all derived expression in closed form. In [8], performance with SSC, square-law combining (SLC) and SLS were considered. In all these cases, only integer *u,* integer Nakagami fading index *m* and integer *Lu* or *Lm* (where *L* is the number of diversity branches) are considered.

In [12], we presented using an alternative form of Marcum Q-function a new approach involving derivatives of moment generating function (MGF) of signal-to-noise ratio (SNR) of fading channels and we applied this method to MRC and SLC in Nakagami-m and Rice fading channels. This method easily tackles scenarios with non-integer *u,* non-integer Nakagami fading index *m*, *Lu* or *Lm* with all expressions involving single infinite summation term.

In this article, we develop novel approaches of analysing the performance of energy detector with SDC and SLS diversity schemes in generalized fading channels. The derived frameworks are based on the canonical series representation of Marcum Q-function with derivatives of MGF of SNR of fading channel and single integral of cumulative distribution function (CDF) of SNR for SLS and SDC diversity schemes respectively. However, simpler expressions were derived for special cases of SDC scheme for both Nakagami-m and Rice fading channels. Using these frameworks, we found out that the performance of both SLS and SDC diversity schemes improves with smaller number of samples, increasing fading index, mean SNR and number of diversity branches. Also, comparing both diversity schemes shows that the choice of better detection diversity scheme varies with mean SNR. Though the derived framework is applicable across fading channels, our numerical result is on Rice fading channel which has not been well researched in literature.

The rest of the paper is organized as follows. In section II we present system model, notations and performance over single channel. Section III extends this to SLS diversity system and develops performance analysis for SDC diversity scheme. Finally, numerical results and concluding remarks are presented in section IV and V respectively.

2 System model and notation

To be consistent, notations similar to [8] are used as listed below.

$s(t)$	: Unknown deterministic signal waveform
$n(t)$	: Noise waveform – White Gaussian random process
s_i	: Unknown deterministic signal waveform
n_i	: Noise waveform – White Gaussian random process
$r(t)$	: Received signal
h	: Channel coefficient amplitude
T	: Observation time interval
W	: One-sided bandwidth
$u = TW$	: Time-Bandwidth product
N_{01}	: One sided noise power spectral density
E_s	: Signal energy over the time interval T
λ	: Energy threshold of the receiver
L	: Number of branches of the receiver combiner
H_0	: Hypothesis 0; no $s(t)$ present
H_1	: Hypothesis 1; $s(t)$ present
χ^2_{2u}	: Central Chi-square distribution with $2u$ degrees of freedom
$\chi^2_{2u}(\epsilon)$	: Non central Chi-square distribution with $2u$ degrees of freedom and non centrality parameter ϵ

The detection of the existence of the unknown deterministic signal $s(t)$ by the receiver, is a binary hypothesis test as shown in [1, eq(1)],

$$y(t) = \begin{cases} n(t) & : H_0 \\ hs(t) + n(t) & : H_1 \end{cases} \tag{1}$$

Therefore, a sample from noise process n_i is a Gaussian random variable with zero mean and $N_{01}W$ variance; $n_i \sim N(0, N_{01}W)$ [1]. The energy detector decision variable Y can be expressed as [1, eq.(2)].

$$Y = \frac{2}{N_{01}} \int_0^T n^2(t)dt = \sum_{i=1}^{2u} \left(\frac{n_i}{\sqrt{N_{01}W}} \right)^2 \quad : H_0 \tag{2}$$

Thus, Y under H_0 is a square sum $2u$ Gaussian random variable of $N(0,1)$ and follows χ^2_{2u}. Similarly, Y under H_1 is formed as given by (3).

$$Y = \frac{2}{N_{01}} \int_0^T y^2(t)dt = \sum_{i=1}^{2u} \left(\frac{hs_i + n_i}{\sqrt{N_{01}W}} \right)^2 \quad : H_1 \tag{3}$$

Here, we assume that the channel coefficient amplitude is Y under constant over the $2u$ samples. It therefore follows that H_1 is $\chi^2_{2u}(\epsilon_j)$ where ϵ is given by [1, eq.(4)],

$$\epsilon = \sum_{i=1}^{2u} \left(\frac{hs_i}{\sqrt{N_{01}W}} \right)^2 = \frac{h^2}{N_{01}W} \sum_{i=1}^{2u} s_i^2 = \frac{2h^2 E_s}{N_{01}} = 2\gamma \tag{4}$$

Here, the SNR is defined by $\gamma = \frac{h^2 E_s}{N_{01}}$.

In [12], it has been shown that the detection and false alarm probabilities of an energy detector in AWGN channel is given by:

$$P_d = Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) \tag{5}$$

and

$$P_f = \frac{\Gamma(u, \frac{\lambda}{2})}{\Gamma(u)} \tag{6}$$

respectively, where $Q_u(.,.)$ is the generalised (u^{th} order) Marcum-Q-function and $\Gamma(.,.)$ is the upper incomplete gamma function which is defined by the integral form $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$ and $\Gamma(a,0) = \Gamma(a)$. P_f is the same over any fading channel since there is no γ in (6). In the other sense P_d has to be averaged over different fading channels and diversity combining.

Therefore, the average detection probability, $\overline{P_d}$ is given by

$$\overline{P_d} = \int_0^\infty Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) f_\gamma(\gamma) d\gamma \tag{7}$$

where $f_\gamma(\gamma)$ is the probability density function of SNR, γ.

Using the alternative representation of generalized Marcum Q- function, we have shown in [12] that it can be written as,

$$Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) = 1 - \sum_{k=0}^{\infty} \frac{\gamma^k e^{-\gamma}}{k!} \frac{G(u+k, \frac{\lambda}{2})}{\Gamma(u+k)} \tag{8}$$

where $G(.,.)$ is the lower incomplete gamma function which is defined by $G(a,x) = \int_0^x t^{a-1}e^{-t}dt$.

Hence, substituting (8) into (6), we obtain the generalized average P_d in this case, $\overline{P_{d\,Gen}}$ as

$$\overline{P_{d\ Gen}} = 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \phi_{\gamma}^{(k)}(s) \Bigg|_{s=1} \tag{9}$$

where $\phi_{\gamma}^{(k)}(s) = \dfrac{\partial^k \phi_{\gamma}}{\partial s^k}$ and $\phi_{\gamma}(s)$ is the MGF of SNR of different stochastic fading channels. The k^{th} derivatives of the MGF of common fading channels are listed in [12, Table 1].

For example, the final expression for detection probability is easily obtained for Nakagami-m and Rice Fading channels respectively as

$$\overline{P_{d\ Nak}} = 1 - \sum_{k=0}^{\infty} \frac{1}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \frac{\Omega^k m^m \Gamma(m+k)}{(m+\Omega)^{m+k}\Gamma(m)} \tag{10}$$

$$\begin{aligned}\overline{P_{d\ \mathrm{R}ic}} &= 1 - \sum_{k=0}^{\infty} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \frac{\Omega^k (k!)(1+K)}{(1+K+\Omega)^{1+k}} \exp\left(\frac{-K\Omega}{1+K+\Omega}\right) \\ &\times \sum_{i=0}^{k} \frac{1}{(i!)^2 (k-1)!} \left(\frac{K(1+K)}{1+K+\Omega}\right)^i\end{aligned} \tag{11}$$

The convergence and truncation error of this infinite series has been well treated in [12] and it has been shown that only few terms are required for four-digit accuracy. In fact, less than 25 terms are required for no diversity case across different system parameter changes. Next, we will apply this result in deriving the expression for the performance of energy detection over generalized fading channel with post-detection square-law selection (SLS).

3 Detection over generalized fading channel with diversity system

3.1 Square-law selection (SLS)

Since the branch with maximum output decision is been chosen, the effective decision variable can be written as in [8]

$$Y_{SLS} = \max(Y_1, Y_2, ..., Y_L) \tag{12}$$

Assuming independent branch statistics,

$$\begin{aligned}P_{f\,SLS} &= \Pr\{Y_{SLS} > \lambda \mid H_0\} = 1 - \prod_{l=1}^{L} \Pr\{Y_l < \lambda \mid H_0\} \\ &= 1 - \left(1 - \frac{\Gamma(u,\frac{\lambda}{2})}{\Gamma(u)}\right)^L\end{aligned} \tag{13}$$

Assuming independent branch statistics, we obtain,

$$\begin{aligned}P_{d\,SLS} &= \Pr\{Y_{SLS} > \lambda \mid H_1\} = 1 - \prod_{l=1}^{L} \Pr\{Y_l < \lambda \mid H_1\} \\ &= 1 - \prod_{l=1}^{L}\left(1 - Q_u(\sqrt{2\gamma_l}, \sqrt{\lambda})\right)\end{aligned} \tag{14}$$

Since there is no γ in (12), $P_{f\,SLS}$ is the same across all channels and branches. Averaging (13) over different channels we obtain similar to (8)

$$\overline{P_{d\ SLS}} = 1 - \prod_{l=1}^{L} \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \phi_{\gamma_l}^{(k)}(s) \Bigg|_{s=1} \right) \tag{15}$$

and for i.i.d branches we obtain

$$\overline{P_{d\ SLS}} = 1 - \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \phi_{\gamma_l}^{(k)}(s) \Bigg|_{s=1} \right)^L . \tag{16}$$

For example substituting the MGF of Nakagami and Rice channels produces respectively

$$\overline{P_{d\ SLS-Nak}} = 1 - \prod_{l=1}^{L} \left(\sum_{k=0}^{\infty} \frac{1}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \frac{\Omega^k {m_l}^{m_l} \Gamma(m_l+k)}{(m_l+\Omega)^{m_l+k}\Gamma(m_l)} \right) \tag{17}$$

$$\begin{aligned}\overline{P_{d\ SLS-Ric}} = 1 - \prod_{l=1}^{L} \Bigg\{ \Bigg(& \sum_{k=0}^{\infty} \frac{1}{k!} \frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)} \frac{\Omega^k (k!)^2 (1+K_l)}{(1+K_l+\Omega)^{1+k}} x \\ & \exp\left(\frac{-K_l\Omega}{1+K_l+\Omega}\right) \sum_{i=0}^{k} \frac{1}{(i!)^2 (k-1)!} \left(\frac{K_l(1+K_l)}{1+K_l+\Omega}\right)^i \Bigg) \Bigg\}\end{aligned} \tag{18}$$

3.2 Selection diversity combining (SDC)

Since the MGF and PDF of i.n.d channels is difficult to obtain in the case of SDC, we abandon the above MGF method to seek a more general solution taking advantage of readily available cumulative distribution function (CDF).

In the selection combining detection, the branch with maximum SNR is selected i.e.

$$\gamma_{SC} = \max(\gamma_1, \gamma_2 ... \gamma_L) \tag{19}$$

Therefore, the CDF of the output SNR of L-branch selection combiner is given by

$$F_{\gamma_{SC}}(\gamma) = \prod_{l=1}^{L} F_{\gamma_l}(\gamma) \tag{20}$$

where $F_{\gamma_l}(\gamma)$ is the CDF of branch l.

Since only the samples from a single branch are being considered at any point in time, the decision variable Y_{SC} is just the i.i.d χ_{2u}^2 for H_0 and $\chi_{2u}^2(\epsilon_j)$ for H_1 and this is defined by

$$Y_{SC} = Y_l \sim \begin{cases} \chi_{2u}^2 & : H_0 \\ \chi_{2u}^2(\epsilon_{SC}) & : H_1 \end{cases} \tag{21}$$

where the non-centrality $\epsilon_{SC} = \epsilon_l = 2\gamma_l = 2\gamma_{SC}$, hence the false alarm probability remains the same i.e.

$$P_{f_{SC}} = \frac{\Gamma(u,\frac{\lambda}{2})}{\Gamma(u)} \tag{22}$$

However, the detection probability is affected due to distribution of the SNR of the combiner's output and could be written from (7) as

$$\overline{P_{d_{SC}}} = \int_0^{\infty} Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) f_{\gamma_{SC}}(\gamma) d\gamma \tag{23}$$

where $f_{\gamma_{SC}}(\gamma)$ is the PDF of SNR of the combiner's output.

Our approach here is to first find a generalized solution to (23) and later present simpler solution for special cases. We will like to mention here that only [5] has considered this problem and their result is limited to i.i.d dual combiner (with non integer m) and i.i.d multiple branches with integer m with both cases resulting to solution with combination of infinite summation term and infinite series function [5, Eq. (5), (17), (21)]. In order to obtain amore compact and general solution for (23) that is applicable to i.n.d channels and mixed fading environment, we use the CDF method.

Using the integration by part, it is straight forward to show that

$$\overline{P_{d_{SC}}} = 1 - \int_0^\infty F_{\gamma_{SC}}(\gamma)\frac{\partial Q_u(\sqrt{2\gamma},\sqrt{\lambda})}{\partial\gamma}d\gamma \qquad (24)$$

Using the identity in [11 eq. (11)], we obtain

$$\overline{P_{d_{SC}}} = 1 - \int_0^\infty \left(\frac{\lambda}{2\gamma}\right)^{u/2} e^{-(\gamma+\frac{\lambda}{2})} I_u\left(\sqrt{2\gamma\lambda}\right)F_{\gamma_{SC}}(\gamma)d\gamma \qquad (25)$$

where $I_u(.)$ is the u-th order modified Bessel function of the first kind. Substituting (20) into (23), we obtain

$$\overline{P_{d_{SC}}} = 1 - \int_0^\infty \left(\left(\frac{\lambda}{2\gamma}\right)^{u/2} e^{-(\gamma+\frac{\lambda}{2})} I_u\left(\sqrt{2\gamma\lambda}\right)\prod_{l=1}^{L}F_{\gamma_l}(\gamma)\right)d\gamma \quad (26)$$

Eq. (26) can be evaluated numerically to obtain $\overline{P_{d_{SC}}}$. The CDF of different fade distributions are listed in [9, Table 3, pp. 420]. Therefore, for Nakagami-m and Rice i.n.d channel, (26) becomes,

$$\overline{P_{d_{SC-Nak}}} = 1 - \int_0^\infty \left(\left(\frac{\lambda}{2\gamma}\right)^{u/2} e^{-(\gamma+\frac{\lambda}{2})} I_u\left(\sqrt{2\gamma\lambda}\right)\prod_{l=1}^{L}\left(1-\frac{\Gamma(m_l,\frac{m_l}{\Omega_l}\gamma)}{\Gamma(m_l)}\right)\right)d\gamma \, (27)$$

and

$$\overline{P_{d_{SC-Ric}}} = 1 - \int_0^\infty \left[\left(\frac{\lambda}{2\gamma}\right)^{u/2} e^{-(\gamma+\frac{\lambda}{2})} I_u\left(\sqrt{2\gamma\lambda}\right)\right. , \qquad (28)$$
$$\left. x\prod_{l=1}^{L}\left(1-Q_1\left(\sqrt{2K_l},\sqrt{\frac{2(1+K_l)}{\Omega_l}\gamma}\right)\right)\right]d\gamma$$

Eq. (26) can be solved numerically using numerical integration methods with the aid of computer (i.e. using MATLAB or Mathematica) and it is easier to program, more compact and general across fading channels than the result in [5, (17), (21)] which holds for only Nakagami-m channel with dual SDC combiner and multiple combiner with integer fading index, m.

For the i.i.d channel scenario, the PDF is given by $f_{\gamma_{SC}}(\gamma) = Lf_\gamma(\gamma)\left(F_\gamma(\gamma)\right)^{L-1}$ and therefore, the average detection probability using (23) is given by

$$\overline{P_{d_{SC}}} = \int_0^\infty LQ_u(\sqrt{2\gamma},\sqrt{\lambda})f_\gamma(\gamma)\left(F_\gamma(\gamma)\right)^{L-1}d\gamma \qquad (29)$$

which after substituting (8) becomes

$$\overline{P_{d_{SC}}} = 1 - \sum_{k=0}^{\infty}\frac{L}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}\int_0^\infty \gamma^k e^{-\gamma} f_\gamma(\gamma)\left(F_\gamma(\gamma)\right)^{L-1}d\gamma \qquad (30)$$

Going forward, we split the derivation for Nakagami and Rice fading channel; though the same principle is applicable to both in order to derive more compact solution for some special cases.

3.2.1 Case 1: Nakagami-m fading channel

For Nakagami-m channel, (33) becomes

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{L}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}$$
$$\times\int_0^\infty \gamma^{m+k-1}e^{-(1+\frac{m}{\Omega})\gamma}\left[1-\frac{\Gamma(m,\frac{m}{\Omega}\gamma)}{\Gamma(m)}\right]^{L-1}d\gamma \qquad (31)$$

For the special case of dual diversity combiner, (31) becomes

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{2}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}$$
$$\times\int_0^\infty \gamma^{m+k-1}e^{-(1+\frac{m}{\Omega})\gamma}\frac{G(m,\frac{m}{\Omega}\gamma)}{\Gamma(m)}d\gamma \qquad (32)$$

Using the identity in [10, 6.455-2], we obtain

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{2}{\left(\Gamma(m)\right)^2}\left(\frac{m}{\Omega}\right)^{m+2} \sum_{k=0}^{\infty}\frac{1}{k!}\frac{\Gamma(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}$$
$$\times\frac{\Gamma(2m+k)}{m(1+\frac{2m}{\Omega})^{2m+k}}\,{}_2F_1\left(1,2m+k;m+1;\frac{m}{2m+\Omega}\right) \qquad (33)$$

where ${}_2F_1(.,.;.;.)$ is Gauss hypergeometric function.

When m is an integer, (31) can be expressed using Binomial expansion as

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{L}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}$$
$$\times\int_0^\infty \gamma^{m+k-1}e^{-(1+\frac{m}{\Omega})\gamma}\sum_{q=1}^{L-1}\frac{(-1)^q}{q!}\left[\frac{\Gamma(m,\frac{m}{\Omega}\gamma)}{\Gamma(m)}\right]^q d\gamma \qquad (34)$$

which using the identity [10, 8.352-4] produces

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{L}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}$$
$$\times\int_0^\infty \gamma^{m+k-1}e^{-(1+\frac{m}{\Omega})\gamma}\sum_{q=1}^{L-1}\frac{(-1)^q}{q!}e^{\left(-q\frac{m}{\Omega}\gamma\right)}\left[\sum_{j=0}^{m-1}\frac{\left(\frac{m}{\Omega}\gamma\right)^j}{j!}\right]^q d\gamma \qquad (35)$$

which after multinomial expansion becomes

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{L}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}\sum_{q=1}^{L-1}\frac{(-1)^q}{q!}$$
$$\times\sum_{j=0}^{q(m-1)}\zeta_j(m,q,\Omega)\int_0^\infty \gamma^{m+k+j-1}e^{-(1+\frac{m}{\Omega}(1+q))\gamma}d\gamma \qquad (36)$$

where $\zeta_j(m,q,\Omega)$ is the coefficient of γ^j in the multinomial expansion of $\left[\sum_{j=0}^{m-1}\frac{\left(\frac{m}{\Omega}\gamma\right)^j}{j!}\right]^q$ which is given by

$$\zeta_j(m,q,\Omega) = \left(\frac{m}{\Omega}\right)^j\sum_{i=j-m+1}^{j}\frac{\beta_{i(q-1)}}{(j-i)!}I_{[0,(q-1)(m-1)}(q),$$
$$I_{[a,c]}(b) = \begin{cases}1 & a\le b\le c\\ 0 & otherwise\end{cases} \qquad (37)$$

where $\beta_{00} = \beta_{0q} = 1, \quad \beta_{j1} = 1/j!, \quad \beta_{1q} = q$.

Using the identity in [10, 3.326-2], (36) then yields,

$$\overline{P_{d_{SC-Nak}}} = 1 - \frac{L}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G(u+k,\frac{\lambda}{2})}{\Gamma(u+k)}\sum_{q=1}^{L-1}\frac{(-1)^q}{q!}$$
$$\times\sum_{j=0}^{q(m-1)}\zeta_j(m,q,\Omega)\frac{\Gamma(m+k+j)}{\left(1+\frac{m}{\Omega}(1+q)\right)^{m+k+j}} \qquad (38)$$

3.2.2 Case 2: Rice fading channel

For Rice channel, (30) becomes

$$\overline{P_{d_{SC-Ric}}} = 1 - \sum_{k=0}^{\infty}\frac{L}{r!}\frac{G(u+r,\frac{\lambda}{2})}{\Gamma(u+r)}\int_0^\infty \gamma^r e^{-\gamma}\frac{1+K}{\Omega}\exp\left(-K-\frac{(1+K)\gamma}{\Omega}\right)$$
$$\times I_o\left(2\sqrt{K(1+K)\gamma/\Omega}\right)\left(1-Q_1\left(\sqrt{2K},\sqrt{\frac{2(1+K)}{\Omega}\gamma}\right)\right)^{L-1}d\gamma \qquad (39)$$

For special case of dual selection combiner (38) becomes

$$\overline{P_{d_{SC-Ric}}} = 1 - \sum_{r=0}^{\infty} \frac{2}{r!} \frac{G(u+r,\frac{\lambda}{2})}{\Gamma(u+r)} \Bigg[(-1)\phi_{\gamma}^{(r)}(s)\,|_{s=1} - \frac{1+K}{\Omega} e^{-K} (-1)^r \frac{\partial^r}{\partial s^r} \quad (40)$$

$$\left(\int_0^{\infty} \mathrm{e}^{-\left(s+\frac{(1+K)\gamma}{\Omega}\right)} I_o\left(2\sqrt{K(1+K)\gamma/\Omega}\right) Q_1\left(\sqrt{2K}, \sqrt{\frac{2(1+K)}{\Omega}\gamma}\right)\right)_{s=1} \Bigg] d\gamma$$

which after manipulation and using [13, eq. (46)], we obtain

$$\overline{P_{d_{SC-Ric}}} = 1 - \sum_{r=0}^{\infty} \frac{2}{r!} \frac{G(u+r,\frac{\lambda}{2})}{\Gamma(u+r)} \left((-1)\phi_{\gamma}^{(r)}(s)\,|_{s=1}\right) - 2Ae^{-K}$$

$$\times(-1)^r \left[\frac{\partial^r}{\partial s^r} I\Big(2\sqrt{A}, \sqrt{2K}, \sqrt{2AK}, \sqrt{2(s+A}\Big) \right]_{s=1} \quad (41)$$

where $A = \frac{1+K}{\Omega}$ and

$$I(a,b,c,d) = \frac{1}{d} e^{\frac{c^2}{2d}} Q_1\left(b\sqrt{\frac{d}{(d+a^2)}}, \frac{ac}{\sqrt{d(d+a^2)}} \right) - .$$

$$\times \frac{a^2}{d(d+a^2)} e^{\frac{c^2-b^2d}{2(d+a^2)}} I_0\left(\frac{abc}{(d+a^2)}\right)$$

Substituting $\phi_{\gamma}^{(r)}(s)\,|_{s=1}$ for Rice fading we obtain,

$$\overline{P_{d_{SC-Ric}}} = 1 - \sum_{r=0}^{\infty} \frac{2}{r!} \frac{G(u+r,\frac{\lambda}{2})}{\Gamma(u+r)} \left(\frac{\Omega^r (r!)(1+K)}{(1+K+\Omega)^{1+r}} \exp\left(\frac{-K\Omega}{1+K+\Omega}\right) \sum_{i=0}^{r} \frac{1}{(i!)^2 (r-1)!} \quad (42)$$

$$\times \left(\frac{K(1+K)}{1+K+\Omega}\right)^i - (-1)^r 2Ae^{-K} \left[\frac{\partial^r}{\partial s^r} I\Big(2\sqrt{A}, \sqrt{2K}, \sqrt{2AK}, \sqrt{2(s+A}\Big) \right]_{s=1} \Bigg)$$

The r^{th} derivative of $I(a,b,c,d)$ could be computed using MAPLE. For higher order configuration, the derivation becomes involved and therefore it is recommended to use numerical evaluation in (28) derived for i.n.d Rice channel.

4 Numerical results

In this article, we analyse the performance of energy detector using the complementary ROC curves and detection probability. Here, we focus on the Rice fading as there is limited information on the performance of energy detector in Rice fading in literature. First we study the effect of detection parameters on the detection performance of SLS and SDC diversity schemes and then compare the two selection schemes. Fig. 1 and 2 shows the effect of change in number of diversity branches, L, fading index, K, and time-bandwidth product, u (which is a function of number of samples). It is observed that that both diversity schemes perform better as the number of diversity branches and fading index is increases. Also, lower number of samples improves the performance. The effect of change in different combinations of the three parameters can also be observed from both figures. Fig. 3 shows the comparison of the two diversity schemes for different mean SNR values. We observe that the choice of better diversity scheme for energy detection varies with mean SNR. At low mean SNR, SDC is the better scheme but as the mean SNR increases, the performance of SLS becomes better than that of SDC while high mean SNR values favours again the SDC scheme. Due to this overlap, we decided to investigate the variation of average detection probability, $\overline{P_d}$ with mean SNR for both schemes. The result is shown in Fig. 4 and it confirms the observation in Fig. 3. Also, Fig. 4 indicates that lower number of samples favours the SLS diversity scheme more than the SDC. This is expected as the performance of SLS is more affected by the sampled signal, because the selection of best detection path or circuitry is done after sampling unlike the SDC, in which the best path is selected before sampling. However, in practical implementation, SLS requires more number of detection circuitry than SDC which requires only one detection circuitry.

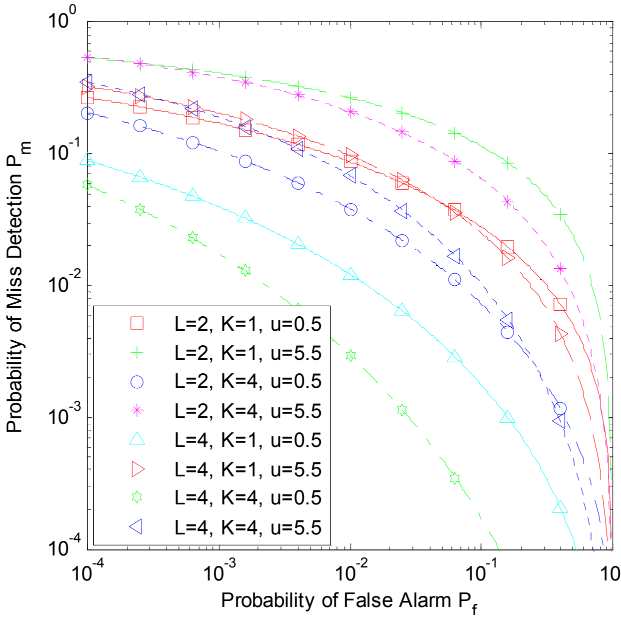


Fig. 1 Complementary ROC curves of SLS diversity scheme over Rice channels K= {1, 4}, for u= {0, 5.5}, L ={2, 4}, and mean SNR=10 dB.

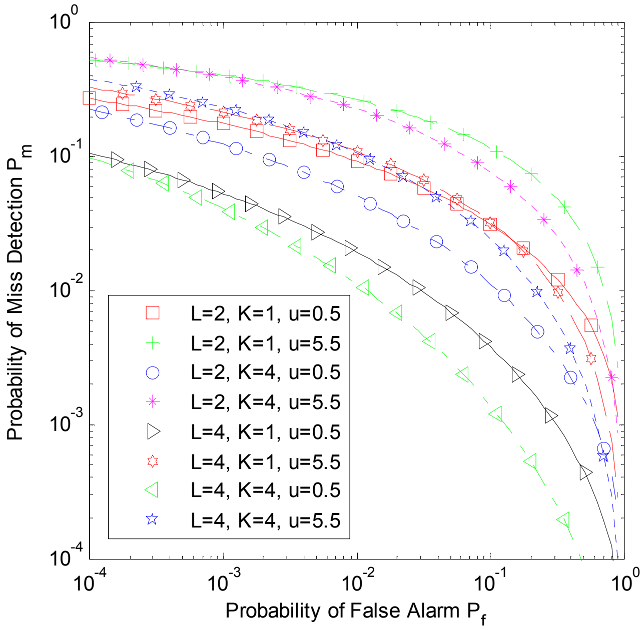


Fig. 2 Complementary ROC curves of SDC diversity scheme over Rice channels K= {1, 4}, for u= {0, 5.5}, L = {2, 4}, and mean SNR=10 dB.

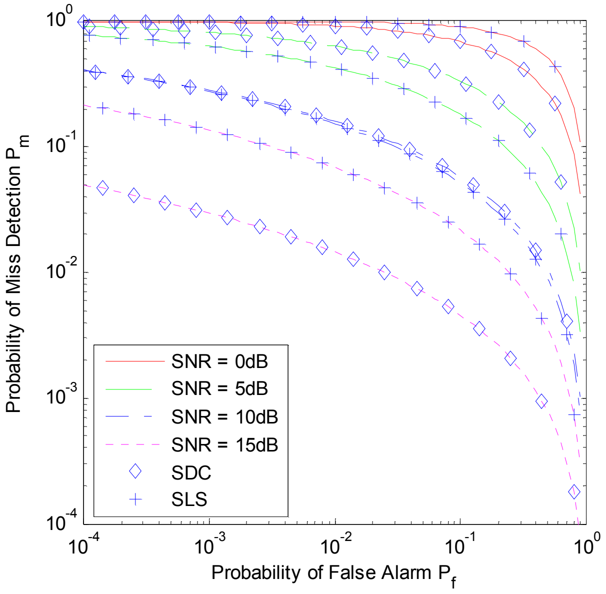


Fig. 3 Complementary ROC curves for dual SLS and SDC over Rice channels (K=2) for u=2.5 and mean SNR= {0 ,5, 10, 15}.

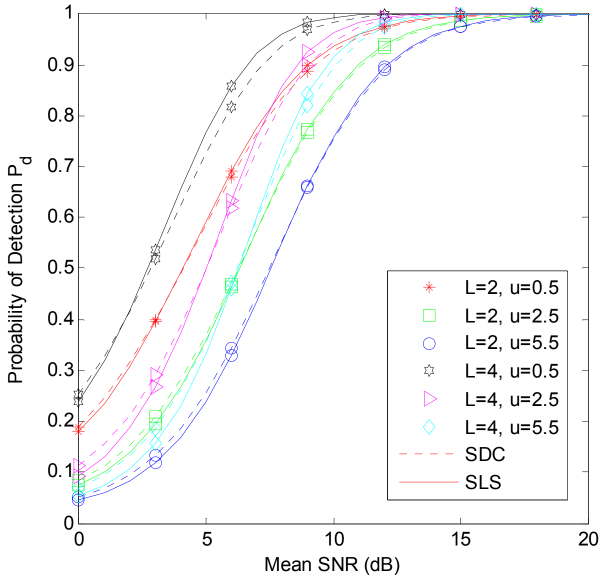


Fig.4 Average probability of detection for SLS and SDC over Rice channel (K=2) for L= {2,4}, u= {0.5, 5}, and P_f=0.01.

5 Conclusions

In this paper, the detection an unknown deterministic signal by using an energy detector has been considered with detectors' selection diversity system. Both SDC and SLS diversity have been analyzed. We derived an expression for the performance analysis of both diversity systems. The derivation for SLS performance is based on the canonical series form of Marcum Q-function in conjunction with derivatives of MGF of SNR while a novel technique resulting into a single integral expression was presented for SDC scheme (for branches with i.n.d statistics) with closed form expressions derived for special cases.

With these mathematical frameworks, we are able to show that the choice of better detection diversity scheme between SLS and SDC depends on the mean SNR values. Careful consideration of related publications revealed that this simplified approaches have not been applied in analyzing the performance of the energy detector until now and to the best of our knowledge, limited analysis has been carried out on the performance of energy detector in Rice fading channel.

We have shown that our approach, unlike other methods which are incapable of handling half-integer u and Lu, produces result for these cases. These results could be readily used in deciding suitable diversity scheme, number of diversity branches and the energy threshold value required to achieve a specified false alarm and detection rate for different scenario of energy detector receiver in cognitive radio system.

Acknowledgment

This work is supported in part by funding from the US Army Research Office (W911NF-04-2-0054), Clarkson Aerospace, and National Science Foundation (0931679 & 1040207).

References

[1] H. Urkowitz, "Energy detection of Unknown deterministic Signals," in *Proc IEEE,* vol. 55, no. 4, pp. 523-531, Apr 1967.

[2] V.I. Kostylev, "Energy detection of a signal with random amplitude," *IEEE Int. Conf. ICC* 2002, vol. 3, pp. 1606 – 1610, Apr-May 2002.

[3] Fadel F. Digham, Mohamed-Slim Alouni and Marvin K. Simon, "On the Energy Detection of Unknown Signals Over Fading Channels," *IEEE Int.Conf. ICC'03*, vol 5, pp. 3575 - 3579, May 2003.

[4] Fadel F. Digham, Mohamed-Slim Alouni and Marvin K. Simon, "On the Energy Detection of Unknown Signals Over Fading Channels," *IEEE Trans. Commun*., vol 55, no.1,pp.21-24, Jan. 2007.

[5] S.P. Herath, N. Rajatheva, C. Tellambura,"On the energy detection of unknown deterministic signal over Nakagami channels with selection combining," *IEEE CCECE'09,*2009, pp. 745–749, May 2009.

[6] S.P. Herath, N. Rajatheva, C. Tellambura, "Unified Approach for Energy Detection of Unknown Deterministic Signal in Cognitive Radio Over Fading Channels" in *Proc. IEEE ICC'09*, 2009, pp. 1-5, June 2005.

[7] S.P. Herath, N. Rajatheva, "Analysis of Equal Gain Combining in Energy Detection for Cognitive Radio over Nakagami Channels," *IEEE GLOBCOM'08,* 2008, pp. 1-5, Dec. 2008.

[8] S.P. Herath, N. Rajatheva, "Analysis of Diversity Combining in Energy Detection for Cognitive Radio over Nakagami Channels," *Proc. IEEE ICC'09,* June 2009, Germany.

[9] M. K. Simon and M-S Alouini, *Digital Communication over Fading Channels*, New York: Wiley, 2 edition, 2005.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products,* 5th ed., San Diego, CA: Academic, 1994.

[11] A. Annamalai, C. Tellambura, "A simple exponential integral representation of the generalized Marcum Q-function QM (a, b) for real-order M with applications," in *Proc* 54th *IEEE MILCOM 2008,* 2008, pp. 1 - 7, Nov. 2008.

[12] A. Annamalai, O. Olabiyi, S. Alam, O. Odejide, and D. Vaman, "Unified Analysis of Energy Detection of Unknown Signals over Generalized Fading Channels," to appear in *Proc. IEEE IWCMC 2011* Conference , Sept, 2011, Turkey, Intanbul.

[13] A. H. Nuttall, "Some integrals involving the Q function," Naval Underwater Systems Center, New London Lab, 1972.

Service Discovery in Urban Cooperative Networks

Yong Shen[+], Shafique Ahmad Chaudhry*
** Department of Computer Science, Al-Imam Muhammad bin Saud University, Saudi Arabia*
[+]Mobile and Internet Systems Laboratory, Department of Computer Science,
University College Cork, Ireland

Abstract

The recent technological developments in communication networks have results into a broad range of networks, e.g.,Wi-Fi, Mobile ad-hoc networks(MANET), 3G and other cellular networks, wireless sensor networks, etc. It is envisioned that such diverse networks must cooperate with each other to form cooperative urban networks, for the sake of making use of all the services and resources. Such integration poses great challenges for service discovery, for example, interoperability, robustness, scalability and context-awareness. In this paper, we propose a novel hierarchical architecture to meet the requirements of service discovery in cooperative urban networks, in which we assume various urban networks are connected to overlay infrastructure through cooperative gateway to form cooperative urban networks. We apply hierarchical architecture to urban networks and overlay infrastructure network in order to discover services using optimal service selection algorithm to take into account context-awareness issue. We also compare the performance of our architecture with distributed and centralized architectures.

Key Words: Service Discovery, Cooperative Urban Networks, Urban Networks, Service Discovery Performance Evaluation

1. Introduction

The recent technological developments in communication networks have resulted into a broad range of networks, e.g., Wi-Fi, Mobile ad-hoc networks, 3G and other cellular networks, and wireless sensor networks, to name a few. Unfortunately, these networks do not cooperate with each other to fully utilize the available data and network resources. It is desired for all different brands of communication technologies to cooperate with each other forming cooperative urban networks. This envisioned integration would make the resource sharing possible within various networks, maximizing the utilization of available data and other resources. Service discovery across the cooperative urban networks (CUN) is essential to achieving such sophistication.

A service discovery protocol enables devices to automatically advertise, discover and configure services, and at the same time facilitate communication with service providers to access desired services. Service discovery and provisioning in CUN brings in various challenges for its architectural framework designers. Firstly, service discovery architecture in CUN must provide a highly sophisticated and seamless interoperability to ensure that users can discover services across different types of networks. Secondly, the service discovery mechanism needs to be scalable and flexible which can adapt well to the change of network size. Thirdly, it also needs to be context-aware which means the service discovery mechanism should be aware of and make changes according to user context.

Over the past few years, many organizations have designed and developed a lot of service discovery protocols but none of existing protocols explicitly addresses the idea in the context of cooperative urban networks.

In this paper, we discuss the suitability of existing service discovery proposal for CUN. We also propose novel service discovery architecture to meet the requirements of service discovery in such environments and validate our solution through simulations.

The rest of the paper is organized as follows. In section 2 we outline the service discovery challenges for CUN, which is followed by our solution architecture in section 3. We present implementation details and preliminary evaluation results in section 4 and 5 respectively. A brief review of the state of the art is given in 6 and paper is concluded in section 7.

2. Service Discovery in Cooperative Urban Networks

Cooperative Urban networks consists of a broad range of networks, e.g., Wi-Fi [3], Mobile ad-hoc networks [4], 3G [5] and other cellular networks, Bluetooth networks[6], wireless sensor networks [7], which cooperate with each other to form a network where information and services can be shared in a ubiquitous fashion . An application of such scenario is U-city - a futuristic city planned to be equipped with services and functionalities of a ubiquitous environment [24]. Consider the scenario presented in figure 1. Fire breaks out in say a shopping mall. Sensing devices embedded in such a commercial centre detect fire and notify it to the municipal machinery through IPv6 network. Context-specific information is further retrieved from the disaster area and disseminated to respective personnel. Later various networks including Wi-Fi, Sensor Networks, Bluthehooth, etc. are involved in extending the operational management and inter-department coordination between the fire department, ambulance service, counter terrorism squad, Forensics intelligent transportation system (ITS), insurance and investigation, police, and other related enterprise networks. With existing protocols, generally, clients and services do not discover services if they do not use a common protocol. This limitation means that there is a need to establish interoperability mechanism among service discovery protocols. So in a CUN a pervasive discovery mechanism should be able to find services within its own network as well as across other networks or Internet. In order to make a service discovery work in pervasive environment, heterogeneity in hardware and software platforms, network protocols, and service discovery protocols is to be considered.

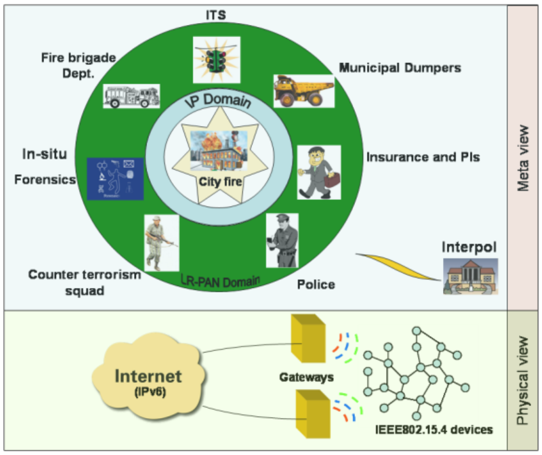


Figure 1. U-City Scenarios

Service discovery in CUN has to meet more challenges than to operate in a single subnet or LAN. The main challenges are:

Interoperability: CUN consists of variety of networks with different bandwidths, different packet sizes, and different capability of nodes resources and so on. Therefore, in order to access services across these networks, we must provide interoperability between different service discovery mechanism and network types.

Scalability: CUN consists of extremely large number of nodes and most service discovery systems' performance decrease sharply when network size goes large. We need an architecture which could localize the communication between entities and reduce the broadcast and multicast and provide scalability.

Robustness: In CUN, various types of networks connected together to provide a large number of services. System should not be vulnerable to one point failure. We must design a robust architecture against one point failure.

Context-awareness: Context-awareness [23] can be described as the ability to use situational information to provide highly relevant information. Context awareness is often a difficult task for designing a system, however, it is essential to consider the context of a situation to provide appropriate services with a desired quality of service.

3. Hierarchical Service Discovery for CUN

Figure 2 gives a conceptual model of cooperative urban networks where various networks are interworking to provide better services to different levels of users.

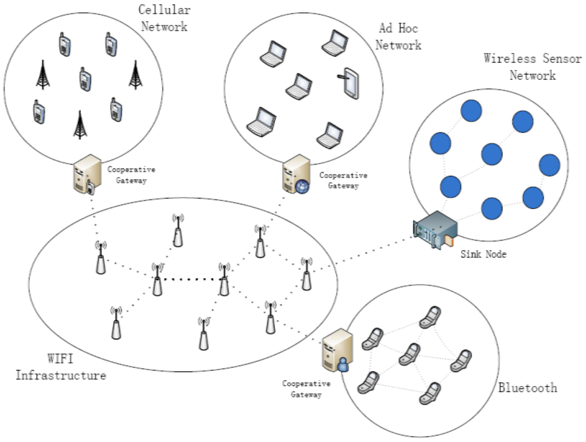


Figure 2. Various Technologies working together

In order to find a tradeoff between scalability and traffic overhead, we propose novel hierarchical architecture for CUN where an overlay network is established to which other sub-networks are attached. For example, in figure 2, Wi-Fi network is the overlay network to which sensor network, buletooth network, mobile adhoc network, and a cellular networks are attached as subnetworks. The overlay networks serves as the backbone of the whole system. Each network may support its own service discovery mechanism but for the cooperating between the networks, we provide *cooperative gateways* which are dedicated components that not only provide interoperability but also can take the role of service directory. And one sub-network's failure won't affect the robustness of the whole discovery system. In our proposed hierarchical architecture, as the simplest situation, the number of urban networks is small, we apply distributed architecture to overlay network and centralized architecture to sub-networks in which cooperative gateways play the role of directory. For extremely large number of networks, a multi-overlay architecture also can be considered. Followings are the major entities in our framework:

User Agent (UA): UA stands for the client requesting for the certain service. In our architecture, we have UAs both from overlay network and sub-networks.

Service Agent (SA): SA stands for the entity providing certain service. There are also different SAs from overlay network and sub-networks with different behavior.

Cooperative Gateway (CG): CG provides services' cache within a sub-network on behalf of the service directory. Also, it has dual network interface for providing interoperability to sub-networks and overlay network. CGs maintain two cache service. The local cache (LC) is used for directory service for sub-networks while the external cache (ENC) is used for historical service discovery record of overlay network.

The following messages are supported to facilitate the communication between the entities:

Directory Advertisement Message (DADV): DADV is only used in sub-networks. DADV is created by CG and broadcasted to the sub-networks to announce the existence of cache service provided by CG.

Service Request Message (SREQ:) SREQ is created when UA request for certain service. It includes the type of requested service and the UA's address. It can be sent by unicasting to CG in sub-networks and also can be sent by broadcasting in overlay network.

Service Reply Message (SREP): SREP is created when certain SREQ is matched. On receiving SREP, UA can establish communication with SA which has the service he requested

Service Registration Message (SREG): SREG is only used in sub-networks. On receiving DADV from CG, SA sends SREG to register service with the CG. SREG specifies the service it provide and lifetime of the registration. SA can send another SREG with new lifetime when the ole lifetime expires or the registration will be removed by the CG.

Service Acknowledgement Message (SACK): SACK is also only used in sub-networks. SACK is created by CG to acknowledge the registration of certain SA.

3.1 Service Discovery Protocol for Cooperative Urban Networks (SDPCUN)

In our novel hierarchical service discovery architecture, UAs stands for clients applications while services are represented as SAs. CGs are nodes with dual network interfaces with both sub-networks and overlay network. CGs are responsible for communication between different types of networks and also provide cache service for certain proximity.

The UA issues SREQ specifying the service type which is required. The UA should receive a SREP specified the most optimal service matching to the SREQ. The hierarchical architecture we used allows UAs to send SREQ to CG using unicast within sub-networks and broadcast SREQ in overlay network. The main flow is shown in figure 3.

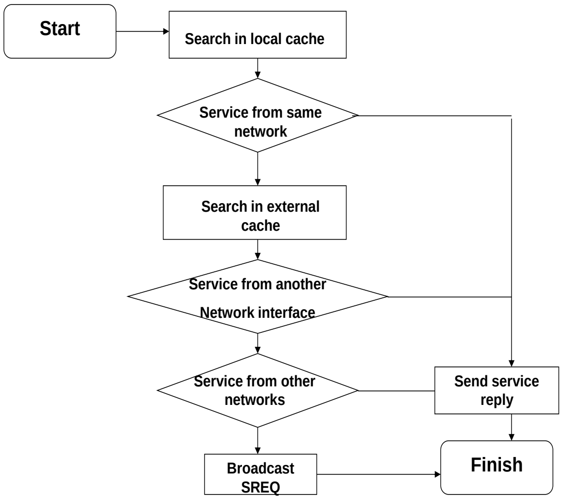

Start
Search in local cache
Service from same network
Search in external cache
Service from another Network interface
Service from other networks
Send service reply
Broadcast SREQ
Finish

Figure 3: Service Discovery

SAs within the sub-network register themselves with CGs by unicasting SREG and receiving unicasting SACK. The registration must be refreshed by sending another SREG or they will expire after the lifetime specified in SREG..

CGs provide directory service for sub-networks. UAs and SAs within the sub-networks discover the CGs by receiving advertisements which are broadcasted periodically from the CGs. There is an alternative way---when a node first comes into the sub-network, it can broadcast SREQ to its neighbors and the neighbors will send back the CG's information they hold.

UAs from the overlay network can discover SAs by two methods---active method and passive method. In active method, UA broadcast SREQ when it needs a certain type of service and SA receiving SREQ sends back SREP when matching the request service. Also, the UAs have caches restoring the historical service discovery. When request a same service next time, UA will unicast the SREQ directly to the SAs. In passive method, SAs broadcast its service information to the network periodically and UAs cache the service information. When UAs request a certain type of service, they search in their caches and unicast the SREQ to SAs. The existence of multiple CGs provides scalability and interoperability to the system, whereas the distributed architecture of overlay network provides the robustness to the system.

3.2 Interoperability

To meet the interoperability challenges we classify the service discovery into four scenarios:

Service discovery inside sub-networks: In this scenario, UA unicasts SREQ to CG and CG finds match of SREQ in its LC. This is a typical scenario of centralized service discovery architecture.

Service discovery from sub-network to overlay network: In this scenario, UA unicasts SREQ to CG and CG can not find match for the request service in LC. Then CG broadcasts SREQ in overlay network and SA from the overlay network match the service requested and send back SREP. The CG provides interoperability to overlay network and sub-networks. Figure shows the scenario.

Service discovery within overlay network: In this scenario, UA in overlay network broadcasts SREQ and the SA also from the overlay network matches the SREQ and sends back SREP. This is a typical scenario of distributed architecture.

Service discovery from overlay network to sub-networks: In this scenario, UA in overlay network broadcasts SREQ and one CG match the requested service and sends back SREP. Then the UA establish communication with SA from the sub-networks through one CG. The CG provides interoperability to overlay network and sub-networks. Figure shows the scenario.

Service discovery from one sub-network to other sub-networks: In this scenario, UA from one sub-network sends SREQ to CG of its sub-network. CG can not find match in its LC and then broadcast SREQ to overlay network. Another CG from another sub-network matches the requested service and sends back SREP. Then the UA establish communication with SA from another sub-network through two CGs. The two CGs provides interoperability to different sub-networks.

3.3 Cooperative Gateway

CG provides interoperability to the system. One CG has two network interfaces and has four interfaces with UAs and SAs from both sub-network and overlay network. Figure 4 shows the CG's architecture.

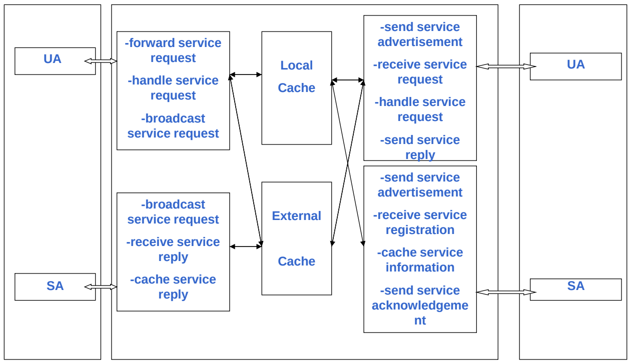

UA
SA
-forward service request
-handle service request
-broadcast service request
-broadcast service request
-receive service reply
-cache service reply
Local Cache
External Cache
-send service advertisement
-receive service request
-handle service request
-send service reply
-send service advertisement
-receive service registration
-cache service information
-send service acknowledgement
UA
SA

Figure 4: Cooperative Gateway Architecture

In sub-networks, we apply centralized service discovery architecture in which CGs localize the communication within the sub-network and produce less traffic. Thus, when the sub-networks go large, the system provides scalability. When a new type of urban network joins in the CUN, we only need to add one CG to allow the urban network to connect to overlay network. In overlay network, we apply distributed architecture in the assumption of the overlay network size is not too large. When overlay network's size increase, distributed architecture will suffer from is bad performance. To deal with this matter, we introduce multi-overlay in which we apply distributed architecture in the top overlay and centralized architecture in low overlay networks.

4. Preliminary Evaluation

We simulate our architecture in OMNET++ version 4 [23]. In our simulation we used an overlay network of 10 nodes, 3 of them are SA, while remaining are UA or sub-networks. Each Sub-network consists of 20 nodes, 7 of them are SA, one CG and others are UA. The link delay of overlay network is 100ms, while 200ms in sub-networks. We have compared the hierarchical architecture with fully-centralized and fully-distributed architectures. Followings are the parameters of our performance metrics:

Service discovery overhead: We consider the messages created and transmitted as the overhead for service discovery because they consume not only the resource of devices but also the network bandwidth. A better service discovery architecture should generate less overhead traffic.

Service discovery delay: Service discovery delay means the time interval between a device issue request for service and it gets reply for the requested service. A better service discovery architecture should show low mean service discovery delay.

Cache operation: Cache stores the historical service request record for future use. The fresh rate and capacity of cache affects the whole performance of service discovery architecture.

4.1 Service discovery overhead

At first, the whole network is the overlay network. Then every time we replace one UA in overlay with one sub-network to enlarge the network size to see the change of service discovery overhead. Figure 5 shows the results of service discovery overhead.

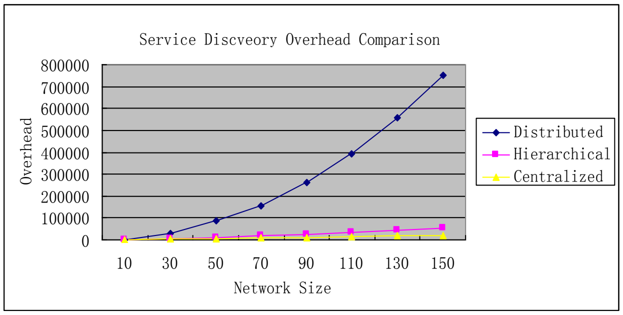


Figure 5: Service discovery overhead

The fully-distributed architecture shows an exponential increase when the network size goes large and the hierarchical and fully-centralized architecture only experience a mild linear increase. Our architecture has little more overhead than centralized architecture because we introduce distributed architecture in overlay network which increases the overhead. However, our architecture provides robustness to the service discovery system while lack of robustness is the fatal drawback of centralized architecture.

4.2 Service discovery delay

As the same, the whole network is the overlay network at first. Then every time we replace one UA in overlay with one sub-network to enlarge the network size to see the change of mean service discovery delay. There is a sharp increase of delay when introducing one, two and three sub-networks. This is because the difference of link delays in sub-networks and overlay network. When network size is small, introducing one sub-network will dramatically increase the average link delay of nodes. Also, the mean service discovery delay is lower in fully-distributed and our architecture than in centralized architecture. However, when the network size goes large, the increase of average link delay is not apparent; we can see from the figure 6 that the mean service discovery delay of fully-distributed architecture shows a sharp increase while mean delay in fully-centralized and our architecture only experience a slow increase.

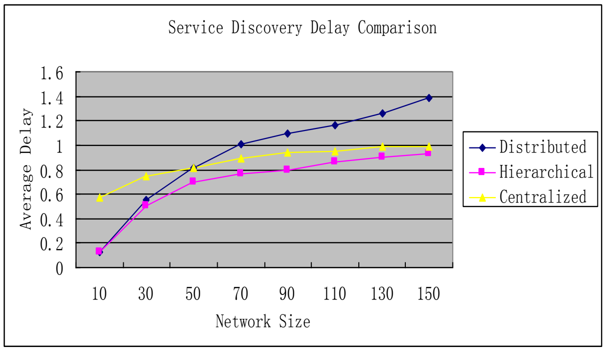


Figure 6: Service discovery delay

4.3 Cache Operation

Cache in CG plays a very important role in our service discovery architecture. When UA requests service not from the same network as itself, without external cache, cooperative gateway must broadcast service request in overlay network, which may induce a lot of traffic and enlarge the service discovery delay. The effect of cache operation varies greatly with different fresh rate and capacity.

To evaluate the cache effect, we set up our simulation environment as follows. The overlay networkconsists of 10 nodes 3 of them are SA 5 of them are UA and 2 sub-networks of 20 nodes each, 7 of them are SA, one CG and others are UA. Each SA provides different services. Therefore, there will be 17 services. UA randomly requests one of 17 services every 5 seconds.

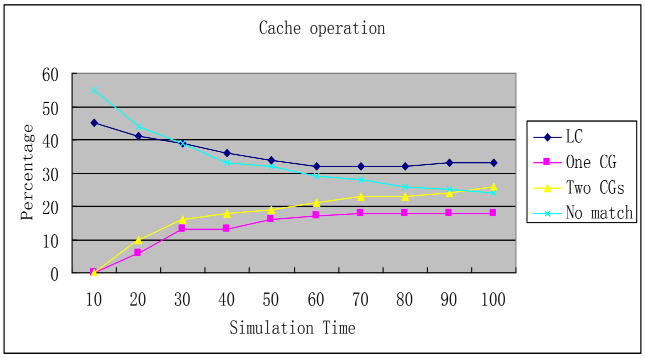


Figure 7: Cache operation

As shown in figure 7, at first, because cache does not come to work, the no match for service request in CG is very high. As the cache operates, the no match rate drops and the rate of available services from overlay network and other sub-network grows. At last, although one sub-network only provides 40% services, there are 75% services are available in CG. The available service will increase if we enlarge the capacity of cache or the fresh rate.

6. Related Work

Various solutions have been designed and developed for service discovery protocols. Some of them are already commercial versions bound to certain operating systems---for instance, Sun Microsystems' Jini Network Technology [8], Microsoft's Universal Plug and Play (UPnP) [9], Apple's Rendezvous [10], Salutation [11] protocol, Internet Engineering Task Force's [12] Service Location Protocol (SLP) [13], and Bluetooth SDP[14].

UPnP is a proposed architecture for service advertisement and discovery supported by the Microsoft. Unlike Jini's mobile code, UPnP aims to standardize the protocols used by devices to communicate, using XML [15].

Simple service discovery protocol (SSDP) [16], the protocol used in UPnP for service discovery and advertisement. Simple object access protocol (SOAP) [17], a protocol for remote procedure calls based on XML and HTTP. Generic Event Notification Architecture (GENA) [18] is a UPnP subscription-based event notification service based on HTTP. Because adopting multicast protocol, the scalability of UPnP is very limited. It is only used for service discovery in small area.

Service Location Protocol is developed by Internet Engineering Task Force (IETF). Unlike UPnP's XML, SLP uses URL [19] to define the service type and address for a particular service.

There are some service discovery protocols that combine the distributed as well as centralized architectures, such as [20] and [21].Service providers within the scope of one or more directories register their services to the directories. At the same time, they are also ready to reply for the broadcasting service request. The clients at first send service request to the directories. And if there is no match in the service reply messages, they broadcast their request to the network.

Most of these works are designed for special purposes and specific networks, therefore, none of such solution can be deployed for a set of heterogeneous networks like CUNs.

6. Conclusion and Future Works

The objective of this work is to develop service discovery architecture in CUN. We list out the requirements for service discovery in CUN and bring forth our novel hierarchical service discovery architecture for CUN and give detailed description of how our architecture meets the requirements of service discovery in CUN. We also present initial performance evaluation results for our solution by using simulator OMNET++.

Although our architecture meets the requirement, it is only the beginning. The future work includes: a) Consider more context-awareness issue in service discovery, b) simulate the multi-overlay architecture to see the scalability of the system, and c) implement the service discovery architecture in real life to see how service discovery in CUN is worked.

References

[1] Service Advertisement and Discovery

[2] Hansmann, Uwe (2003). Pervasive Computing: The Mobile World. Springer. ISBN 3540002189.

[3] http://en.wikipedia.org/wiki/Wifi.

[4] http://en.wikipedia.org/wiki/MANET.

[5] http://en.wikipedia.org/wiki/3G

[6] http://www.bluetooth.com/Pages/Bluetooth-Home.aspx

[7] http://en.wikipedia.org/wiki/WSN

[8] K.Arnold et al., The Jini Specification, Addison-Wesley Longman, Reading, Mass., 1999.

[9] Universal Plug and Play specification v1.0; http://www.upnp.org/.

[10] Apple, Rendezvous, 2004.

[11] Salutation Architecture Specification; http://www.salutation.org/.

[12] Internet Engineering Task Force http://www.ietf.org/

[13] E.Guttman, "Service Location Protocol: Automatic Discovery of IP Network Services," IEEE Internet Computing, vol.3, no.4, July/Aug. 1999, pp. 71-80.

[14] Specification of the Bluetooth System; http://www.bluetooth.com/.

[15] Extensible Markup Language; http:// www.w3.org/TR/WD-xml/.

[16] Yaron Y. Goland , Ting Cai , Paul Leach , Ye Gu , Shivaun Albright .Simple Service Discovery Protocol. Internet Draft-a work in progress, draft-cai-ssdp-v1-03.txt

[17] Simple Object Access Protocol (SOAP) 1.1 W3C Note; http://w3.org/TR/SOAP

[18] J. Cohen, S. Aggarwal, Y. Y. Goland. General Event Notification Architecture Base: Client to Arbiter. Internet Draft - a work in progress, draft-cohen-gena-client-00.txt.

[19] RFC1738---Uniform Resource Locator.

[20] Christian Frank. A Hybrid Service Discovery Approach for Mobile Ad-Hoc Networks. Diplomarbeit, Technische Universit¨at Berlin, September 2003.

[21] Hybrid 2

[22] Daniel Salber, Anind K. Dey, and Gregory D. Abowd, The context toolkit:aiding the development of context-enabled applications, Proceedings of the SIGCHI con- ference on Human factors in computing systems, ACM Press,1999, pp. 434–441.

[23] http://www.omnetpp.org/.

DTN Routing with Localized Distribution of Mobility Plans of Nodes

Yo Chigira and Hiroaki Higaki
Department of Robotics and Mechatronics
Tokyo Denki University, Tokyo, Japan

Abstract - *In an environment with sparse distribution of mobile wireless nodes, conventional wireless multihop adhoc routing protocols are inefficient due to less available neighbor nodes for detection of next-hop nodes. Thus, DTN (Delay-Tolerant Network) routing is required, which supports combination of wireless multihop transmissions and a store-carry-forward method. For avoidance of communication overhead caused by copies of data messages, a unique next-hop node is selected based on locations, velocities, mobility plans and so on of neighbor nodes in distributed methods, and based on mobility plans of all nodes in global methods. However, in the former, due to lack of information about the future topology of the network, it is difficult for intermediate nodes to select their next-hop nodes with high reachability of data messages to the destination node. On the other hand, in the latter, high communication overhead for distribution of mobility plans and high computation overhead to determine next-hop node are required. This paper proposes a localized distribution method of mobility plans where each node distributes all the achieved mobility plans to its neighbor node and a routing method for data message transmissions where each node determines its next-hop node based on the achieved mobility plans. The methods are expected to realize higher reachability of data messages with lower communication and computation overheads.*

Keywords: Wireless Multihop Networks, Ad-Hoc Networks, DTN, Routing, Store-Carry-Forward, Mobility Plan

1 Introduction

Wireless multihop transmissions of data messages are one of the most important techniques for the future wireless computer networks such as mobile ad-hoc networks, sensor networks and mesh networks. Even though a source wireless node and a destination one are not included in their wireless transmission ranges each other, data messages are transmitted from the source wireless node to the destination one with help of intermediate wireless nodes which forward the data messages. Each intermediate wireless node stores data messages received from its previous-hop wireless node and forwards them to its next-hop wireless node. For higher reachability of date massages, i.e. for higher connectivity between wireless nodes, it is assumed that density of mobile wireless nodes are enough high in the conventional routing protocols for wireless multihop transmissions.

On the other hand in wireless multihop networks with sparsely distributed mobile wireless nodes, it is not always possible to detect a wireless multihop transmission route from a source wireless node to a destination one. In addition, it is difficult for the wireless multihop transmission route to be stable for enough long period for transmissions of data messages. Hence, recently, DTN (Delay Tolerant Networks) are being researched. DTN is based on the Store-Carry-Forward method in which data message transmissions from a source wireless node to a destination one is realized by the combination of wireless multihop transmissions and mobility of wireless nodes carrying the data messages. Here, the DTN routing method is expected to require less communication overhead and to achieve high reachability of data messages.

One of the methods is to transmit data messages based on mobility plans of wireless nodes. Some DTN routing protocols are designed based on an assumption that mobility plans of all the wireless nodes are determined and shared among them in advance. Wireless multihop communication among spaceships is one of the applications. However, in wireless multihop networks in which mobility plans are determined autonomously in every wireless node in realtime manner, it is difficult or impossible to share their mobility plans globally among all the wireless nodes.

In this paper, we propose a novel DTN routing method in which each intermediate wireless node determines its next-hop wireless node for each data message based on achieved mobility plans of wireless nodes which are advertised and distributed locally and autonomously in also DTN manner. The distribution of the mobility plans are independent of the requirements of data message transmissions and higher reachability of data messages are expected with lower communication overhead.

2 Related Works

Let $\mathcal{N} := \langle \mathcal{M}, \mathcal{L} \rangle$ be a mobile wireless multihop network where $\mathcal{M} := \{M_i\}$ is a set of mobile wireless nodes and $\mathcal{L} := \{\langle M_i M_j \rangle\}$ is a set of bi-directional wireless communication link $\langle M_i M_j \rangle$ between mobile wireless nodes M_i and M_j. $\langle M_i M_j \rangle$ is available only while the distance $|M_i M_j|$ between M_i and M_j is less than the wireless signal transmission range R. Since $\mathcal{L}$ depends on relative locations of the mobile wireless nodes in M, $\mathcal{L}$ is variable even though $\mathcal{M}$ is stable. In wireless multihop networks such as ad-hoc networks, sensor networks and mesh networks, in case that a destination node M^d is out of the wireless signal transmission range of a source node M^s and vice versa, data messages are transmitted along a wireless multihop transmission route $\mathcal{R} := ||M^s M_1 \dots M_{n-1} M^d\rangle\rangle$, a sequence of intermediate nodes M_i $(i = 1 \dots n-1)$ which forwards data messages.Various ad-hoc routing protocol for detection of $\mathcal{R}$ have been proposed [6]. Here, it is assumed that the density of wireless nodes is enough high to detect wireless multihop transmission route $\mathcal{R}$ with high probability and $\mathcal{R}$ is stable while a sequence of data messages are transmitted from M^s to M^d or even if a wireless link $\langle M_i M_{i+1}\rangle \in \mathcal{L}$ becomes unavailable, another wireless multihop transmission route $\mathcal{R}'$ is configured by re-detection or localized restoration.

However, it is difficult or impossible to detect wireless multihop transmission routes in a wireless multihop network with low density of wireless nodes[1]where each wireless nodes has only a few neighbor wireless nodes or has not always its neighbor wireless nodes. In order to solve this problem, a novel wireless multihop transmission technique, DTN (Delay-Tolerant Network) routing based on the Store-Carry-Forward method, has been proposed[5]. If an intermediate mobile wireless node M_i which receives a data message from its previous-hop wireless node M_{i-1} cannot detect its next-hop wireless node M_{i+1} in its wireless signal transmission range, it carries the data message until it detects M_{i+1}. That is, the combination of wireless multihop transmission and mobility of nodes provides stable data message transmission from a source wireless node M^s to a destination one M^d even in the environment where detection and maintenance of wireless multihop transmission route from M^s to M^d. However, longer transmission delay and higher communication overhead are required generally in DTN routing than in the conventional ad-hoc routing and it is required for them to be reduced.

Until now, various DTN routing methods have been proposed under the following different assumptions:

- Mobility of all or part of the wireless nodes are under control.
- Mobility plans of all the wireless nodes are advertised to all or part of the wireless nodes in advance or in realtime manner.
- All the wireless nodes determines their mobility plans autonomously and in realtime manner.

Message Ferrying[10] is one of the DTN routing method designed under an assumption that mobility of some wireless nodes are under control. Here, some mobile wireless nodes serve a role to store data messages into their storage and to carry them to their destination wireless nodes. The key technique to realize this routing method is to make mobility plans of the mobile wireless nodes carrying data messages depending on the requirements of data message transmissions.

In [3], the mobility plans of all the wireless nodes in the DTN are assumed to be always available globally in advance or in realtime manner. Here, a source wireless node of a sequence of data messages determines whole the wireless multihop transmission routes to a destination wireless nodes, i.e. it determines all the intermediate wireless nodes in the transmission route and the time when each intermediate wireless node forwards the carrying data messages to its next-hop wireless node. It is possible for the source wireless node to determine the plan of the transmission of the data messages since it has mobility plans of all the wireless nodes in the DTN. This routing method is adoptable to the DTNs consisting of spaceships and aircrafts since the mobility plans of all the mobile wireless nodes have been determined in advance and distributed to all the nodes. However, in DTNs in which each wireless nodes determines its mobility plan autonomously and in realtime manner, it requires higher communication overhead to advertise the dynamically determined mobility plans to all the wireless nodes in the DTN.

On the other hand in [2], another variation of the routing method has been proposed. Here, a DTN consists of not only wireless nodes moving autonomously but also some wireless nodes moving in accordance with the predetermined mobility plans which have been advertised to all the wireless nodes in the DTN in advance. Each intermediate wireless node carrying data messages in transmission determines its next-hop wireless node based not only on the information about its neighbor wireless nodes but also on the information about the wireless nodes whose mobility plans have already been advertised. Thus, the routing protocol achieves high reachability of data messages to their destination nodes.

There are two kinds of routing protocols for DTNs in which all the wireless nodes determines their mobility plan autonomously and dynamically as follows:

- Data messages are copied in intermediate wireless nodes for their higher reachability to their destination nodes.
- No data messages are copied and each intermediate wireless node forwards them only to its next-hop wireless node for lower communication overhead.

Epidemic Routing[8] is one of the most well-known routing protocol using copies of data messages. Each time a wireless node M is included in an intermediate wireless node M_i carrying data messages, M_i confirms whether M is carrying copies of the data messages. If it is not carrying them, M_i forwards copies of the data messages to M with the predetermined probability called infection probability. That is, it is a variation of flooding of the data messages which is based on store-carry-forward manner and realizes reduction of communication overhead caused by

[1] [7] shows that more than 8 neighbor wireless nodes are required for each wireless node to detect wireless multihop transmission routes with probability higher than 90%.

the copies of the data messages by the restrict broadcasts of data messages in intermediate wireless nodes. On the other hand in Probabilistic Routing[4], based on the probability of the neighbor wireless nodes to become neighbor wireless nodes of a destination wireless node, an intermediate wireless node carrying data messages forwards to its neighbor wireless nodes only if the evaluated probability is higher than the threshold probability. It is also a variation of flooding of data messages and is designed for DTNs with autonomous mobility of wireless nodes. Until now, such various DTN routing protocols have been proposed. However, most of them are based on certain limitations on mobility of wireless nodes. For DTNs with autonomous mobility of wireless nodes, higher communication overhead is required in some DTN routing protocols due to copies of data messages since it requires many broadcasts of copies of data messages and higher storage overhead to store copies of the data messages. In addition, it is difficult for each wireless nodes to determine when it discards the storing data messages. Hence, this paper proposes a novel DTN routing protocols for support of autonomous and dynamic mobility of wireless nodes with less communication overhead. Here, no copies of data messages are transmitted and each intermediate mobile wireless node carries data messages and forwards them to only one neighbor wireless node based on the achieved mobility plans of multiple wireless nodes. For lower communication overhead and higher reachability of data messages, mobility plans are shared according to localized distribution, i.e. each wireless nodes advertises sharing mobility plans to its neighbor wireless nodes independently of the requirements of transmissions of data messages.

3 Proposal

3.1 Localized Distribution of Mobility Plans

As discussed in the previous section, in case that each wireless node moves based on its mobility plan determined independently of the other nodes, DTN routing protocol by which next-hop wireless node is determined in each intermediate wireless node based on its achieved mobility plans of the other nodes is expected to achieve higher reachability and shorter transmission delay of data messages with less communication overhead. Since the mobility plans are determined autonomously and dynamically, i.e. not in advance but in realtime manner, the DTN routing protocols based on an assumption that all the wireless nodes achieve mobility plans of all the wireless nodes in the DTN in advance or in realtime manner with low communication overhead are not adopted. Thus, a novel DTN routing protocol by which an intermediate mobile wireless node carrying data messages determines to carry them until it forwards them to its neighbor wireless node which it includes in its wireless signal transmission range later based on its achieved mobility plans is required to be developed. In DTN routing protocols proposed in [9] and some others, an intermediate mobile wireless nodes determines whether it continue to carry data messages or to forward them to one of its neighbor wireless nodes based on the mobility plans of its own and its neighbor wireless nodes. However in these routing protocols, it does not consider the cases where the intermediate node carries the data messages and forwards them to a wireless node which becomes to be included in its wireless signal transmission range later. In addition, it also does not consider the cases where one of its neighbor wireless node to which it forwards the data messages moves and other wireless nodes become to be included in its wireless signal transmission range and the data messages are also forwarded since each intermediate wireless node determines its next-hop wireless node only based on the mobility plans of its own and its 1-hop neighbor wireless nodes. For example in Figure 1, a mobile wireless node M_i carries data messages and forward them to its neighbor mobile wireless node M_j at time t_{ij}. Here, mobility plans are exchanged between M_i and M_j and these wireless nodes detect that the data messages are expected to be transmitted to nearer to a destination wireless node M^d by M_j than by M_i. However, if M_i has achieved the mobility plan of M_k which will be included in its wireless signal transmission range at time $t_{ik} > t_{ij}$ and will carry data messages nearer to M^d than M_i and M_j, M_i does not forward the data messages to M_j at t_{ij} but carries and forwards them to M_j at t_{ik}. In addition, if M_i has achieved mobility plan of another mobile wireless node M_l which will become a neighbor wireless node of M_j at time $t_{jl} > t_{ij}$ and it detects that M_l carries the data messages nearer to M^d by M_l than by M_k, M_i forwards the data messages to M_j at t_{ij}[2]

For realizing such DTN routing, this paper proposes a method by which each wireless node holds not only its mobility plan but also mobility plans of the other wireless nodes and exchanges them with its possible neighbor wireless nodes even when it has not any data messages to forward to one of its neighbor wireless nodes. A mobility plan MP^j is a 4-tuple $\langle M_j, t_b^j, t_e^j, l^j(t)\rangle$ where M_j is a node identifier of M_j and $l^j(t)$ is a location of M_j at time t between the mobility beginning time t_b^j and the mobility ending time t_e^j, i.e. $t_b^j < t < t_e^j$. Each wireless node M_i holds mobility plans $\{MP^j\}$ ($M_j \in MS^i$) which M_i achieves by exchange with its neighbor wireless nodes where MS^i is a set of wireless nodes whose mobility plan M_i has achieved and $M_i \in MS^i$. Each time M_i and another wireless node M_k become neighbor nodes one another, M_i sends all or part of the holding mobility plans $\{MP^j\}$ ($M_j \in MS^j$) to M_k and M_i receives all or part of the mobility plans $\{MP^j\}$ ($M_j \in MS^k$) held by M_k

[2] Though M_i expects M_j to forward the data messages to M_l at t_{jl}, M_j does not always forward the data messages to M_l. According to the distribution method of advertised mobility plans of wireless nodes, M_j achieves mobility plans of M_l from M_i with the data messages. Hence, it is possible for M_j to forward the data messages to M_l. However, since M_j also achieves mobility plans of other wireless nodes, it may forward the messages to one of the other wireless nodes if the messages will be transmitted nearer to M^d.

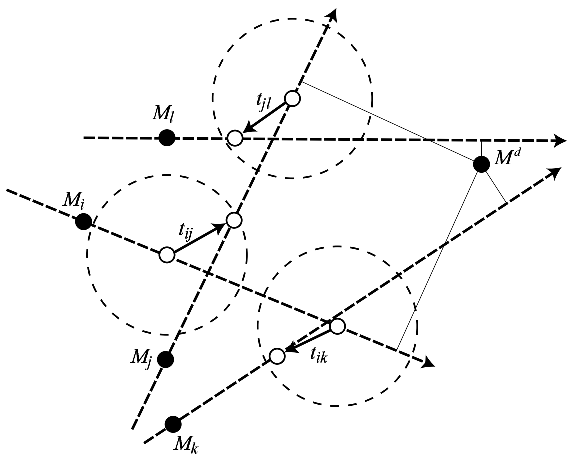


Figure 1. DTN Routing with Mobility Plans Advertised Locally.

from M_k as shown in Figure 2. Here, if M_i holds a mobility plan $\langle M_l, t_b^l, t_e^l, l^l(t)\rangle$ of a mobile wireless node M_l and achieves another mobility plan $\langle M_l', t_b^{l'}, t_e^{l'}, l^l(t)'\rangle$ of M_l from M_k, M_i holds the more recent mobility plan of M_l and discards the other. That is, if $t_b^l < t_b^{l'}$ is satisfied, M_i holds $\langle M_l', t_b^{l'}, t_e^{l'}, l^l(t)'\rangle$ from M_k and otherwise, i.e. $t_b^l \geq t_b^{l'}$ is satisfied, M_i discards $\langle M_l', t_b^{l'}, t_e^{l'}, l^l(t)'\rangle$ from M_k. According to the procedure, each wireless node holds not only mobility plans of its own and its 1-hop neighbor mobile wireless nodes but also mobility plans of wireless nodes which has not yet been included in its wireless signal transmission range, i.e. it also achieves mobility plans of wireless nodes which have not yet been its neighbor wireless nodes.

At time $t > t_e^l$, M_i also discard holding mobility plan $\langle M_l, t_b^l, t_e^l, l^l(t)\rangle$ of M_l since it never provides useful information for future routing of data messages.

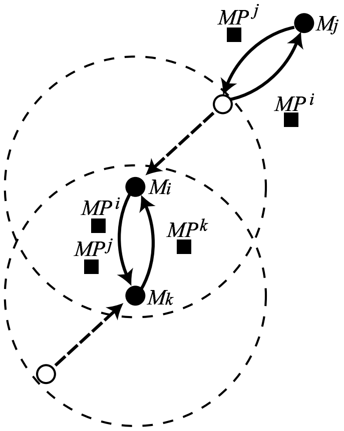


Figure 2. Localized Advertisement of Mobility Plans based on Exchange between Neighbor Nodes.

3.2 DTN Routing

A mobile wireless node M_i carrying data messages in transmission determines to continue to hold the data messages without forwarding to one of the current neighbor wireless nodes or to forward the data messages to one of the current neighbor wireless nodes $M_j \in MS^i$ based on the set $SMP^i := \{MP^j\}$ of mobility plans of its own and of some other wireless nodes. This section proposes a DTN routing protocol in which data messages are transmitted in Store-Carry-Forward manner to a destination wireless node M^d based on the mobility plans achieved by the method proposed in the previous subsection. Each intermediate wireless node determines its next-hop wireless node according to the following 4 steps.

[Step 1] Calculation of Neighboring Duration of Pairs of Nodes

For every pair $\{M_j, M_k\}$ of mobile wireless nodes in MS^i, time duration while M_j and M_k are neighbor wireless nodes, i.e. $|M_j M_k| \leq R$ where R is the wireless signal transmission range, is calculated. Since mobility plans of M_j and M_k held by M_i are $\langle M_j, t_b^j, t_e^j, l^j(t)\rangle$ and $\langle M_k, t_b^k, t_e^k, l^k(t)\rangle$ respectively, if one of the following conditions is satisfied, there are no common t where both mobility plans of M_j and M_k are available and M_i does not detects the time when M_j and M_k exchange data messages with each other.

- $t_e^j < t_b^k$ is satisfied.
- $t_e^k < t_b^j$ is satisfied.

By solving the inequality $|l^j(t) - l^k(t)| \leq R$ on t, the solution is a set of time when M_j and M_k are included in their wireless signal transmission range and are neighbor wireless nodes each other, i.e. it is possible for them to exchange data messages. Hence, if there are no reasonable solutions of the inequality, M_i does not also detect the time when M_j and M_k exchange data messages. The solution is a set of closed intervals $TI_u^{jk} := [t_{b(u)}^{jk}, t_{e(u)}^{jk}]$ $(u = 1, 2, \ldots)$. If data messages are exchanged between M_j and M_k while one of the intervals TI_u^{jk}, there should be a common interval among TI_u^{jk} and intervals $[t_b^j, t_e^j]$ and $[t_b^k, t_e^k]$ while mobility plans of M_j and M_k are available. That is, in a closed interval TI_u^{jk} which satisfies one of the following conditions, M_i does not determine that M_j and M_k can exchange data messages each other:

- $\max(t_b^j, t_b^k) > t_{e(u)}^{jk}$ is satisfied.
- $\min(t_e^j, t_e^k) < t_{b(u)}^{jk}$ is satisfied.

In a closed interval TI_u^{jk} which does not satisfy all the above conditions, it is possible for data messages to be transmitted between M_j and M_k at time t as the following:

$$t \in [\min(t_{e(u)}^{jk}, t_e^j, t_e^k), \max(t_{b(u)}^{jk}, t_b^j, t_b^k)]$$

[Step 2] Calculation of Shortest Time DTN Transmission Route to Each Node

Based on the calculation result in Step 1, M_i achieves the shortest time DTN transmission route from M_i to $M_j \in MS^i$ according to the Dijkstra's SPF algorithm[1].

As shown in Figure 3, let T^{jk} be a time when a data message transmitted from M_i is forwarded from M_j to M_k. In addition, let $TI^{kl}_u := [t^{kl}_{b(u)}, t^{kl}_{e(u)}]$ $(u = 1, 2, \ldots)$ be closed intervals while data messages are transmitted from M_k to M_l. Here, if there are any u where $T^{jk} < t^{kl}_{e(u)}$ is satisfied, it is possible for M_k to forward the data messages forwarded by M_j at T^{jk}. Let u' be the minimum u where $T^{jk} < t\ kl_{e(u)}$ is satisfied. Now, the earliest time T^{kl} when M_l receives data messages from M_k is as follows:

$$T^{kl} := \max(T^{jk}, t^{kl}_{b(u')})$$

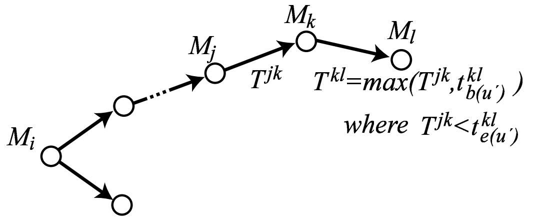

M_k
M_l
M_j
T^{jk}
T^{kl}=max(T^{jk},$t^{kl}_{b(u')}$)
where T^{jk}<$t^{kl}_{e(u')}$
M_i

Figure 3. Calculation of Earliest Arrival Time of Data Message.

According to this formula, by applying the Dijkstra's SPF algorithm, for all wireless nodes $M_j \in MS^i$, the earliest time when M_j receives data messages which M_i is currently carrying and the DTN transmission route to M_j are calculated. If M_i does not detect that it is possible for data messages to be transmitted between M_j and M_k, T^{jk} is evaluated as infinite, i.e. $T^{jk} := \infty$.

Here, a tree S whose root is M_i is configured. A path on S from M_i to each wireless node M_j is the shortest time DTN transmission route for data messages. Let $D[M_j]$ be the temporary earliest time when data messages from M_i reach M_j and initially $D[M_j] := \infty$. In addition, let $P[M_j]$ be the temporary previous-hop wireless node of M_i in the shortest time DTN transmission route and initially $P[M_j] := \emptyset$. When the algorithm terminates, $D[M_j]$ is the determined earliest time when the data messages reach M_j and $P[M_j]$ is the determined previous-hop wireless node along the shortest time DTN transmission route.

1. Let S be a tree only with a root node M_i.
2. $T := MS^i - \{M_i\}$.
3. $D[M_i] := t$ where t is the current time.
4. For all the wireless nodes M_j in T, $D[M_j] := T^{ij}$ and $P[M_j] := M_i$.
5. While $D[M_j] \neq \infty$ for $\exists M_j \in T$, the following procedures are repeated:
    (a) $M_j \in T$ with the minimum $D[M_j]$ is added to S as a child node of $P[M_j]$.
    (b) $T := T - \{M_j\}$.
    (c) For $\forall M_k \in T$, the earliest time T^{jk} when data messages reach M_k from M_i through M_j is calculated and if $T^{jk} < D[M_k]$ is satisfied, $D[M_k] := T^{jk}$ and $P[M_k] := M_j$.

When the algorithm terminates, M_i detects that it is possible for data messages to reach wireless nodes M_j included in the tree S at $D[M_j]$ and M_i does not determine whether it is possible for data messages to reach wireless nodes in T. In order to transmit data messages from M_i to $M_j \in S$ along the shortest time DTN transmission route, M_i forwards data messages to its child wireless node M_k along the unique path from M_i to M_j in S at time T^{ik}. This does not mean that the data messages are transmitted along the path in S as discussed in the previous section.

[Step 3] Calculation of the Shortest Distance between Nodes Carrying Data Messages and the Destination Node

It is possible for data messages to reach the wireless nodes included in S with the root node M_i configured in Step 2. Each mobile wireless node $M_j \in S$ can receive the data messages at time $D[M_j]$ from its previous-hop wireless node $P[M_j]$ along the detected DTN transmission route. Hence, M_i detects that during a closed interval $[D[M_j], t^j_e]$, M_j can carry the data messages and moves along $l^j(t)$. At time t'_j during the closed interval $[D[M_j], t^j_e]$, i.e. $t'_j \in [D[M_j], t^j_e]$, the distance $|l^j(t'_j)M_d|$ between M_j and a destination wireless node M^d is the shortest. Here, $Dist(M_j) := |l^j(t'_j)M_d|$.

[Step 4] Determination of Next-Hop Wireless Node According to the calculation in Step 3, M_i detects the mobile wireless node M_j where $Dist(M_j)$ $(M_j \in S)$ is the shortest to which data messages are transmitted from M_i by combination of wireless multihop transmission and mobility and which carries the messages to the nearest to the destination wireless node. Thus, the path from M_i to M_j in the configured tree S is the best DTN transmission route along which the data messages reach the nearest to the destination wireless node which M_i can detect based on the mobility plans which M_i has achieved. M_i forwards the data messages to its child wireless node M_k in S at time $D[M_k]$. As discussed before, M_k does not always transmits the received data messages to its child wireless node in S which is a tree configured by M_i. On receipt of the data messages and each time M_k achieves new mobility plans from its neighbor nodes while it carries the data messages, M_k determines its next-hop wireless node according to the procedure, i.e. Steps 1 to 4, and carries and forwards the data messages.

4 Performance Evaluation

This section evaluate some performance of the proposed DTN routing protocol based on the localized distribution of mobility plans. Here, reachability of data messages and expected delay for DTN transmission are evaluated in simulation. The performance of the proposed protocol is compared to that of the conventional protocol in which it is assumed that mobility plans of all the wireless nodes are shared among all the wireless nodes without communication overhead and transmission delay for the distribution.

A simulation field is a 3,000m × 3,000m square and a source wireless node M^s and a destination one M^d are stationary at (1,000m, 1,000m) and (2,000m, 2,000m) respectively as shown in Figure 4. 10–100 mobile wireless nodes with 100m wireless signal transmission range are initially distributed randomly according to the unique distribution randomness. Each mobile node moves according to the Random-Way-Point mobility model where it moves in straight line with randomly determined constant speed 0–10m/s. Thus, the location $(x^i(t), y^i(t))$ of a mobile wireless node M_i at time $t \in [t^i_b, t^i_e]$ moving from (x^i_b, y^i_b) at time t^i_b to (x^i_e, y^i_e) at time t^i_e is as follows:

$$x^i(t) = x^i_b + (x^i_e - x^i_b)(t - t^i_b)/(t^i_e - t^i_b)$$

$$y^i(t) = y^i_b + (y^i_e - y^i_b)(t - t^i_b)/(t^i_e - t^i_b)$$

Each mobile wireless node M_i has randomly determined 0–50s interval time at (x^i_e, y^i_e) and randomly determines its next mobility plan. During the interval time, though M_i does not move, M_i receives, stores and forwards data messages and exchanges mobility plans with its neighbor wireless nodes. In the conventional method, the newly determined mobility plans are assumed to be advertised at the moment without communication overhead. The lifetime of all the data messages is assumed to be 1,000s. If the lifetime is expired during the transmission without reaching a destination wireless node, a data messages is discarded by the carrying intermediate wireless node.

Figure 5 shows reachability of data messages in the proposed routing protocol and in the conventional one in various numbers of mobile wireless nodes. The reachability in the proposed protocol is averagely only 13.7% lower than the conventional one even though each wireless node holds only part of mobility plans in the mobile wireless network. Figure 6 shows average DTN transmission delay of data messages. In cases with various numbers of mobile wireless nodes, almost the same transmission delay is required in the proposed and conventional protocols.

5 Concluding Remarks

This paper proposes a novel DTN routing protocol in which each intermediate wireless node determines its next-hop wireless node based on the achieved mobility plans of the other wireless nodes which are locally advertised among the neighbor wireless nodes. It is expected to achieve higher reachability of data messages with shorter transmission delay and lower communication overhead. In brief simulation experiments, it achieves 13.7% lower reachability with the same transmission delay in comparison with the impractical conventional protocol based on the assumption with the ideal sharing of mobility plans among all the mobile wireless nodes in the network.

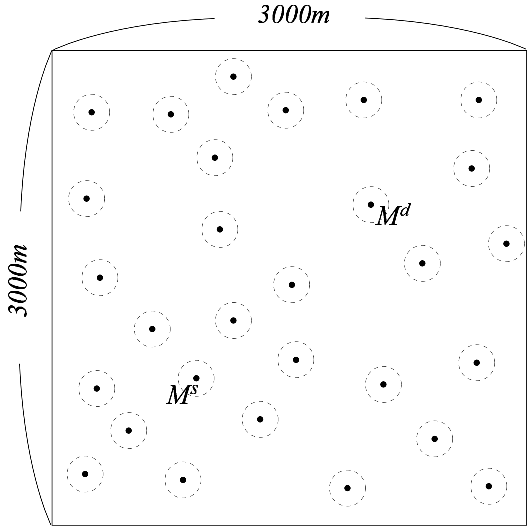


Figure 4. Simulation Field.

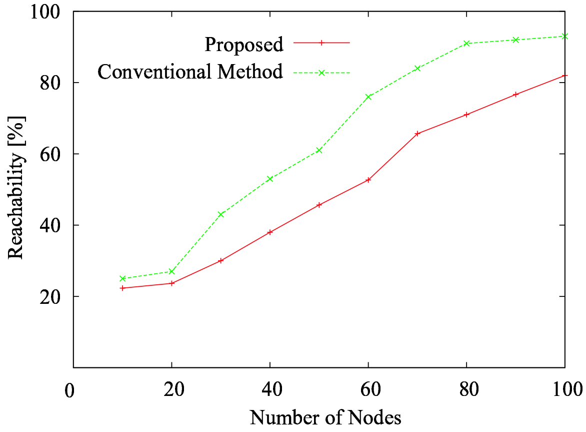


Figure 5. Reachability of Data Messages.

Our future works are as follows:

- Evaluation of communication overhead including that for distribution of mobility plans.
- Performance comparison with the other conventional protocol in which mobility plans are only exchanged between neighbor wireless nodes, i.e. next-hop wireless nodes are determined only based on the mobility plan of 1-hop neighbor nodes.
- Extension of the proposed protocol based on the assumption of variable (unstable) mobility speed of wireless nodes.

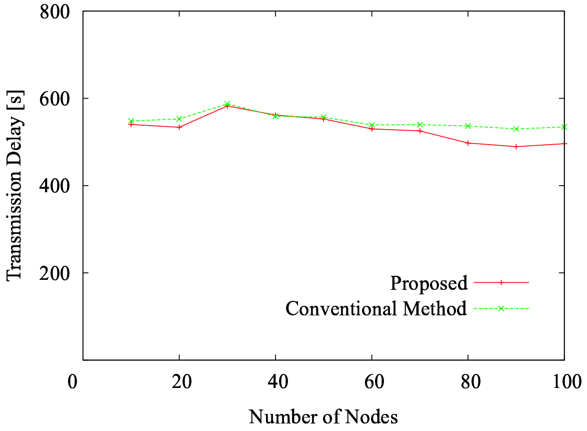


Figure 6. Transmission Delay of Data Messages.

- A method to discard the achieved mobility plans based on the limitation of storages in wireless nodes.

References

[1] Aho, A.V., Hopcroft, J.E. and Ullman, J.D., “The Design and Analysis of Computer Algorithms,” Addison Wesley (1974).

[2] Chen, Z.D., Kung, H.T. and Vlah, D., “Ad Hoc Relay Wireless Networks over Moving Vehicles on Highway,” Proceedings of the 2nd ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 247–250 (2001).

[3] Jain, S., Fall, K. and Patra, R., ”Routing in a Delay Tolerant Network,” Proceedings of the ACM SIGCOMM 2004, pp. 145–158 (2004).

[4] Lindgren, A., Doria, A and Schelen, O., ”Probabilistic Routing in Intermittently Connected Networks,” Lecture Notes in Conputer Science, No. 3126, pp. 239–254 (2004).

[5] Parrell, S. and Cahill, V., “Delay- and Disruption-Tolerant Networking,” Artech House (2006).

[6] Perkins, C.E., “Ad Hoc Networking,” Addison-Wesley (2000).

[7] Seyama, T. and Higaki, H., “G-AODV+PCMTAG: Routing in MANET with Low Overhead Flooding and Route-Shortening,” Proceedings of the International Conference on Parallel and Distributed Computing and Networks, pp. 103–110 (2008).

[8] Vahdat, A. and Becker, D., “Epidemic Routing for Partially-Connected Ad Hoc Networks,” Technical Report CS-200006, Duke University (2000).

[9] Yamanaka, M. and Ishihara, S., “A Location Dependent Information Access Technique by Push/Pull Combination in VANET,” IPSJ SIG Report, Vol. 2008, No. 227, pp. 25–32 (2008).

[10] Zhao, W., Ammar, M. and Zegura, E., “A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks,” Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing, pp. 187–198 (2004).

CBC Based Inter-LMA Route Optimization for the Race Condition Problem

Eunyoung Oh[1], Eunjoo Jeong[1], Sunme Ryu[1], and Byunggi Kim[1]
[1]Department of Computing, Soongsil University, Seoul, Republic of Korea

Abstract - *For the inter-LMA network environment in PMIPv6, location information of communicating Mobile Node(MN) and Correspondent Node(CN) is maintained in the different Local Mobility Anchor(LMA). Accordingly, when both MN and CN move to different MAGs simultaneously, race condition occurs in which updated location information of MN is forwarded to the previous MAG of CN. In this paper, an inter-LMA Route Optimization(RO) scheme is proposed to solve the race condition. It solves the race condition using the Correspondent Binding Cache(CBC) that keep track of location information of MNs managed by neighboring MAGs.*

Keywords: PMIPv6; Inter-LMA; Router Optimization; Race Condition; Correspondent Binding Cache.

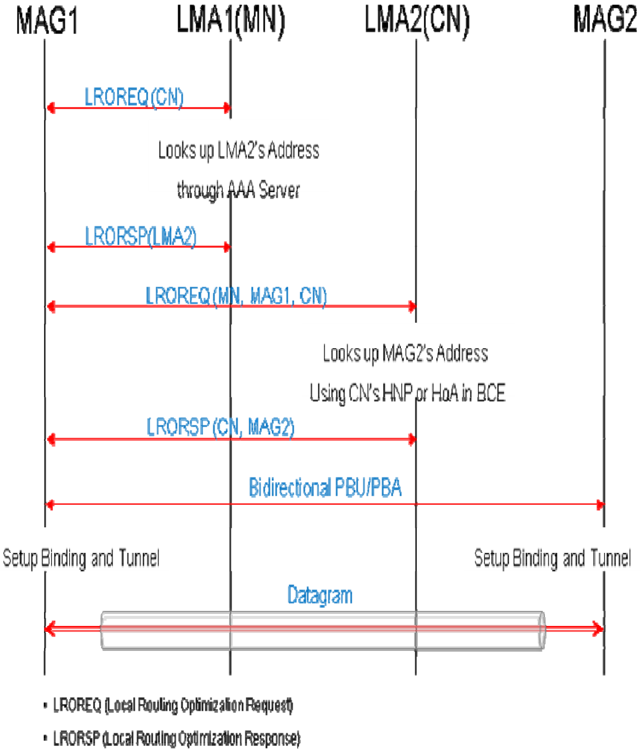
1 Introduction

Recently, with development wireless network technologies, user needs to access Internet for receiving the services anytime, anywhere even while on the move as well as fixed place. To ensure the Mobile Node's mobility, Internet Engineering Task Force(IETF) has been researched. Proxy Mobile IPv6(PMIPv6)[2] is RFC 5213.

PMIPv6[1] is network-based mobility management protocol based on MIPv6[3]. In PMIPv6, it doesn't need to exchange the signaling messages between MN and Home Agent(HA). In other worlds, it has a good advantage that MN doesn't have to do any modifying the protocol stack. Although this, PMIPv6 has a problem, triangular routing problem, that all packet passes by the LMA. To solve this problem, many researches has been studied various RO schemes. However, they mostly couldn't treat the race conditions which might occur in the Inter-LMA local routing. Therefore, we propose the PMIPv6 Inter-LMA RO scheme for solving the race condition in this paper. In chapter 2, we explain the proposed the Inter-LMA in Internet draft[4] , and in chapter 3, we propose the PMIPv6 Inter-LMA RO scheme. We will make a conclusion in chapter 4.

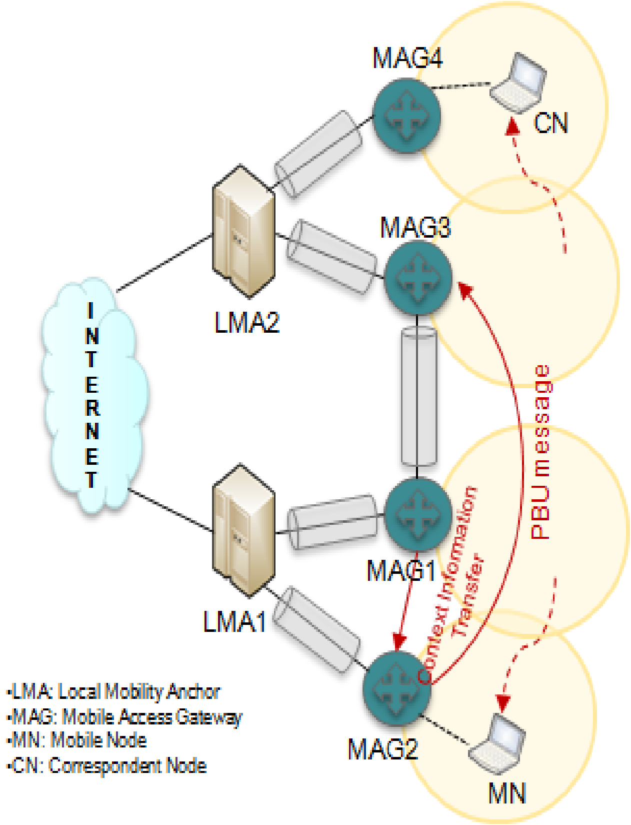
2 PMIPv6

This chapter addresses shortly about the route optimization scheme[4] proposed by the Network based Localized Mobility Management(NetLMM) working group. In [4], the Local Route Optimization Indication(LRI) field values in the Local Routing Optimization Mobility Option is used for the route optimization. Local Optimization Mobility Option is added in the route optimization message.



[Figure 1] LMA in the path of two or more optimization methods

[Figure 1] shows the processes of the Inter-LMA Local Routing in [4]. This RO scheme is more efficient than the others, nevertheless they couldn't handle the race conditions which might occur with the Inter-LMA networks. [Figure 2] represents the race conditions in the Inter-LMA networks. The location information of the CN and MN is managed by each LMA, if they belong to the different LMAs. In this Inter-LMA network environments, when two nodes move into the new MAG area at the same time, a MN's location information is transmitted to a CN's previous MAG. This situation is so called the race condition and it leads to the packet loss and the needless signaling exchanges[5].



[Figure 2] Race conditions in the Inter-LMA Local Routing

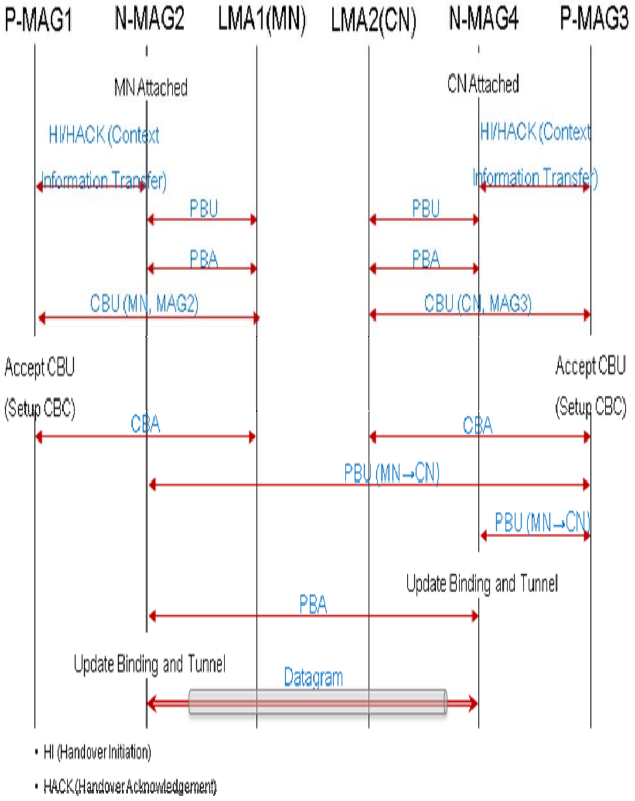
3 Proposed Route Optimization Scheme

We solved the problem, the race condition, by applying the CBC proposed in [6]. A CBC maintains the location information of MNs, which is managed by the neighbor MAG, at the Binding Cache in the MAG. [Table 1] is a data structure of a MAG's CBC Entry. MN is managed by the neighboring MAGs in this table.

Entry	Description
Proxy-CoA of MAG	Proxy Care-of Address of neighboring MAG
MN-Identifier	The Identifier of the MN in the neighboring MAG
MN-Link layer Identifier	Link-layer Identifier of the MN's connected interface on the access link
MN-HNP	Home Network Prefix of the MN's connected interface
Link-local address of the MAG	Link-local address of the MAG on the access link shared with the MN
LMA Address	Address of the LMA serving the attached MN
Tunnel Interface Identifier	Interface Identifier of the tunnel between LMA and MAG of the MN

[Table 1]Correspondent Binding Cache Entry data structure of MAG

[Figure 3] shows our proposal route optimization method. Above the process, the data messages transmission from a MN to a CN is as follows.


P-MAG1
N-MAG2
LMA1(MN)
LMA2(CN)
N-MAG4
P-MAG3
MN Attached
CN Attached
HI/HACK (Context Information Transfer)
HI/HACK (Context Information Transfer)
PBU
PBU
PBA
PBA
CBU (MN, MAG2)
CBU (CN, MAG3)
Accept CBU (Setup CBC)
Accept CBU (Setup CBC)
CBA
CBA
PBU (MN→CN)
PBU (MN→CN)
Update Binding and Tunnel
PBA
Update Binding and Tunnel
Datagram
• HI (Handover Initiation)
• HACK (Handover Acknowledgement)

[Figure 3] The proposed Inter-LMA route optimization scheme

When a MN and a CN simultaneously goes into the new MAG area, the bi-directional tunnel between MN's MAG1 and CN's MAG3 is deleted, and MAG1 transmits the context information including the address of all CN's MAG communicated with a MN to the MAG2. A MAG2 exchanges the Proxy Binding Update (PBU) / Proxy Binding Acknowledgement (PBA) messages with a LMA for updating the MN's location information. A LMA sends the Correspondent Binding Update(CBU) message with a MAG1 for a notifying the MN's location information, then the MAG1 is created CBC. MAG2 is MN to communicate with the PBU message delivery to all CN in order to update the location information. MAG3 is sent the PBU message to the MAG4 using CBC. MAG4 is an update location information of MN and while send a PBU messages, create a tunnel between the MAG2. MAG2 that received PBA message from MAG4 also update the location information of CN and make a tunnel to MAG4, local data path between two MAG is made. Therefore, proposed route optimization technique with additional CBC in MAG can be used to avoid the race condition.

4 Conclusions

The triangle routing problem in PMIPv6 occurs additional delay during the packet transmissions and then causes packet loss or reordering. As a result, the cause packets re-transmissions reduce the quality of the transmission to provide facilitate service to give trouble. To solve this triangle routing problem various RO or local routing scheme has been proposed. But, most of them cannot handle the race condition which might occur with the Inter-LMA network environment. By solving problems we can expect higher performance than proposed different route optimization scheme. In addition, it can provides efficient transport environment in PMIPv6.

5 Acknowledgements

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2011-C1090-1121-0010)

6 References

[1] S. Gundavelli et al., "Proxy Mobile IPv6," RFC 5213, May 2008.

[2] Jae-Min Lee et al., "Performance Analysis of Route Optimization on Proxy Mobile IPv6," ICSNC'08, Malta, pp. 280-285, Oct. 2008.

[3] D. Johnson et al., "Mobility Support in IPv6," RFC 3775, Jun. 2004.

[4] Q. Wu et al., "An Extension to Proxy Mobile IPv6 for Local Routing Optimization," draft-wu-netext-local-ro-05, Feb. 2010.

[5] M. Liebsch et al., "PMIPv6 Localized Routing Problem Statement," draft-ietf-netext-pmip6-lr-ps-02, Jan. 2010.

[6] A. Dutta et al., "ProxyMIP Extension for Inter-MAG Route Optimization," draft-dutta-netlmm-pmipro-01, Jul. 2008.

A Study on the Reduction of Overhead based on Message Lists in DTN

YoungBhin Song[1], Eunjoo Jeong[1], WooYeon Joo[1], and Byunggi Kim[1]
[1]Department of Computing, Soongsil University, Seoul, Republic of Korea

Abstract - *After a message is arrived at the destination, it's copies can remain in the DTN. So, we propose the duplicated message reduction scheme based on the delivered message lists. Each node in the network maintains a list of messages that already arrived at the destination. Whenever two nodes are connected, they update their delivered message lists by exchanging the lists information. By removing duplicated messages in the list, are reduced and the network traffic is reduced too. This eventually reduce network overhead and improve transmission success rate on the DTN network.*

Keywords: DTN, Duplicated message, Delivered message list, Epidemic, Spray and Wait

1 Introduction

Recently, as technology has been advancing rapidly, communicative technology has been developed in various environments. However, there are still some problems such as a communication gap due to lack of infrastructure in many areas, and difficulty in communicating between different networks because of diversity in communicative technology. A new network model is required, not just for an earth but for the communication between satellites or planets. Delay Tolerant Network(DTN) [1] is the network for supporting communication in such environments.

DTN supports communication in environment which has many communication gaps, errors, delay, and bandwidth. In this paper, we researched the ways to reduce network's overhead and improve transmission success rate by exchanging the list of messages reached the destination. In chapter 2, we explain the previous Epidemic Routing Protocols and Spray and Wait Routing Protocols, and in chapter 3, we will discuss the way of reducing overhead and improving transmission success rate by using the list of the messages with ID that reached the destination. We will make a conclusion in chapter 4.

2 Routhing Protocol Of DTN

There are various routing protocols in DTN. Especially, Epidemic [2] and Spray and Wait[3] Routing Protocol are often used for it because of their simple structure.

Epidemic Routing Protocol exchanges messages with different nodes whenever communication is possible. [Fig.1] shows the procedure of how Epidemic Routing delivers the message. (a) Source node S does not have a connection to its destination, node D. But it sends the message to C1 and C2, and in (b) C1 and C2 move to reach C3. C2 sends the message to C3, and C3 delivers it to node D. Thus, the message can reach the right destination. Epidemic Routing uses the Flooding method, which is the way of sending the message whenever possible, so it causes many overheads in the network and exhausts the each node's storage so quickly. In contrast, Spray and Wait Routing protocol keeps messages from being spread more than certain number in whole network, using two modes of Spray phase and Wait phase. Spray phase is similar to Epidemic Routing Protocol and Wait phase is similar to the Direct Transmission [4].

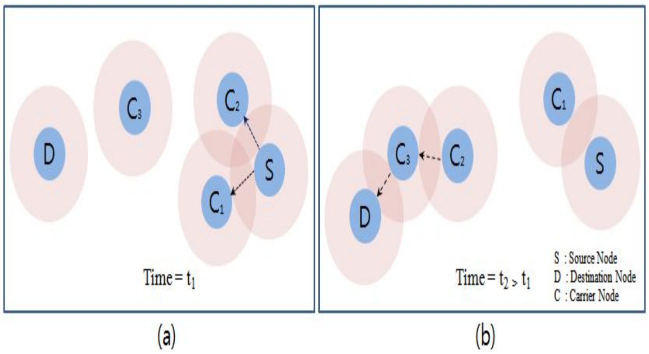


Fig. 1. Epidemic Routing

[Fig.2] shows the case of Spray phase and Wait phase, respectively. In (a) Spray phase, when a Source node happens to meet the other nodes, it sends the fixed number of messages and turns into the Wait Phase. On the other hand, in (b) Wait phase, a Source node must send messages only when it meet the destination node directly.

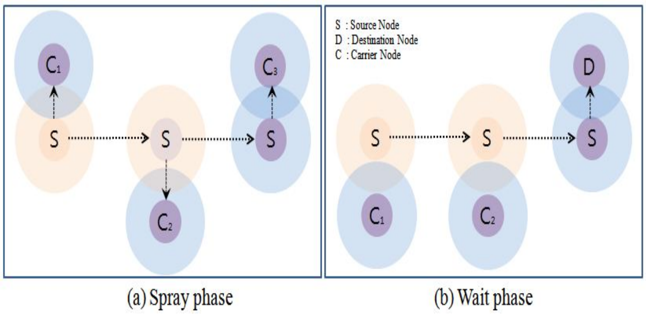


Fig. 2. Spray and Wait Routing

3 Deleting Duplicated Messages Using The Message List

In this paper, we propose the method of deleting the messages already delivered from the network. Previous methods mainly focused on how to deliver the message using the Routing between nodes. But the messages that already reached the destination in network, unless its TTL expires or every message is deleted, make it difficult to deliver other messages and waste the storage space of the node.

The suggested method is characterized as follows :

1) Every node has the list of the messages reached the destination.
2) When a node receives a message headed for itself, it updates the message ID and TTL value at the list.
3) The node exchanges and updates the list every time, if it meets another node.
4) The node deletes the message from a storage space, if it already has the messages on the list.
5) When a TTL value of the message expires, it deletes the message from the list.
6) The list's size is not limited, because it does not have a great effect on the storage space of node.

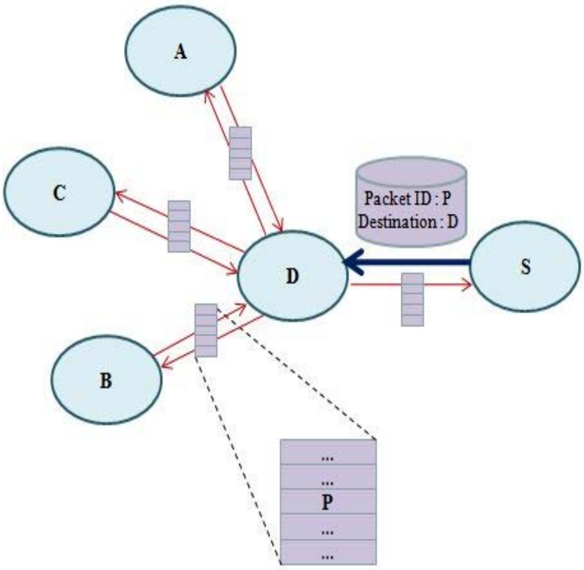


Fig. 3. Exchanging and creating the item of the message list arrived at the destination

With these six assumptions, when a message arrived at the destination node D, Node D updates its' list with message's ID and TTL value as shown in [Fig.3]. After then, whenever it meets by chance the other nodes, it exchanged the messages with Node A, B, and C for removing the expired messages. That is, the mechanism will be widespread in the network, helping the message be deleted, and when TTL expires the nodes will delete the message automatically, causing the message to disappear from the list.

If two nodes are connected as shown in (a) of [Fig.4], they exchange the each owned messages (message delivered to the destination). The node which updated its list checks the storage space if it has the ID of that message and if it has, deletes the message (b). They continue exchanging message with message exchanging procedure (c). When they finish exchange, the process ends as like (d).

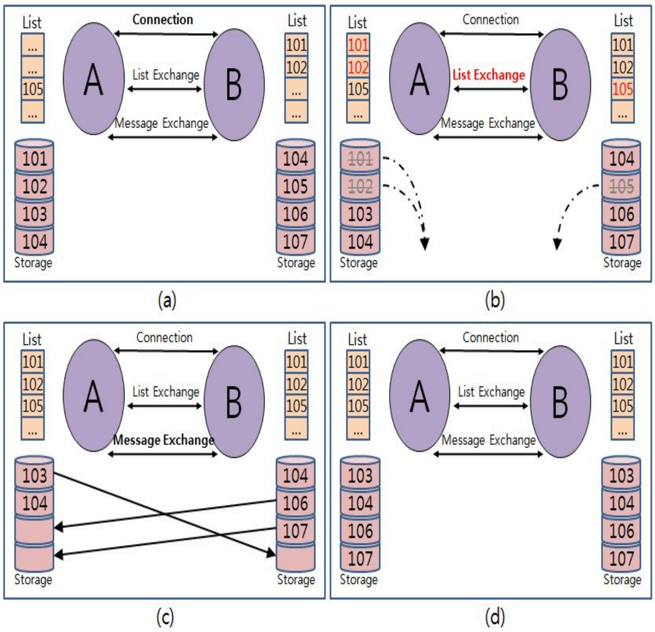


Fig. 4. List exchange process between two nodes

4 Conclusions

The suggested method cannot have a positive effect in existing network because of increased network traffic when exchanging the lists. But in DTN, it could be very effective due to the response message or the method of deleting messages with TTL does not have great effects. Also, it is necessary to delete the duplicated message because a single message take a lot of space compared to previous communication.

In this study, we can reduce overhead of network using the method suggested on the previous chapter in the network with special environment like DTN, reducing the case that the delivered massages keep other messages from being delivered. Through this method, we can improve the transmission success rate of the messages and expect to improve the performance by applying to various Routing Protocol researched in previous studies.

5 Acknowledgment

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2011-C1090-1121-0010)

6 Reference

[1] Forrest Warthman, “Delay-Tolerant Networks (DTNs) : A Tutorial v1.1,” http://www.dtnrg.org/docs/tutorials/warthman-1.1.pdf, March 2003.

[2] Amin Vahdat and David Becker, "Epidemic Routing for Partially-Connected Ad Hoc Networks," Technical Report CS-200006, Department of Computer Science Duke University, April 2000.

[3] Thrasyvoulos Spyropoulos, et al., "Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile networks," In proceedings of the ACM Annual Conference of the Special Interest Group on Data Communication,pp.252-259,Philadelphia, Pennsylvania, USA, August 2005.

[4] Thrasyvoulos Spyropoulos, et al., "Single-Copy Routing in Intermittently Connected Mobile Networks," Inproceedings of 2004 First Annual IEEE Communications society Conference on Sensor and Ad Hoc Communications and Networks, pp. 235-244, Santa Clara, Canada, October 2004.

On Packet Tracer Implementation of Virtual Network: A Demonstration of Packet Travel and Collision Occurrences in Networks

Ekabua, Obeten O[1]., Isong, Bassey E[1]. and Mbodila, Munienge[2]

1. Department of Computer Science and Information Systems, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa.
2. Science Foundation - Computer Science Unit, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa.
(obeten.ekabua@univen.ac.za, bassey.isong@univen.ac.za, munienge.mbodila@univen.ac.za)

Abstract: *Packet Tracer uses two representation schemes to represent a network: logical workspace and physical workspace. In this paper, our implementation uses these two schemes to build networks and to demonstrate our experiments. Logical workspace is used to build a logical topology and physical workspace is used to arrange devices physically. Packet Tracer as demonstrated in this paper has an operating mode that reflects the network time scheme and demonstrates the behavior of packet travels over the network and how collision occurs in networks. The work presented in this paper shows the need to use packet tracer as an e-learning tool to demonstrate the behavior of network as it concern packet travels and collision occurrences.*

Keywords*: Network, Packet tracer, Collision, Host, Topology, Subnet.*

1. Introduction

One of the strongest engines with immense mixtures of application driving the evolution witness in our modern society is the internet network. People's life, techniques for capturing data and the way of information transmission is drastically being affected on daily basis through internet usage. Internet has become a huge market for communication, education, business and entertainment. With a growing internet network, the size of data resources on the Internet is increasing exploding and it is important for every organisation to retain well-trained staff to support its IT infrastructure. Without the correct skills, time and money can be wasted in trying to solve seemingly complex or difficult problems.

Packet Tracer supports internet users in simulations, visualizations, and animations of networking phenomena. The instrument was developed by Cisco to assist in addressing the "digital divide" in networking learning. In this paper, Packet Tracer is the tool used to demonstrations or clarify the concept of Internet and to define different networking media, cabling testing, Ethernet technologies, how layers operate, Ethernet switching, some protocols used in the network, TCP/IP transport and how the software works. By that we mean some of the features that are used in Packet Tracer, such as the workplace basics (logical and physical) and the operating modes (real-time and simulation time) functioned in Packet Tracer. With the aid of a series of demonstrations, we will show how the network looks in real-time mode and simulation time mode.

These are the keys that one should know in order to understand what happens when packets travel over the network. Different network devices (hubs, switches, routers) will be used to demonstrate the experiments with different technologies such as star topology, ring topology and a mix of star and ring called hybrid topology.

2. Background Information

For clarity, it is important for us to lay some foundation concerning the realities in networks -how do networks work, what is necessary for a network to operate. Therefore we need a network media, and various kinds of cabling that are used in the network for hosts to communicate. Also, we would define the behavior of each device and the layer in which the devices operate.

Packet Tracer uses two representation schemes to represent a network: logical workspace and physical workspace. These two will be used to build networks and to demonstrate our experiments. Logical workspace is used to build a logical topology and physical workspace is used to arrange devices physically. Packet Tracer operating modes reflect the network time scheme. For us to be able to demonstrate the behaviour of the network in reality we will need these operating modes (Real-time Mode and Simulation Mode). In real-time mode, the network responds to actions immediately as they would in a real device. In Simulation Mode, you have direct control over time related to the flow of PDUs. You can see the network run step by step, or event by event, however quickly or slowly you like. Here we will set up scenarios, such as sending a ping packet from one device to another. This will help us to demonstrate most of our experiments in this piece of software Packet Tracer.

We'll also use different topologies to represents our little networks that will be built using Packet Tracer4.0, such as star topology, where one device is connected to the hosts through a central point. Ring and hybrid topologies will also be used with a mix of ring and star topologies. This will help us to come up with some experiments to explain network behaviour in real-time mode and simulation mode using Packet Tracer 4.0. Consequently, the report on this paper will consist of a number of demonstrations with each of them containing some theoretical explanation.

3. System Model

In this section, we present the underlying system model, focusing particularly on issues related to virtual networks and its infrastructure. We assumed that communication is synchronous, and that at each round of communication a node may choose to broadcast or not. But eventually, during intervals in which the network is well-behaved, nearby nodes can communicate, as long as there is no contention. During intervals when the network is badly behaved, however, all messages may be lost. Also, we assumed that nodes have limited collision detection capability (without which collision detection is impossible), and also that collision detection are complete (i.e. there are no false negatives; when a collision occurs, the detector reports a collision) and eventually accurate (i.e. there are eventually no false positives; eventually, the collision detector reports a collision only when a message has been lost - which is usually due to high channel contention).

Our virtual infrastructure are containing a set of deterministic virtual nodes which are distributed throughout the network and residing at a fixed location. There is an interaction between the virtual nodes and the clients behaving like any other mobile device. We would like to emphasize that the system containing virtual nodes appears from a client's perspective, more or less like a system in which each virtual nodes is replaced with a reliable, real mobile device.

As part of our model, we'll like to refer to a network configuration as a network topology and our implementation would be using packet tracer 4.0. Packet Tracer 4.0 uses two representation schemes for a network: the Logical Workspace and the Physical workspace. The logical workspace allows you to build a logical network topology, without regard to its physical scale and arrangement. While the Physical Workspace allows you to arrange devices physically in Workspace (the area in which we use to create our network in Packet Tracer 4.0, watch simulations, and view many kinds of information and statistic) cities, building and wiring closet. Network topology defines the structure of the network. One part of the topology definition is the physical topology, which is the layout of the wire or media. The other part is the logical topology, which defines how the hosts access the media to send data. In designing a network in reality, the network designer has three major goals when establishing the topology of a network:

To provide the maximum reliability, route network traffic through the least cost path within the network and to give the end user the best possible response time and throughput [1]

In order to provide the flexibility required to support differently sized networks, IP addresses come in three classes, A, B, and C. Every class fixes the boundary between the network portion and the host portion of the IP address at a different point. This makes them appropriate for different size networks. [2]

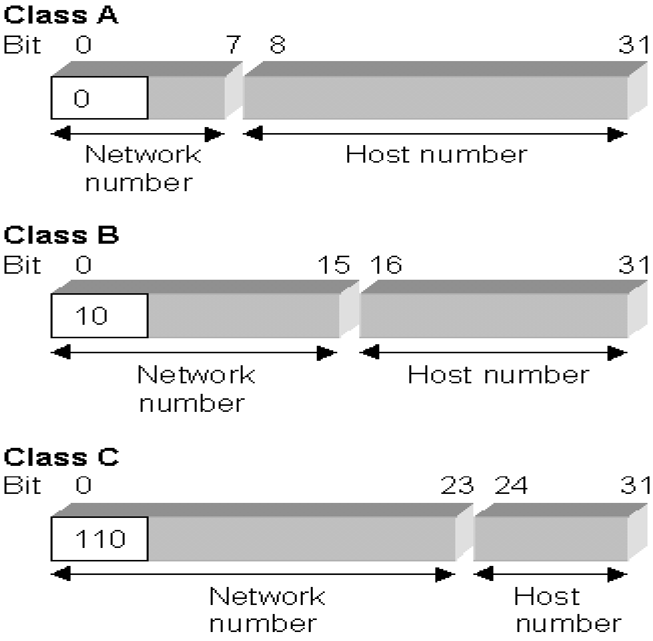


Figure 1: Classes of IP addresses

Class C addresses allow 254 hosts per network and are typically used by smaller and middle-sized companies. Class B networks allow a maximum of 16,384 (65534) hosts, while Class A networks allow more than 16 million hosts. As a consequence, Class A networks are only used by really large organisations. In this project we will be using only class B addresses and Class C addresses.

Calculating the number of possible hosts requires a closer look at the IP classes in their binary form. (The binary system is a base-2 number system, just like the base-10 number system is known as the decimal number system). It is done as follows:

- In a Class C network only the last octet is used to designate the hosts. The maximum decimal number that you can write using eight bits is 256 (28). The host calculation now requires that 2 is subtracted, because two host addresses must be reserved for a network address and a broadcast address. The maximum number of hosts on a Class C network is 256-2=254.
- A class B network allows a maximum of 16,384 hosts (216-2) per network (two octets designate the hosts).
- A class A network allows up to 16,777,214 (224-2) hosts per network (three octets are used to designate the hosts).

The table below shows the range of dotted-decimal values that can be assigned to each of the three address classes. An x represents the host number field of the address which is assigned by the network administrator.

Address class notation	IP address range in dotted-decimal
A (/8 prefixes)	1xxx.xxx.xxx through 126.xxx.xxx.xxx
B (/16 prefixes)	128.0.xxx.xxx through 191.255.xxx.xxx
C (/24 prefixes)	192.0.0.xxx. through 223.255.255.xxx

Table 1: IP address range in dotted-decimal notation

It is very important to understand the notation in the above table, because, when you are using Packet Tracer 4.0, some time when you point your mouse on a device it will give you this notation. Class A networks are also referred to as '/8's (pronounced slash eight's or just eight's) since they have an 8-bit network prefix (one octet is used to designate the network). Following the same convention, Class B networks are called '/16s' and Class C networks '/24s'. [2]

3.1. Subnet

Formula
2N, where N is equal to number of bits borrowed, this is to get the number of total subnets created. 2N-2 Number of valid subnets created 2 H, Where H is equal to number of host bits; this is to get the number of total hosts per subnet. The last formula is 2H-2 to get the number of valid hosts per subnet. The classification bellow shows you how to get your N and H from different IP addresses classes.

Class A Address	N	H	H	H
Class B Address	N	N	H	H
Class C Address	N	N	N	H

N = Network bits
H = Host bits

All 0s in host portion = Network or subnetwork address
All 1s in host portion = Broadcast address
Combination of 1s and 0s in host portion = Valid host address.

3.2. List of Subnets

Class B Host/Subnet Table

Class B	Subnet	Effective	Effective	Number of Subnet
Bits	Mask	Subnets	Hosts	Mask Bits
1	255.255.128.0	2	32766	/17
2	255.255.192.0	4	16382	/18
3	255.255.224.0	8	8190	/19
4	255.255.240.0	16	4094	/20
5	255.255.248.0	32	2046	/21
6	255.255.252.0	64	1022	/22
7	255.255.254.0	128	510	/23
8	255.255.255.0	256	254	/24
9	255.255.255.128	512	126	/25
10	255.255.255.192	1024	62	/26
11	255.255.255.224	2048	30	/27
12	255.255.255.240	4096	14	/28
13	255.255.255.248	8192	6	/29
14	255.255.255.252	16384	2	/30
15	255.255.255.254	32768	2*	/31

Table 2: Class B Host / subnet Table

Class C Host/Subnet Table

Class C	Subnet	Effective	Effective	Number of Subnet
Bits	Mask	Subnets	Hosts	Mask Bits
1	255.255.255.128	2	126	/25
2	255.255.255.192	4	62	/26
3	255.255.255.224	8	30	/27
4	255.255.255.240	16	14	/28
5	255.255.255.248	32	6	/29
6	255.255.255.252	64	2	/30
7	255.255.255.254	128	2*	/31

Table 3: Class C Host / subnet Table

3.3. Subnet Mask

3.3.1. How to Calculate Subnet Mask

The default subnet mask for a Class B network is as follows:

Decimal Binary
255.255.0.0
11111111.11111111.00000000.00000000
1 = Network or subnetwork bit
0 = Host bit

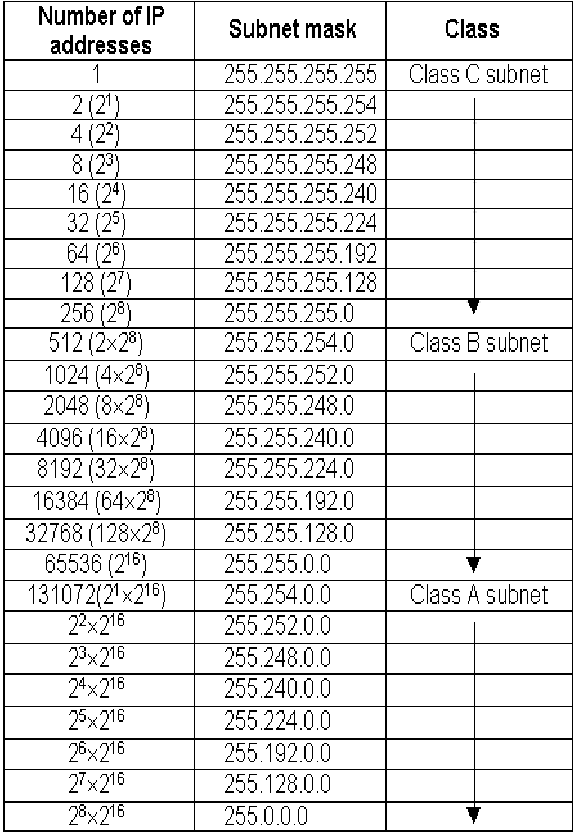
You borrowed 4 bits; therefore, the new subnet mask is the following:
11111111.11111111.11110000.00000000 255.255.240.0
The default subnet mask for a Class C network is as follows:

Decimal Binary
255.255.255.0
11111111.11111111.11111111.00000000
1 = Network or subnetwork bit
0 = Host bit
You borrowed 4 bits; therefore, the new subnet mask is the following:
11111111.11111111.11111111.11110000
255.255.255.240

Note: Subnet a Class B or a Class A network with exactly the same steps as for a Class C network; the only difference is that you start with more H bits. Since we won't be using Class A there is no need for us to show how you get your subnet or subnet mask but the procedures still the same. [1]

3.3.2. List of Subnet Masks

The list below can be used as a fast track when subnetting. It describes the relationship between the number of host IP addresses required and the corresponding subnet mask.



Number of IP addresses	Subnet mask	Class
1	255.255.255.255	Class C subnet
2 (2^1)	255.255.255.254	
4 (2^2)	255.255.255.252	
8 (2^3)	255.255.255.248	
16 (2^4)	255.255.255.240	
32 (2^5)	255.255.255.224	
64 (2^6)	255.255.255.192	
128 (2^7)	255.255.255.128	
256 (2^8)	255.255.255.0	↓
512 (2×2^8)	255.255.254.0	Class B subnet
1024 (4×2^8)	255.255.252.0	
2048 (8×2^8)	255.255.248.0	
4096 (16×2^8)	255.255.240.0	
8192 (32×2^8)	255.255.224.0	
16384 (64×2^8)	255.255.192.0	
32768 (128×2^8)	255.255.128.0	
65536 (2^{16})	255.255.0.0	↓
131072($2^1\times2^{16}$)	255.254.0.0	Class A subnet
$2^2\times2^{16}$	255.252.0.0	
$2^3\times2^{16}$	255.248.0.0	
$2^4\times2^{16}$	255.240.0.0	
$2^5\times2^{16}$	255.224.0.0	
$2^6\times2^{16}$	255.192.0.0	
$2^7\times2^{16}$	255.128.0.0	
$2^8\times2^{16}$	255.0.0.0	↓

Table 4: List of subnet Mark

4. Conclusion

In networking, packet tracer is an enabling tool that supports internet users in simulations, visualizations and animations of phenomena. It helps to show what happens when packets travel over the network and when collisions occur. Therefore, in this paper, we have demonstrated how packet tracer implementation is useful in determining packet travel and collision occurrences in Networks. A system model was instantiated to clarify the implementation and an assumption that communication is synchronous, that at each round of communication a node may choose to broadcast or not. But as long as there is no contention between nodes and the network is well-behaved, during such intervals the nearby nodes can communicate. This also implies that when the network is badly behaved with increased collision, messages may be lost. In submission, network behavior can be determined using packet tracer and our model implementation demonstrates this behavior.

References

[1] http://www.cisco.com/warp/public/473/lan-switch-cisco.pdf. Updated Aug. 01, 2007.

[2] Todd, Lammle, Cisco Certified Networking, Sybex Publishing, 3rd Edition Jones, ISBN: 0782141676, 2004.

A Novel VANET-Based Approach to Determine the Position of the Last Vehicle Waiting at a Traffic Light

Eric Gamess[1] and Imad Mahgoub[2]

[1] Escuela de Computación, Universidad Central de Venezuela, Los Chaguaramos, Caracas 1040, Venezuela
[2] College of Engineering and Computer Science, Florida Atlantic University, Boca Raton, Florida 33431, USA

Abstract—*Vehicular Ad hoc Networks (VANETs) are recognized as an important component of the Intelligent Transportation Systems. Thanks to this emerging technology, new information will be available to design better adaptive traffic light control systems which will dramatically improve the traffic flow. In this paper, we present a novel VANET-based approach to obtain (1) the position of the last vehicle and (2) the number of vehicles, in a line of vehicles stopped at a traffic light. We also show that our algorithm to estimate the position of the last vehicle will still function even if just a few vehicles are equipped with VANET technology (low penetration rate), which is important since it will take years for all vehicles to be equipped with VANET technology*.

Keywords: DSRC, Adaptive Traffic Light Control Systems, VANET, V2V (Vehicle-to-Vehicle Communication), V2I (Vehicle-to-Infrastructure Communication), Vehicular Networks, Intelligent Transportation Systems.

1. Introduction

Vehicular Ad hoc Network (VANET) is a promising, emerging, new technology that integrates the capabilities of wireless networks to vehicles. VANET shares some common characteristics with the general Mobile Ad hoc Network (MANET), such as the movement and self-organization of the nodes. Indeed, some researches consider VANET as a kind of MANET. However, they also have some significant differences that tend to separate them and motivate the development of new specific algorithms, especially for routing protocols [13]. MANETs can contain many nodes that are power constrained and are subject to uncontrolled moving patterns [10]. On the other hand, most of the VANET nodes do not have battery issue since their indirect source of power is the gas in the fuel tank of the vehicle, and their movements are constrained by the road and traffic pattern. In most case, the mobility of vehicles in VANET is assumed to be the car following model [21], in which cars follow one after the other in a streamlined fashion.

There are two types of special electronic devices in VANET: On Board Units (OBUs) and Road Side Units (RSUs). OBUs are placed inside each vehicle and therefore are mobile. RSUs are fixed and installed near the road. Vehicle-to-Vehicle (V2V) communications take place between OBUs, while Vehicle-to-Infrastructure (V2I) communications involve OBUs and RSUs. A VANET-enabled vehicle should be able to receive and relay messages to other VANET-enabled vehicles in its neighborhood (also known as multi-hop relaying), as shown in Fig. 1 where car A is sending a message to car C through car B.

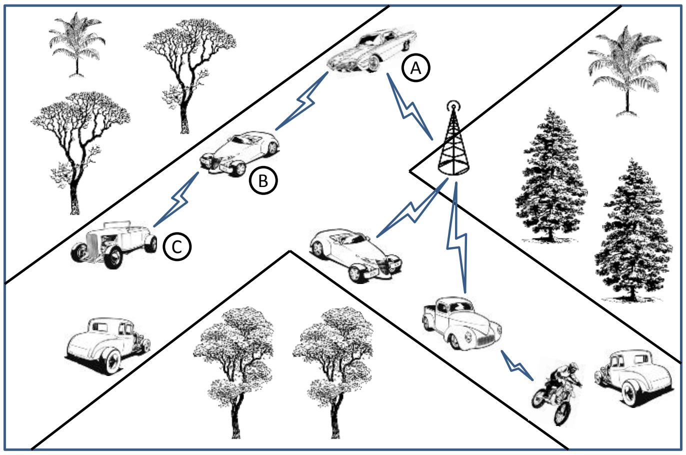


Fig. 1. Example of a Vehicular Ad hoc Network

In the US, the Federal Communication Commission (FCC) allocated a 75 MHz of spectrum in the 5.850-5.925 GHz band for Dedicated Short Range Communication (DSRC) for the Intelligent Transportation System (ITS) [11]. The usage of VANET in the ITS is mainly focused on safety and will be extensively used for collision prevention, forward obstacle detection and avoidance [7], blind crossing, and approaching emergency vehicle warning (Blue Waves). Other important applications are traffic congestion detection, adaptive traffic light control systems, accident reconstruction [6], mapping and navigation guidance, electronic payment (e.g., electronic toll collection, electronic parking payment), weather advisory dissemination, and gas station and hotel information dissemination. Comfort applications will also be present in VANET and aim to improve the comfort of the passengers and include the communication with other vehicles or with the Internet, as well as various social or multimedia entertainment applications. Some of these applications use broadcast communications extensively [18][22][23][26].

Traffic congestion continues to worsen in cities all over the world. In the US alone, traffic congestion creates \$114.8 billion annual drain on the economy in the form of 4.8 billion lost hours and 3.9 billion gallons of wasted fuel [25]. Consequently, a lot of research has been done in the field of adaptive traffic light control systems [29] and traffic congestion detection [4][8][19] to try to improve the flow of vehicles. At the present, most of the proposals for adaptive traffic light control systems rely on information gained from inductive loops [24] or video cameras. An inductive loop is a coil of wire buried in the road’s surface. The two ends of the loop wire are connected to a loop extension cable, which in turn connects to a vehicle detector. The detector powers the loop causing a magnetic field in the loop area. The loop resonates at a constant frequency that the detector monitors. A base frequency is established when there is no vehicle over the loop. When a vehicle passes over the loop, the frequency increases and is sensed by the detector. So inductive loops are generally used in adaptive traffic light control systems to detect the presence of a vehicle and trigger the red light to change to green. Video cameras are another alternative for vehicle detection. They usually require some initial configuration, are processor intensive, have visibility issues, and the information gathered from them is limited to the vision range of the camera.

In this paper, we focus on the use of VANET in adaptive traffic light control systems. We highlight some important information that we consider useful for an adaptive traffic light control system and propose some algorithms to obtain this information.

The rest of this paper is organized as follows. In Section 2, we introduce our novel algorithms to collect useful information from VANET to be used in adaptive traffic light control systems. In Section 3, we discuss some issues and strengths of our approach. Section 4 presents the related works. Finally, Section 5 concludes the paper.

2. Collecting Information from VANET for Adaptive Traffic Light Control Systems

Important information that can be gathered from a line of vehicles stopped at the red light is the number of vehicles in the line and the location of the last vehicle in the line. Priority levels of the vehicles can also be considered. For now, and probably for a few years to come, just a few vehicles will be equipped with a VANET device. So it is important to keep this fact present in our solutions. We will assume that all VANET-enabled vehicles are also capable of determining their current position on the road using a location service like the Global Positioning System [5][9] (GPS). For our algorithms, the location is specified through the latitude and the longitude. However, the same algorithms can be modified to use Cartesian coordinates, by choosing the origin and the direction of the axis.

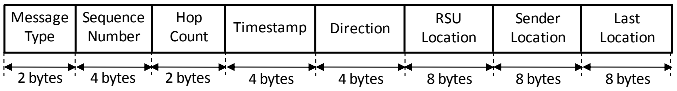
2.1 Obtaining the Position of the Last Vehicle

We will first focus on obtaining the position of the last vehicle in a line of vehicles stopped at a light. Having the position of the RSU and the last vehicle in the line, it is easy to infer the length of the line. We propose to obtain this information by propagating a first message (Last Location Request or LLRequest) from the RSU toward the last vehicle in the line and a second message (Last Location Reply or LLReply) from the last vehicle toward the RSU. These messages are sent as broadcast. The direction of propagation of an LLRequest will be indicated by the RSU using a field called *Direction*. To restrict the rebroadcast of a message and control unnecessary redundant messages which may trigger the familiar broadcast storm problem [17], nodes will rebroadcast the messages according to a “Rebroadcast-Wait-Time” called T and computed by each node as:

$$T = T_{\max}\left(1 - \frac{D}{D_{\max}}\right) \qquad (1)$$

where T_{max} is the “Maximum Rebroadcast-Wait-Time”, D_{max} is the “Maximum Transmission Range” of a node, and D is the distance between the node and the node that sent the LLRequest or the LLReply. Note that T goes to zero for nodes that are far-away (at a distance of D_{max}) from the sender and to T_{max} for nodes that are nearby (at a distance of 0) the sender. Therefore, the further away a node is from the sender, the smaller is his Rebroadcast-Wait-Time (T). This strategy will minimize the number of rebroadcast and give a total propagation time closed to the best one that can be achieved.

The Last Location Request (LLRequest) message and the Last Location Reply (LLReply) message have the same Protocol Data Unit (PDU) and are composed of 8 fields (see Fig. 2).



Message Type	Sequence Number	Hop Count	Timestamp	Direction	RSU Location	Sender Location	Last Location
2 bytes	4 bytes	2 bytes	4 bytes	4 bytes	8 bytes	8 bytes	8 bytes

Fig. 2. LLRequest and LLReply Message

Message Type can be either 0 or 1, and identifies an

LLRequest message and an LLReply message, respectively. *Sequence Number* is used to match a request with a reply and to distinguish between different requests (LLRequests). When the RSU sent an LLRequest message, it must specify a value of *Hop Count* greater than or equal to 1. Each OBU that rebroadcast the message will decrement this field by 1, until 0 where the message is discarded. *Hop Count* is used to control the extent of the propagation of the messages. *Timestamp* is set by the RSU when it sends an LLRequest message and is aimed to control out-of-date messages and replay attacks. By indicating a direction (field *Direction*) in an LLRequest, an RSU can obtain information about different lines of vehicles in the same intersection. In other words, it is not necessary to install several RSUs in an intersection, since one can do all the required work. *RSU Location* (latitude and longitude) are filled by the RSU when it sends an LLRequest message and must contain its latitude and longitude. Each VANET-enabled device that sends or resends a message has to fill *Sender Location* with its own latitude and longitude. This information is used by the recipients to compute their distance (D) from the sender. *Last Location* is the location (latitude and longitude) of the last-known vehicle in the line.

To explain the algorithm, let us take the example of Fig. 3 where 11 vehicles are in a line waiting for the green light. To make the explanation easier, we assume that the radius of the range of a message is equivalent to 3 vehicles. In a real scenario, it will be bigger than 3 vehicles since the range of DSRC is targeted to be up to 1 km. We will assume that each VANET-enabled vehicle periodically broadcast beacon messages that include its position (latitude and longitude), speed, direction, and a timestamp (for dating the spread information). This assumption is fair since the main objective of VANET is safety, so that vehicles have to send regular beaconing messages in order to advertise their presence and prevent collisions. Also, we will presume that it is possible to infer if a vehicle is stopped in the queue of a light from its location, speed, and direction. Before the initiation of the process, the RSU will listen to beaconing messages to discover vehicles waiting in the line, and will find out that within its range (assumed to be 3 vehicles), vehicles A, B, and C are waiting in the line. The RSU initiates the process by sending an LLRequest message with its own position for both *RSU Location* and *Sender Location* fields. It will also fill *Last Location* with the position of vehicle C, since it is the further away vehicle that the RSU knows. It starts a timer for T_{max} seconds, in case it has to resend the LLRequest. Vehicles A, B, and C receive the LLRequest, so they compute their Rebroadcast-Wait-Time (T) to schedule a rebroadcast at T seconds. The Rebroadcast-Wait-Time of vehicle C will be the smallest (closed to 0 seconds) since it is almost out of the range of the RSU (i.e., its distance D from the source of the message is almost D_{max}). Vehicle B will have the second smallest Rebroadcast-Wait-Time, and vehicle A will have the biggest Rebroadcast-Wait-Time (closed to T_{max}) since it is near the RSU. So vehicle C will rebroadcast the LLRequest message almost immediately, with the adequate modifications; that is, decrementing by 1 the *Hop Count* field, replacing the *Sender Location* with its own position, and *Last Location* with the position of vehicle F (discovered by vehicle C thanks to the beacon messages). It also starts a timer for T_{max} seconds, in case it has to resend the LLRequest. Six nodes are within the range of vehicle C (the RSU, vehicles A, B, D, E, and F). When the RSU, vehicles A and B receive the second broadcast (the one from vehicle C), they realize that the propagation of the broadcast is already further away than their own position, by comparing their location with the *Sender Location* field in the broadcast. Hence, they cancel the scheduled rebroadcast. When vehicles D, E, and F receive the second broadcast (the one from vehicle C), they compute their Rebroadcast-Wait-Time (T) to schedule a rebroadcast; vehicle F having the smallest one and vehicle D the biggest one. So vehicle F will be the first to rebroadcast the LLRequest message, with the adequate modifications; that is, decrementing by 1 the *Hop Count* field, replacing the *Sender Location* with its own position, and setting *Last Location* field with the position of vehicle I. The process will continue with the rebroadcast of vehicle I, followed by the one of vehicle K. Vehicle K, the last in the line, will not receive a rebroadcast of the LLRequest with a *Sender Location* field that is further away than its own location. So after the first rebroadcast of the LLRequest, vehicle K will send up to MAXREQUEST-1 copies of the LLRequest separated by T_{max} seconds (for a total of MAXREQUEST messages of type LLRequest). If during this period, vehicle K does not receive an LLRequest further away than its own location, it will assume that it is the last vehicle in the line, and therefore will start to propagate a LLReply back to the RSU.

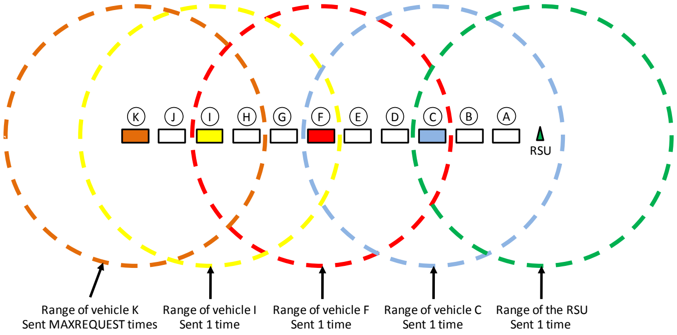


Fig. 3. Propagation of a Last Location Request Message

Fig. 4 shows the propagation of the LLReply. Vehicle K initiates the process by sending an LLReply message, which is a copy of the LLRequest it previously sent several times with *Message Type* in 1 (indicating a LLReply) and a decremented value of *Hop Count*. It also starts a timer for T_{max} seconds, in case it has to resend the LLReply. Vehicles H, I, and J are within the range of vehicle K, and therefore compute their Rebroadcast-Wait-Time (T); vehicle H having the smallest one and vehicle J the biggest one. So vehicle H will be the first to rebroadcast the LLReply message, with the adequate modifications; that is, decrementing by 1 the *Hop Count* field and replacing the *Sender Location* field with its own position. Vehicles I, J, and K will cancel their scheduled rebroadcast since they realize that the propagation of the broadcast is already closer to the RSU than their own location. Note that the *Last Location* field will not be changed by LLReply messages. The process will continue with the rebroadcast of the LLReply message by vehicles E, B, and the RSU, in sequence. The RSU rebroadcasts the LLReply just to inform vehicles A, B and C that they can cancel their scheduled rebroadcast of the LLReply.

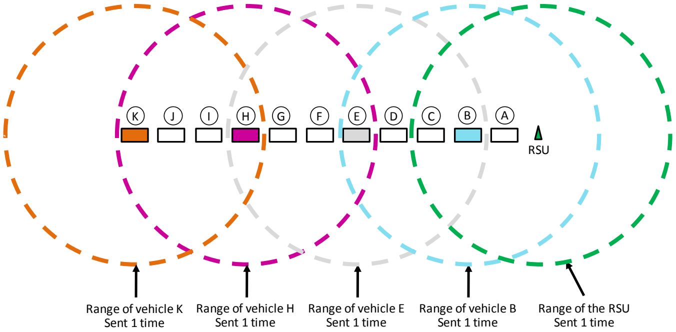


Fig. 4. Propagation of a Last Location Reply Message

Our algorithm to find the last position of a vehicle in a line of vehicles waiting for the green light will still work even if a few vehicles have a VANET device installed. Let us come back to Fig. 3 and assume that vehicle C does not have an OBU. When the RSU initiate the propagation of the LLRequest, only vehicles A and B (which are VANET-enabled nodes) will compute the Rebroadcast-Wait-Time (*T*), and vehicle B will rebroadcast first since it has the smallest Rebroadcast-Wait-Time. Therefore, the LLRequest will still advance toward the end of the line of vehicles. As it can be easily inferred, the algorithm will still work if at least one vehicle in the range of the LLRequest or the LLReply is a VANET-enabled vehicle, since this VANET-enabled vehicle will allow the progress of the message. Given that DSRC is expected to have a range of up to 1 km, the algorithm will work if at least one vehicle, every 1 km, has VANET technology. The drawback will be in an increased Round-Trip-Time (RTT). Also note that in the case that the last vehicles in a line do not have an OBU, the algorithm will report the location of the last VANET-enabled vehicle, and not the location of the last vehicle.

It is also important to notice that the algorithm will still function if the vehicles do not send periodic beacon messages. In the example of Fig. 3 and without beaconing messages, the LLRequest sent by the RSU will be the same, except for the *Last Location* field that will have the position of the RSU, instead of the location of vehicle C, since the RSU will not be able to discover vehicles A, B, and C. Then, almost immediately after the reception of the LLRequest, vehicle C will rebroadcast the request with the field *Last Location* sets with its own position; and the LLRequest will continue to progress towards the end of the line.

2.2 Obtaining the Number of Vehicles

We will now focus on obtaining the number of vehicles in a line of vehicles stopped at a light. For this purpose, we introduce two new PDUs called Count Vehicle Request (CVRequest) and Count Vehicle Reply (CVReply) as shown in Fig. 5.

Message Type	Sequence Number	Hop Count	Timestamp	Direction	RSU Location	Sender Location	Last Location	Number Vehicles
2 bytes	4 bytes	2 bytes	4 bytes	4 bytes	8 bytes	8 bytes	8 bytes	4 bytes

Fig. 5. CVRequest and CVReply Message

These new PDUs (see Fig. 5) are similar to the previous ones (see Fig. 2), but also include a field called *Number Vehicles* to store the actual number of vehicles. The RSU will be in charge of starting the process by sending a CVRequest. To simplify the explanation, we will show the process with the vehicles of Fig. 3. Before the initiation of the process, the RSU listens to beaconing messages to discover vehicles waiting in the line, and will find out that vehicles A, B, and C are waiting in the line. So the RSU will send a CVRequest with *Last Location* filed with the location of vehicle C, and *Number Vehicles* set to 3. That is, up to the position of vehicle C, 3 vehicles are waiting in the line. It also starts a timer for T_{max} seconds, in case it has to resend the CVRequest. Vehicles A, B, and C are within the range of the RSU, and therefore compute their Rebroadcast-Wait-Time (*T*); vehicle C having the smallest one and vehicle A the biggest one. So vehicle C will be the first to rebroadcast the CVRequest message. By discovering through the beaconing messages, vehicle C has determined that there are 3 vehicles (D, E, and F) that are waiting on the line and are further away from the position specified in *Last Location*. So vehicle C rebroadcast the message with the adequate modifications; that is, decrementing by 1 the *Hop Count* field, replacing the *Sender Location* field with its own position, setting *Last Location* to the one of vehicle F, and adding 3 to *Number Vehicles* (the new value is now 6). This message means that up to the position of vehicle F, 6 vehicles are waiting in the line. The CVRequest will continue its propagation through vehicles F and I, in sequence. Vehicle I will rebroadcast the CVRequest with a *Last Location* set to the position of vehicle K and *Number Vehicles* set to 11. Since the Rebroadcast-Wait-Time of vehicle K is smaller, it will act before vehicles I and J. By discovering through the beaconing messages, vehicle K has determined that no vehicles are further away from it waiting in the line. So it will rebroadcast the CVRequest with modifications in *Hop Count* and *Sender Location* up to MAXREQUEST times, before sending an CVReply in the direction of the RSU. The CVReply will be forwarded by vehicles H, E, and B, before reaching the RSU, which in turn will repeat it as a way to inform vehicles A, B, and C of the finalization of the process.

3. Discussion over the Approaches

In our proposal, determining if a vehicle is waiting in a line for the green light to come is done based on the location of the vehicle (latitude and longitude), its speed and direction. However for some applications or in some intersections, this information may be insufficient or the way to determine whether or not a vehicle has to be counted may need to be more specific. In this case, the RSU may also propagate the code to be executed for the determination of inclusion or exclusion in the LLRequest or the CVRequest.

As stated in Section 2.1, the algorithm to obtain the position of the last vehicle in a line of vehicles will still function well even if just a few vehicles have VANET technology. However, the algorithm that counts the number of vehicles will work only when most of the vehicles are equipped with a VANET device, since beaconing messages are needed to sense the presence of a vehicle. According to [1], it can take 15 to 20 years for the vehicle fleet to be equipped with V2V technology. So another option to count the number of vehicles is necessary, and can be achieved by using the proposed algorithm to obtain the location of the last vehicle to compute the distance between the RSU and the last vehicle, before dividing the obtained value by a constant (e.g., 7 m) that represents the space to accommodate one vehicle in the line.

4. Related Works

Most of the works that have been done for vehicle detection and vehicle counting are based on video cameras, and a few used sensors. For example, Wu and Gu [28] presented a new algorithm that takes advantage of digital image processing and camera optics to automatically estimate vehicle speed and vehicle count in real-time. The algorithm requires only a single video camera and a computer to operate. The camera has to be set up directly above the target road section (at least 6 meters above the road to assure satisfactory accuracy) with its optical axis tilting a certain angle downward from the highway forward direction. Daigavane and Bajaj [3] introduced a background subtraction and image segmentation using morphological transformation to detect and count dynamic objects on highways efficiently. Their proposed algorithm segments the image by preserving important edges which improves the adaptive background mixture model and makes the system learn faster and more accurately, as well as adapt effectively to changing environments. Pornpanomchai, Liamsanguan, and Vannakosit [20] proposed a system consisting of a PC connected to a video camera to detect and count vehicles. An input of the system must be a background frame without any moving vehicles and a foreground frame with moving objects. The basic idea of the system is to calculate the number of vehicles that move through the detection area from the difference between the background frame and foreground frame. Lei et al. [14] presented a video-based solution for real time vehicle detection and counting, using a surveillance camera mounted on a relatively high place to acquire the traffic video stream. The two main methods applied in their solution are (1) the adaptive background estimation for robust moving detection and (2) the Gaussian shadow elimination to deal with different size and intensity of shadows. Wang et al. [27] described an automated vehicle counting system using DSP board and image processing techniques. The counting algorithm is based on virtual loops that the end users need to initially set in the detection zones in the video image. Bas, Tekalp, and Salman [2] proposed a new video analysis method for counting vehicles using an adaptive bounding box size to detect and track vehicles according to their estimated distance from the camera given the scene-camera geometry. They employed adaptive background subtraction and Kalman filtering for road/vehicle detection and tracking, respectively. Some other works also use sensors. For example, Litzenberger et al. [16] presented an embedded system comprising a motion-sensitive optical sensor and low-cost DSP to detect and count vehicles. The detection is based on monitoring of the optical sensor output within configurable regions of interest in the sensor’s field-of-view.

As shown by this related work study, a lot of work has been done in this area, but mostly based in cameras. Sensors are another technology that is sometime also used. It seems that no work has been done to count vehicles using VANET technology. We strongly believe that many other algorithms will be soon proposed, especially algorithms that will used hybrid technology.

5. Conclusions and Future Work

In this paper, we proposed some algorithms to get the location of the last vehicle in a line of vehicles waiting for a green light. We also modify the algorithms to count the number of vehicles that are present in the line of vehicles. Our proposal minimizes the number of rebroadcasts necessary, with the use of a broadcast scheme based on position and time (Rebroadcast-Wait-Time), and has a response time closed to the best that can be achieved. To the best of our knowledge, this is the first work based on VANET in this direction.

Many public sites have important parking space problems. It can be very frustrating for a driver to be forced to circle several times in search for a vacant spot in a huge parking lot. If the driver were notified in advance of how many parking slots are available and their position in a parking lot, he or she would not have to waste time and gasoline looking for a vacant space. Parking lot management systems had been proposed [12][15] based on different technologies such as sensors and video cameras. However, up to now, just a few works have been done in this area using VANET. As future work, we plan to develop some methods to count the number of VANET-enabled vehicles in a parking lot with the aim of developing a new parking lot management system based on VANET.

6. References

[1] Achieving the Vision: From VII to IntelliDrive. http://www.its.dot.gov/press/2010/vii2intellidrive.htm.

[2] E. Bas, M. Tekalp, and F. Salman. Automatic Vehicle Counting from Video for Traffic Flow Analysis. 2007 IEEE Intelligent Vehicles Symposium. Istanbul, Turkey. June 2007.

[3] P. Daigavane and P. Bajaj. Real Time Vehicle Detection and Counting Method for Unsupervised Traffic Video on Highways. International Journal of Computer Science and Network Security, Volume.10, No.8. August 2010.

[4] S. Dornbush and A. Joshi. StreetSmart Traffic: Discovering and Disseminating Automobile Congestion Using VANETs. Vehicular Technology Conference (VTC-2007). Dublin, Ireland. April 2007.

[5] A. El-Rabbany. Introduction to GPS: The Global Positioning System, Second Edition, Artech House Publishers. August 2006.

[6] C. Farkas and Y. Kopylova. Application Level Protocol for Accident Reconstruction in VANETs. Master’s Proposal, University of South Carolina. September 2007.

[7] E. Fasolo, A. Zanella, and M. Zorzi. An Effective Broadcast Scheme for Alert Message Propagation in Vehicular Ad hoc Networks. In Proceedings of the IEEE International Conference on Communications (ICC’06). Istanbul, Turkey. June 2006.

[8] A. Ghazy and T. Ozkul. Design and Simulation of an Artificially Intelligent VANET for Solving Traffic Congestion. International Symposium on Mechatronics and its Applications (ISMA’09). Sharjah, United Arab Emirates. March 2009.

[9] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. Global Positioning System: Theory and Practice, 5th Revised Edition, Springer. September 2004.

[10] M. Ilyas and I. Mahgoub. Handbook of Mobile Computing. CRC Press. 2005.

[11] Intelligent Transportation Systems. US Department of Transportation. http://www.its.dot.gov.

[12] H. Jung, D. Kim, P. Yoon, and J. Kim. Light Stripe Projection based Parking Space Detection for Intelligent Parking Assist System. In Proceedings of the 2007 IEEE Intelligent Vehicles Symposium.Istanbul, Turkey. June 2007.

[13] S. Kohli, B. Kaur, and S. Bindra. A comparative study of Routing Protocols in VANET. In Proceedings of the 2010 International Symposium on Computer Engineering & Technology (ISCET 2010). Mandi Gobindgarh, Punjab, India. March 2010.

[14] M. Lei, D. Lefloch; P. Gouton, and K. Madani. A Video-Based Real-Time Vehicle Counting System Using Adaptive Background Method. The Fourth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS '08). IEEE International Conference on. Bali, Indonesia. December 2008.

[15] S-F. Lin, Y-Y. Chen, and S-C. Liu. A Vision-Based Parking Lot Management System. In Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics. Taipei, Taiwan. October 2006.

[16] M. Litzenberger, B. Kohn, G. Gritsch, N. Donath, C. Posch, N.A. Belbachir, and H. Garn. Vehicle Counting with an Embedded Traffic Data System using an Optical Transient Sensor. The 10th International IEEE Conference on Intelligent Transportation Systems (ITSC’07). Seattle, WA, USA. September 2007.

[17] S-Y. Ni, Y-C. Tseng, Y-S. Chen, and J-P. Sheu. The Broadcast Storm Problem in a Mobile Ad hoc Network. In Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 151-162. August 1999.

[18] T. Osafune, L. Lin, and M. Lenardi. Multi-Hop Vehicular Broadcast (MHVB). In Proceedings of the 6th International Conference on ITS Telecommunication. Chengdu, China. June 2006.

[19] F. Padron and I. Mahgoub, Traffic Congestion Detection Using VANET. Florida Atlantic University, Tech. Rep., 2010.

[20] C. Pornpanomchai, T. Liamsanguan, and V. Vannakosit. Vehicle Detection and Counting from a Video Frame. International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR '08). Hong Kong. September 2008.

[21] W. Rothery. Car Following Models. Traffic Flow Theory, Transportation Research Board, Special Report 165, 1992.

[22] M. Slavik and I. Mahgoub. Stochastic Broadcast for VANET. In Proceedings of the 7th IEEE Consumer Communications and Networking Conference (CCNC'10). Las Vegas, Nevada, USA. January 2010.

[23] M. Slavik and I. Mahgoub. Statistical Broadcast Protocol Design for Unreliable Channels in Wireless Ad-Hoc Networks. IEEE Globecom 2010 - Wireless Networking Symposium (GC10 - WN), Miami, Florida, USA, December 2010.

[24] S. Soloman. Sensors Handbook. McGraw-Hill Professional; 2nd Edition. November 2009.

[25] Texas Transportation Institute. TTI’s 2010 Urban Mobility Report. The Texas A&M University System, Tech. Rep., Dec. 2010. http://mobility.tamu.edu.

[26] O. Tonguz, N. Wisitpongphan, F. Bai, P. Mudalige, and V. Sadekar. Broadcasting in VANET. Mobile Networking for Vehicular Environments. Anchorage, Alaska, USA. May 2007.

[27] K. Wang, Z. Li, Q. Yao, W. Huang, and F-Y. Wang. An Automated Vehicle Counting System for Traffic

Surveillance. 2007 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2007). Beijing, China. December 2007.

[28] J. Wu and C. Gu. The Design and Implementation of Real-Time Automatic Vehicle Detection and Counting System. International Conference on Information Engineering and Computer Science (ICIECS 2009). Wuhan City, China. December 2009.

[29] S. Yi. Design and Construction of LAN based Car Traffic Control System. World Academy of Science, Engineering and Technology. Issue 46, pp. 612-615. October 2008.

LDPC Generation and New Decoding Process based on Convolutional Encoder

Jongsu Lee[1], Youil Lee [2], Joon-Young Jung[3] and Sangseob Song[4]
[1]Department of Electronics Engineering, Chonbuk University, Jeonju, South Korea
[2]Department of Electronics Engineering, Chonbuk University, Jeonju, South Korea
[3]Electronics and Telecommunication Research Institute, Daejeon, South Korea
[3]Department of Electronics Engineering, Chonbuk University, Jeonju, South Korea

Abstract - *In this paper, we suggest a new technique for LDPC parity-check matrix (H-matrix) generation and a corresponding decoding process. The key idea is to construct LDPC H-matrix by using a convolutional encoder. It is easy to have many different coderates from a mother code with convolutional codes. However, it is difficult to have many different coderates with LDPC codes. Constructing LDPC H-matrix based on a convolutional code can easily bring the advantage of convolutional codes to have different coderates. Moreover, both LDPC and convolutional decoding algorithms can be applied altogether in the decoding part. This process prevents the performance degradation of short-length LDPC code.*

Keywords: Systematic convolutional encoder, Low Density Parity Check code, Parity-check matrix, Sum-product algorithm, Viterbi decoder

1 Introduction

For the last several decades or so, many researchers have studied low-density parity check codes (LDPC) [1] and turbo codes [2] that is possible to achieve the practical realizations of Shannon's theory [3], which have revolutionized the field of error correction coding [4].

Shannon's claim can be realized by a technique referred to as forward error correction. The basic idea is that of incorporating redundant bits, or check bits, thus creating what is known as a codeword. If the check bits are introduced in an appropriate manner so as to make each codeword sufficiently distinct from each other, the receiver then becomes capable of determining the most likely codeword that has been transmitted.

Our main interest, LDPC codes are typically referred to as linear block codes. A code is termed a block code, if the original information bit-sequence is segmented into fixed-length message blocks, denoted by $\boldsymbol{u}=u_1,u_2,...,u_K$, each having K information bits. The LDPC encoder is then capable of transforming each input message block $\boldsymbol{u}$ according to a predefined set of rules into a distinct N-tuple $\boldsymbol{c}$, which is typically referred to as the codeword. In this way, there are $2K$ distinct legitimate codewords among $2N$ codewords.

This paper shows how to generate LDPC generator matrix and parity matrix by using convolutional codes. Moreover, in a receiver, both Viterbi decoding and Sum-product algorithm can be applied together. This is the reason why we suggest here applying two decoding process at the same time with one encoding process.

The status of wireless channel is unpredictable and changing continuously. For the transmission efficiency, the coderate of a codeword should be changed every moment according to the channel status. In this sense, it is very important to make different coderates from one mother code in forward error correcting system. However, it is difficult to make many different coderates with LDPC codes from a mother code. This problem can be easily solved by constructing LDPC codes with convolutional codes.

Also codeword size of a codeword should be changed every transmitting moment according to the channel status for the transmission efficiency. For any channel codes, there exists bit error (BER) performance degradation in case of short-length frame size. Applying both Viterbi and sum-product algorithm together a little bit prevents this BER performance degradation which is caused by using short-length frame size codeword.

Section 2 describes how to generate LDPC generator matrix and parity matrix for a block channel code by a convolutional encoder. Section 3 shows new decoding process corresponding to the LDPC generator matrix which is constructed in section 2 by a convolutional encoder. Section 4 shows simulation results and section 5 concludes our paper.

2 LDPC based on a convolutional code

The unique and distinctive nature of the codewords implies that there is a one-to-one mapping between a K-bit information sequence and the corresponding N-bit codeword described by the set of rules of the encoder. This LDPC codes are linear codes and thus the codeword **c** can be calculated by multiplying the input message sequence **u** with a K by N element G-matrix, which is referred to as the generator matrix. We also note that G can also be transformed into what is referred as the systematic matrix form, i.e., to G=[I_K P], where I_K is a K by K element identity matrix and P is K by N-K matrix.

There is another useful matrix associated with a linear block code. This matrix is referred to as the parity-check matrix, which is typically denoted by H and contains N-K by N elements. If the generator matrix is in the systematic matrix form, then the H-matrix of the code is given by H=[-P^T I_{N-K}], where I_{N-K} is an identity matrix of dimension N-K by N-K. A characteristic of the H-matrix of LDPC codes is that it is sparse, i.e., there are fewer ones than there are zeros. As a result, their H matrix is said to have a low-density hence the terminology of low-density parity check codes.

2.1 LDPC H-matrix generation

Figure 1 shows a Forward Error Correcting (FEC) system over Additive White Gaussian Noise (AWGN) channel.

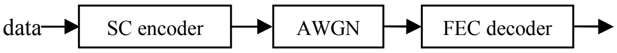


Fig. 1. Block diagram of new encoding and decoding process with AWGN channel

By truncating at every frame size, any convolutional encoder can be seen as a generator matrix (G-matrix) for a block channel code. In our work specifically for the simulation result, a systematic convolutional (SC) encoder is used with which code rate is 1/2, constraint length is 3, and code vector is (4, 5) to generate corresponding G-matrix and H-matrix. Here, code vector (4, 5) is represented by an octal number. This SC encoder is depicted in figure 2.

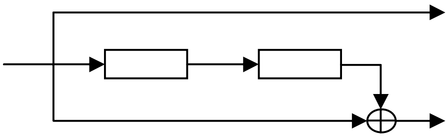


Fig. 2. Systematic convolutional encoder (rate=1/2, constraint length=3, code vector=(4,5))

Here, this SC code is not optimum code. The code vector for the optimum convolutional code with constraint length of 3 and code rate of 1/2 is (5,7) as an octal number [5]. This encoder is non-systematic convolutional encoder. The reason why the optimum non systematic convolutional code is not used for generating corresponding G-matrix and H-matrix is the 4-cycle factor in H-matrix. 4-cycle factor exists when there are at least 2 ones in the same columns across every each row. This 4-cycle factor is called a girth in LDPC code. Figure 3 shows this 4-cycle factor with H-matrix and Tanner graph.

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

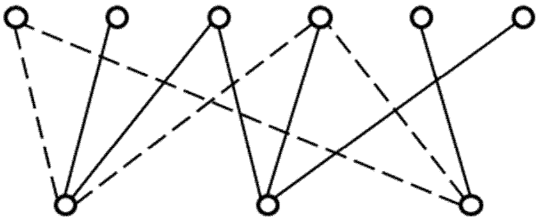


Fig. 3. 4-cycle factor with H-matrix and Tanner graph

In figure 3, there are 2 ones in the third and forth columns across first row and second row from the H-matrix. This causes 4-cycle factor. Also 4-cycle factor can be seen as a dashed line from the Tanner graph starting from one node to itself.

If there exits 4-cycle factor in H-matrix or in Tanner graph, there is no guarantee for the performance of LDPC decoder. Considered to this 4-cycle factor, SC code as shown in figure 2 is used. As an example, the identical G-matrix and H-matrix are generated in figure 4 according to the above SC encoder for the frame size of 6 bits.

$$G = \left[\begin{array}{cccccc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$H = \left[\begin{array}{cccccc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Fig. 4. Identical generator matrix and parity-check matrix to the above non-systematic convolutional code (N=12,K=6)

Code vector of the above SC encoder is (4, 5) as an octal number. This vector can be seen to (100, 101) as a binary number. In figure 4 on G-matrix, this each code vector is located on the first row at left side and right side with (K-3) zeros padding. From the second row on G-matrix, this each code vector is right-shifted. Then, this G-matrix is exactly identical to SC encoder by truncating every each K-bit frame size except the arrangement of codeword. Let left side of G-matrix be identical matrix I with K by K size and right side of G-matrix be P with K by N-K size. To generate the corresponding H-matrix, the left side of H-matrix becomes P^T

and right side of H-matrix becomes *I* with *N-K* by *N* size. For a block channel code, the below equation must be satisfied between G-matrix and H-matrix [6].

$$G^{\cdot}H^{\mathrm{T}} = 0 \tag{1}$$

The identical G-matrix and H-matrix generated by the above SC encoder are satisfied with the equation.

3 Decoding process

In FEC encoder, a codeword is made by using SC encoder. Even though a codeword is made by SC, it is considered to be the same as a codeword by using the identical G-matrix. That is how we can apply LDPC decoder to received codeword over AWGN channel in a receiver. Also, since a codeword is made by a convolutional encoder, Viterbi decoder can be applied to decode a received codeword too. In this sense, we suggest a new decoding approach to use both Viterbi decoder and LDPC decoder to a received codeword encoded once by SC encoder in a transmitter. Figure 5 depicts that the block diagram of FEC decoder consists of Viterbi decoder and LDPC decoder.

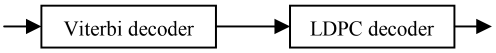


Fig. 5. Block diagram of FEC decoder

First, each frame of a received codeword through AWGN channel is decoded by Viterbi decoder. Log-likelihood Ratio (LLR) of information bits is taken as the output value from Viterbi decoder. Second, a codeword once decoded in Viterbi decoder becomes the input to LDPC decoder. This LDPC decoder uses H-matrix generated by the convolutional encoder.

For example, if received codeword size is 96 bits, then the input to Viterbi decoder is 96-bit codeword. If the coderate is 1/2 and encoder is systematic, then half of 96-bit codeword (46 bits) is information part, the other part is parity part. If the number of received signals to Viterbi decoder is 96 bits, then the output from Viterbi decoder is 48 bits. The decoded output value becomes now input value to LDPC decoder. However in LDPC decoder whole received signals are needed to be decoded. In this reason, the input value is added between the output LLR value of Viterbi decoder and the half of originally received signal, the parity part, from the AWGN channel. Then the input size to LDPC decoder is 96 bits. After the iterative decoding in LDPC decoder, the final decoded codewords are retrieved. Because of the property of the systematic codes, the first half of the decoded codewords becomes the estimated information messages.

3.1 Viterbi decoding

A convolutional encoder is a finite state machine. Hence the optimum decoder is a maximum likelihood sequence estimator (MLSE) for signals with memory. Therefore, optimum decoding of a convolutional code involves a search through the trellis for the most probable sequence. First branch metric (BM) value is calculated with the received signals. Then, forward metric (FM) value and backward metric (KM) value are calculated. Finally the log likelihood ratio (LLR) value can be taken by using these three metric values. Since our paper is focused on the LDPC, we here skip the specific Viterbi decoding algorithm.

3.2 Sum-Product algorithm

The SP algorithm [7] is a multi-iteration procedure that exchanges the extrinsic log likelihood ratio (LLR) information between check and variable nodes until the stopping criterion is satisfied. Usually this means that either the parity check equations are satisfied or the preset maximum iteration number is reached. At every iteration step, the SP algorithm consists of two sequential steps: variable node decoder followed by check node decoder. Let p_j^0 and p_j^1 be a priori probabilities of bit values 0 and 1 for each bit *j*. Let P^0_{ij} and P^1_{ij} be probabilities that bit *j* is 0 or 1, given the parity checks other than check *i*. Also, let Q^0_{ij} and Q^1_{ij} be probabilities that check *i* is satisfied by a value of 0 in bit *j* given the current values of the other bits. The standard SP decoding algorithm works as follows:

1. The horizontal (row) step at *P* check nodes

Define $dP_{ij}=P^0_{ij}-P^1_{ij}$

For each i,

Compute $dQ_{ij'}$ as the product of dP_{ij} for all *j'* is not *j*.

Set Q^0_{ij} be $1/2(1+dQ_{ij})$, Set Q^1_{ij} be $1/2(1-dQ_{ij})$.

2. The vertical (column) step at *N* variable nodes

For each j,

Compute P^0_{ij} as p^0_j times the product of $Q^0_{i'j}$ for all $i'\neq i$

Compute P^1_{ij} as p^1_j times the product of $Q^1_{i'j}$ for all $i'\neq i$.

Scale the values of P^0_{ij} and P^1_{ij} such that $P^0_{ij} + P^1_{ij}=1$

For each *j*,

Compute P^0_j as p^0_j times the product of Q^0_{ij} for all *i*.

Compute P^1_j as p^1_j times the product of Q^1_{ij} for all *i*.

Scale the values of P^0_j and P^1_j such that $P^0_j + P^1_j=1$.

3. Last step

Check $c \cdot H^T = 0$ (use hard decision value), if fails the values of P^0_{ij} and P^1_{ij} are fed back to another iteration of the horizontal step.

4 Simulation result

In this paper, we choose the parameter of convolutional encoder with which code rate is 1/2, constraint length is 9, and code vector is (400, 435). With this convolutional encoder, we constructed corresponding G-matrix and H-matrix for our simulation as explained in section 2. After analyzing the error correcting performance of the suggested decoding process through AWGN channel, the BER performance versus Eb/N0 (dB) is depicted in figure 5 and 6.

In figure 5 during the simulation, a convolutional encoder is used for encoding information bit sequence into a codeword and only sum-product algorithm is used for decoding received signal sequence into an information bit sequence. In order to compare BER performance of the suggested LDPC codes, Mackay LDPC codes are selected of which the coderate is 1/2 and the block size is the same as the suggested code (*N*=96, *K*=48) [8]. As it is shown as figure 5, there is no performance difference between Mackay LDPC codes and suggested LDPC codes.

In figure 6, a convolutional encoder is used for encoding information bit sequence into a codeword and both Viterbi decoder and sum-product algorithm are used for decoding received signal sequence into an information bit sequence at the same time. In this case, there is about 0.5 dB of Eb/N0 coding gain versus Mackay LDPC code at BER is equal to 10^{-5} with which *N*=96, and *K*=48, and coderate is 1/2.

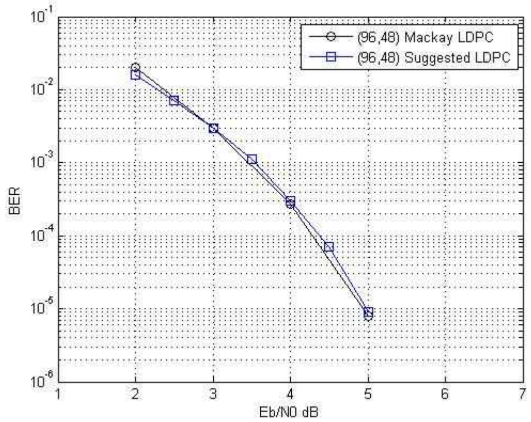


Fig. 5. BER performance of (*N*=96, *K*=48) LDPC codes for Mackay H-matrix and the suggested LDPC with SP algorithm

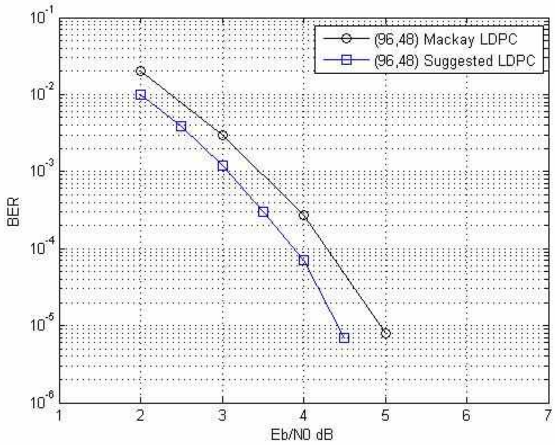


Fig. 6. BER performance of (*N*=96, *K*=48) LDPC codes for Mackay H-matrix and the suggested LDPC with Viterbi and SP algorithm

5 Conclusion

LDPC code selected for the simulation has the property of *N*=96, *K*=48 and coderate is 1/2. Mackay LDPC code and convolutional code basis LDPC code are simulated for comparison. As expected, LDPC code we suggested in our paper shows better BER performance in a short-length LDPC codes. We analyzed and simulated the code with which *N*=96, *K*=48 and coderate is 1/2. As a future work, different length of LDPC codes should be studied and found. Then because of the property that many different coderate can be made from the mother code in case of convolutional codes, BER performance analysis is necessary for many different coderates.

6 Acknolegement

This work was supported by the IT R&D program of KCC. [10035709, Development of Next Generation Digital Cable Transmission Technology]

7 References

[1] R.G.Gallager. “Low density parity check codes,” IRE Trans. Inform Theory, vol.IT-8, pp.21-28, Jan. 1962

[2] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error-correcting coding and decoding: Turbo-codes,” in Proc. IEEE International Conf. Commun., Geneva Technical Program, vol. 2, (Geneva, Swizerland), pp. 1064-1070, May 23-26, 1993

[3] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical J., vol. 27, pp. 379-423, 1948

[4] G.D. Forney Jr. and D. J. Costello Jr., “Channel coding: The road to channel capacity,” in Proc.IEEE,vol.95,pp.1150-1177,June 2007

[5] Bernard Sklar, “Digital Communications”, Second edition, pp. 418-419

[6] J. G. Proakis, M. Salehi, “Digital Communication” Fifth edition, pp. 412-413

[7] R.M.Tanner,”A recursive approach to low complexity codes, ”IEEE Trans. Inform. Theory,, vol.IT-27, no. 5, pp. 533-547

[8] Mackay, “http://www.inference.phy.cam.ac.uk/mackay/codes/da ta. html#l1”

Introducing the concepts of swarm intelligence and genetic algorithms in cognitive networks

Laiti Apoorva
Undergraduate student, Bachelor of engineering
Birla Institute of Technology
Patna, India.

Abstract:

At the outset, I am to concentrate on the concepts of swarm intelligence and genetic algorithms, which enhance the scope of usage of cognitive networks. There are an ever-increasing number of wireless mobile nodes. This necessitates the instillation of an element of cognition at the mobile node. Here evolves cognitive network that can detect its operating wireless environment and learn how to adapt and evolve. The DARWIN THEORY of evolution-the basis for GENETIC ALGORITHM, models this evolution. The world is moving towards vastly higher bandwidth applications. The need for flexible and dynamic allocation of frequency bands opens the door for the novel SWARM INTELLIGENCE based approach. This satiates the thirst for bandwidth. The detailed approach for both the above concepts is mentioned by introducing cognitive chromosome and cognitive sub-network. A solution for the problem of re-initialization in cognitive sub-networks is also proposed.

Keywords: Genetic algorithm, swarm intelligence, fitness measure, re-initialization, optimization.

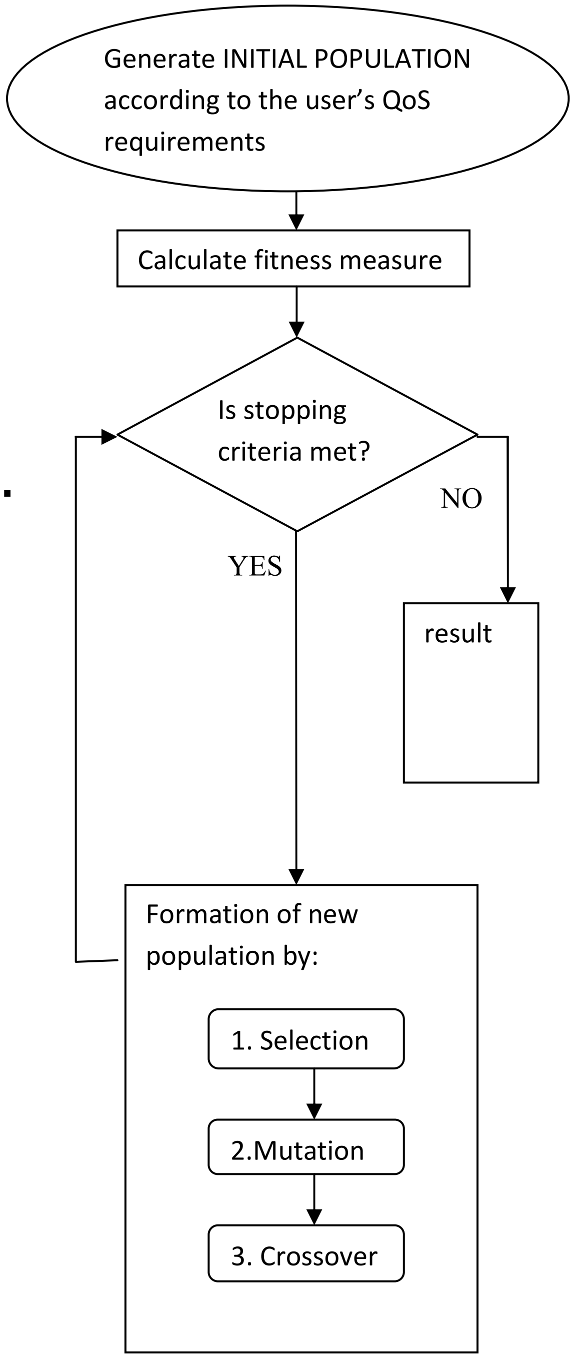
1 Introduction

The emerging wireless technologies and their QoS (Quality of Service) requirements lead to dramatic rise in the demand for bandwidth. The frequency bands are allocated by selling access rights exclusively to users. According to Federal Communications Commission's (FCC's) report, only 15% to 85% of licensed spectrum is used, i.e. about 90 percent of the licensed spectrum is unused. Fundamental physical laws limit the availability of free spectrum .Spectrum scarcity is mainly due to the inefficient use of spectrum. Hence, the availability of free spectrum is limited. The FCC unanimously agreed in November 2008 to open up the unused spectrum for unlicensed use. Cognitive network, an adaptive data network, has the potential to open up secondary or complementary spectrum markets. A cognitive network provides efficient utilization of spectrum and network resources and thereby provides competitive wireless services to the consumer. Cognitive Wireless Network technologies employ intelligent decision-making for finding unoccupied spectrum or channel. It dynamically reconfigures the networks and gets itself adapted to the available Networks. The dynamic reconfiguration is possible due to the autonomic multiplexing in cognitive networks [1]. Cognitive networks adopt the idea of coupling the network devices with sensors to sense network conditions and perform autonomously with the minimum possible user intervention. Cognitive networks exploit self-learning policy. The network elements observe the network conditions and then, using prior knowledge gained from previous interactions with the network, plans, decides and acts on this information. These actions are taken with respect to the end-to-end goals of a data flow. In the modern communication networks, emphasis is laid on network centric goals rather than node based goals. The element of cognition organizes the network nodes to allow for cooperation and distributed reasoning and learning and achieving the required goals [2].

2 Genetic Algorithm (GA)

Genetic algorithm is an optimization algorithm that tries to imitate the evolutionary behavior of the living creatures [3]. The algorithm applies the rules of nature. Genetic algorithms (GAs) are algorithms rooted in biological functions like reproduction and evolution. Darwin's theory is the base for these algorithms.

Figure1A flowchart depicting GA


Generate INITIAL POPULATION according to the user's QoS requirements
Calculate fitness measure
Is stopping criteria met?
NO
YES
result
Formation of new population by:
1. Selection
2.Mutation
3. Crossover

It involves all the methods described in Darwin's theory like, selection, crossover and mutation. These algorithms are capable of exploring large solution spaces. The main concept behind genetic algorithm is trial and error [3]. The algorithms work best for centralized problems where the environment is well known. Hence, they can be used to achieve the required network centric goals [4].

2.1 Explanation of the flow chart

First we generate an initial population. we take the user's QoS requirements. The initial population is generated according to the QoS requirements. The next step is evaluating fitness measure. It is calculated from a fixed set of formulae given in section 3.2. The third step is checking the condition of stopping criteria. The stopping criteria can be set according to our own decision. If time is a matter of consideration, then stopping criteria is set to some fixed number of generations. Stopping criteria can also be set as the time required to achieve the required result. Formation of new population takes place in three steps.

"Selection" Genetic algorithms take a population of chromosomes using this genetic operation. Its purpose is to choose the individuals from the current population, which will go into an intermediate population. Only individuals in this intermediate population will be chosen to mate with each other.

"Crossover", it involves choosing two individuals to swap segments of their code, producing artificial "offspring" that are combinations of their parents. This process is intended to simulate the analogous process of recombination that occurs to chromosomes during sexual reproduction. There are two types of crossover. Single-point crossover, in which a point of exchange is set at a random location in the two individuals' genomes, and one individual contributes all its code from before that point and the other contributes all its code from after that point to produce an offspring. Uniform crossover, in which the value at any given location in the offspring's genome is either the value of one parent's genome at that location or the value of the other parent's genome at that location, chosen with 50/50 probability [5].

"Mutation", just as mutation in living things changes one gene to another, so mutation in a genetic algorithm causes small alterations at single points in an individual's code.

3. Proposing a cognitive chromosome

The genetic algorithms operate on chromosomes. A chromosome is the basic structure of the GA. The chromosomes represent a multi-dimensional solution search space. Chromosomes are comprised of numerous individual “genes”. Each gene represents different parameters. In our consideration we assumed a chromosome with six different genes namely data rate, error rate, bandwidth, operating frequency, signal power and modulation technique [6].

Table 1: The genes of a typical chromosome of our consideration-
THE COGNITIVE CHROMOSOME

DATA RATE	ERROR RATE	BANDWIDTH	OPERATING FREQUENCY	SIGNAL POWER	MODULATION TECHNIQUE

Table 2: A table showing the genes of the cognitive chromosome and the range of values of the genes and the encoding of the values in binary and decimal formats.

	DATA RATE	ERROR RATE	BAND-WIDTH	FREQUENCY	POWER	MODU-LATION
RANGE	0-4Mbps	10^(-1) to 10^(-16)	0 to 8 MHz	0 to 40MHz	-30dBm to30dBm	given below
NUMBER OF INTERVALS	16	16	8	500	61	8
VALUE OF INTERVAL	250 Kbps	10^(-1)	1MHz	80KHz	1dBm	given below
DECIMAL VALUES	0-15	0-15	0-7	0-499	0-60	0-7
NUMBER OF BINARY BITS	4	4	3	9	6	3
GENE WEIGHT	14%	14%	10%	31%	21%	10%
Fitness Point value	7	8	4	200	20	1

Table 3: A table of the gene, modulation technique, with decimal values for each technique.

DECIMALVALUES:	0	1	2	3	4	5	6	7
MODULATION TECHNIQUE:	BPSK	QPSK	GMSK	16 QAM	DPSK	MSK	OFDM	OOK

3.1 Analyzing the cognitive chromosome:

The proper representation of the chromosome will increase the performance of the GA by limiting the search space. The range of each gene is divided into intervals. Each interval is represented in two ways. Decimal representation and binary representation. The gene weight is found by dividing the number of binary bits needed to represent a gene by the total number of binary bits required. Let the fitness point value be ‘n’. To find an optimal value for a particular gene, the genetic algorithm considers n values in each direction of the user’s requested value. If the value of n is half the interval size, it will cover the whole search space. The fitness point value is generally approximated to the half the value of the interval size in order to optimize the solution. The fitness point value is taken as one for modulation technique because we cannot compromise on the modulation technique used. The modulation technique used is the same as the one requested by the user.

The modulation techniques involve BPSK binary phase shift keying, QPSK quadrature phase shift keying, GMSK Gaussian minimum shift keying, 16 QAM quadrature amplitude modulation, DPSK differential phase shift keying, MSK minimum shift keying, OFDM orthogonal frequency division multiplexing, and OOK on-off keying.

3.2 FORMULAE USED IN DETERMINING FITNESS MEASURE OF THE GENES

Let the absolute value of gene be ABS.

ABS of a gene = (the value of our cognitive chromosome's gene)-(the value of the gene requested by the user) (1)

FP=fitness point value

$$\text{if FP} > ABS$$
$$\text{Fitness measure of gene} = [(\text{Gene weight} \times \text{ABS}) \div (\text{FP})] \qquad (2)$$

$$\text{if FP} \leq \text{ABS}$$
$$\text{Fitness measure of gene} = \text{Gene weight} \qquad (3)$$

Total fitness measure of a chromosome= 100-(sum of fitness measures of all the genes) (4)

The selection process can be done by Roulette-wheel selection. It is a form of fitness-proportionate selection. The selection probability p(i) of an individual is proportional to its fitness measure f(i) [6].

$$p(i) = f(i) / \sum_{i=1}^{n} f(i) \qquad (5)$$

p(i)=selection probability of ith chromosome
f(i)=fitness measure of the ith chromosome
n=number of chromosomes

Secondly, crossover is performed with a crossover rate 80%. Thirdly, mutation is performed with a mutation rate 2%. This generates a new population. The stopping criterion is set either by limiting the number of generations or by limiting the process until we achieve the required fitness value.

3.3 RESULTS:

To achieve satisfactory results, the QoS requirements of the user should be within the specified range (given in the Table2). The above-mentioned process can be simulated in MATLAB. The QoS (Quality of Service) requirements (i.e., data rate, error rate, bandwidth, power etc.,) of the user can be applied on different possible sets of initial population. The fitness measure of each chromosome can be calculated using genetic algorithm toolbox in the MATLAB model. The graph of the fitness measurement can be plotted in MATLAB [6].

4 Swarm Intelligence:

Have you ever been marveled by the brilliant motion of the flocks of birds that fly in a cooperative behavior and perfect synchronization? The underlying intelligent principle is the SWARM INTELLIGENCE [7]. Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial [8]. The present networks employ 'fundamentally dumb' sensors. 'Dumb' in the sense, they require a centralized server for the information processing. By extending the principle of swarm intelligence, in which every bird looks through the eyes of all other birds, to our cognitive networks, we make an 'intelligent sensor'. We beneficially utilize the cooperation and coordination among all our 'intelligent sensors' for the effective spectrum hole detection [9][10].

4.1 Advantages of swarm intelligence

1. Swarm intelligence matches the dynamic nature of the Cognitive network very well.
2. A Secondary User is allowed to access a spectrum hole only when causing tolerant interference to the Primary Users working nearby on that hole. This protects the interests of primary users.
3. The communications among the Secondary Users are not falsely recognized as Primary User signals, due to the synchronized spectrum sensing process.
4. Due to flexibility, the swarms can quickly adapt to a changing environment.
5. The swarm network is robust: even when one or more individuals fail, the group can still perform its tasks.
6. Self organization: the group needs relatively little supervision or top down control. They have little or no centralized control.
7. The subunits are effectively autonomous. High connectivity exists between the subunits [11].
8. Swarms are not goal directed; they react rather than plan extensively.
9. Swarms are simple, with minimal behavior and memory.
10. Control is decentralized; there is no global information in the system.
11. Swarms can react to dynamically changing environments.
12. Direct swarm interaction is not required [12].

5 Particle swarm optimization (PSO)

PSO is a versatile population-based optimization algorithm, in which population individuals are modeled as particles in the multidimensional search space.PSO is inspired by nature. PSO is a computational method that optimizes a problem iteratively. PSO is suitable for optimization problems that are relatively irregular, noisy, or dynamic [13]. The large numbers of members that make up the particle swarm make the technique impressively resilient to the problem of local minima.

5.1 Simplicity of PSO

1. PSO is conceptually very simple.
2. PSO can be implemented in just a few lines of code, requiring no operations more advanced than scalar multiplication [14].
3. PSO is utilized in many fields of optimization [15], [16].

5.2 Comparison between PSO and GA:

5.2.1 Similarities:

1. PSO and GA do not require the objective function to be differentiable. Hence, they are suitable for cognitive networks.
2. Both algorithms start with a group of a randomly generated population.
3. Both have fitness values to evaluate the population.
4. Both systems do not guarantee success [17].

Differences:

1. PSO is prone to premature convergence than GAs
2. PSO does not have genetic operators like crossover and mutation.
3. Particles update themselves with the internal velocity.
4. PSO has memory, which is important to the algorithm.
5. The information sharing mechanism in PSO is different from GA. In GAs, chromosomes share information with each other. So the whole population moves like a one group towards an optimal area. It is a one -way information sharing mechanism in PSO.
6. In PSO, the evolution only looks for the best solution. All the particles tend to converge to the best solution quickly [17].

6 A swarm sub-network and its significance:

A single swarm of particles represents the cognitive sub-network and the search within the sub-network is achieved through PSO. Each sub swarm in PSO has a fixed number of particles. The cognitive network is divided into several sub-networks, each searching in a different region of the spectrum. Information sharing within a swarm is global adopt the idea of coupling the network devices (often nodes with software defined radios) with sensors to sense network conditions a network that can perform autonomously with the minimum possible user intervention. PSO starts with a single swarm and adds additional swarms as needed (all consisting of a predefined number of particles)

6.1 Significance

1. Multiple vacant bands can be found.
2. As the cognitive sub–networks work simultaneously the overall processing time can be reduced.
3. Effects due to multi–path propagation can be met in a more efficient way.

7 A problem in a swarm sub-network:

Two different swarm sub-networks start their search. Their search proceeds in different directions. Maximum probability is that, their search leads to different frequency bands. However, there are chances that both the sub-networks may allocate the same frequency band to the secondary user. This leads to collision of sub-networks.

7.1 The Solution:

The search space of our cognitive network is the entire RF spectrum .it is indeed a tedious and time-consuming task for a single network to search the whole spectrum. The only plausible option at our disposal is the cognitive sub-network. The requirement for a swarm sub-network is that the swarm sub-networks collaborate among each other and scan the frequency range simultaneously. If this collaboration is effective then several vacant frequency bands can be found.

7.1.1 The present day solution

This solution is based on the principle of exclusion [18]. It states that if two sub-networks collide there exists a competition between the sub-networks. Taking the help of the genetic algorithm the fitness of both the competing sub-networks is calculated. The sub-network with the minimum fitness is reinitialized. Here minimum fitness implies maximum fitness measure. This is not a viable option. This solution poses a serious bottleneck to the whole process.

7.1.2 The proposed solution:

To plug in the bottleneck we consider the time based sharing of the allocated spectrum. In this method, we consider the fitness of each swarm sub-network. The sub-network with the maximum fitness (minimum fitness measure) gets 'more' time compared to the one with the minimum fitness. The word 'more' here is a measure of the difference between the fitness measures of the two sub-networks.

7.1.2.1 Drawbacks of present day solution

The spectrum resources depend on frequency, time, and location.
1. This method increases the search time. The search time is the time taken by a swarm sub-network to find a vacant frequency band.
2. It decreases the probability of a successful search, posing a serious bottleneck to the whole process.
3. This solution cannot cope up with the dynamic real world problems.

7.1.2.2 Advantages of proposed solution

1. There is no problem of re-initialization. Initialization is a crucial aspect for effective spectrum detection. A good initialization decreases the search time. A bad initialization leads to improper setup of communication [18].
2. It is suitable for dynamic real world problems where frequency allocation will change in time.
3. It increases the probability of successful search because both the swarm sub-networks get a chance in time based sharing.

8 Conclusion

In this paper, we emulated the natural processes like genetic algorithm and swarm intelligence to our cognitive networks. We proposed a 'Cognitive Chromosome', a 'Swarm Sub-network', and an 'Intelligent Sensor' in the cognitive network. We gave a detailed description of all the genes present in our cognitive chromosome. We solved the problem of collision between two sub-networks. Moreover, we proposed a solution to the problem of re-initialization as the time based spectrum sharing. There exists a large scope for future research in the cognitive field, which involves the intermingling of cognitive science and the wireless network technology. NOESIS, which is the essence of existence of we human beings, will surely succeed in making the cognitive networks ubiquitous in the nearby future.

References:

[1] Self adaptive and cognitive network elements by Mikhail Smirnov. Fraunhofer Institute FOKUS 2009.

[2] Island Genetic Algorithm-based Cognitive Networks by Mustafa Y. El-Nainay. Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University2009

[3] COGNITIVE RADIO AND GAME THEORY: OVERVIEW AND SIMULATION by Mohamed Gafar Ahmed Elnourani. Thesis presented as part of Degree of Master of Science in Electrical Engineering Blekinge Institute of Technology December 2008.

[4] Cognitive Networks by Captain Ryan W. Thomas. Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University June 15, 2007

[5] Genetic Algorithms and Evolutionary Computation by Ms.Nagori Meghna, Ms. KaleJyoti IJCSNS International Journal of Computer Science and Network 126 Security December 2010.

[6] Spectrum Optimization in Cognitive Radio Networks using Genetic Algorithms by Taimoor Siddique and AdnanAzam. Thesis presented as part of Degree of Master of Science in Electrical Engineering Blekinge Institute of Technology 2010.

[7] E. Bonabeau, G. Theraulaz, and M. Dorigo, "Swarm intelligence: From Natural to Artificial Systems," Oxford University Press, NewYork, 1999.

[8]www.wikipedia.org

[9] Q. Bi, G. I. Zysman, and H. Menkes, "Wireless Mobile Communications at the Start of the 21st Century," IEEE Communications Magazine, January 2001, pp. 110–116.

[10] R. Berezdivin, R. Breinig, and R. Topp, "Next-Generation Wireless Communications Concepts and Technologies," IEEE Communications Magazine, vol. 40, no. 3, March 2002, pp. 108–116.

[11] Swarm Intelligence: From Natural to Artificial Systems (Oxford University Press, 1999) by Rodney Brooks professor MIT

[12]Expert Assessment of Stigmergy: A Report for the Department of National Defence by Tony White Associate Professor Carleton University 16th May 2005

[13] Particle Swarm Optimization (PSO) of Power Allocation in Cognitive Radio Systems with Interference Constraints .By Saeed Motiian1, Mohammad Aghababaie1 And Hamid Soltanian-Zadeh, 1,2,3 Senior Member, IEEE

[14] Swarm-Intelligent Localization by Pontus Ekberg August2009

[15] G.A. Laguna-Sanchez, R. Barron-Fernandez, "Blind channel estimation for power-line communications by a PSO-inspired algorithm," Communications, LATINCOM '09. IEEE Latin-American Conference, Sept. 2009.

[16] Wen-Chung Liu "Design of a multiband CPW-fed monopole antenna using a particle swarm optimization approach", Antennas and Propagation, IEEE Transactions, vol.53, pp.3273-3283, 2005.

[17]PSO tutorial from www.google.com

[18] A Cognitive Approach to the Detection of Spectrum Holes in Wireless Networks Tobias Renk, Clemens Kloeck, and Friedrich K. Jondral

The other references which are not cited in the main text:

[19] Genetic Algorithms and Evolutionary Computation by Adam Marczyk Copyright © 2004

[20] Biologically inspired cognitive radio engine model utilizing distributed genetic algorithms for secure and robust wireless communications and networking-Christian James Rieser 2004 Doctoral Dissertation

[21] Genetic algorithm-based optimization for cognitive radio networks Published in: · Proceeding Sarnoff'10 Proceedings of the 33rd IEEE conference on Sarnoff IEEE Press Piscataway, NJ, USA ©2010

[22] J. Kennedy, R.C. Eberhart. "Particle Swarm Optimization", In Proc .of the IEEE Int Conf. Neural Networks, 1995.

[23] Bio-Inspired Algorithms for Dynamic Resource Allocation in Cognitive Wireless Networks T. Renk, C.Kloeck, D. Burgkhardt, F. K. Jondral, D. Grandblaise, S. Gault and J.-C. Dunat.

[24] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. IEEE International Conference on Neural Networks, vol. 4, (Perth, WA, Australia), pp. 1942–1948, 27 Nov.-1 Dec. 1995.

[25] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie, "Cognitive networks," in First IEEE International Symposium *on* New Frontiers in DySPAR.

SESSION

APPLICATIONS

Chair(s)

TBA

Novel Data Harvesting Scheme for Efficient Data Aggregation

Mohammad Zuheir Hourani, Nasir Hussain, Ridha OUNI
Department of Computer Engineering
College of Computer and Information Sciences, KSU
P.O.Box 51178, Riyadh 11543, KSA

***Abstract*—The basic idea behind intelligent transportation system is how to deploy vehicular sensor network that have many characteristic such as high computation power, enough storage space and mobile sensor node in order to design an effective and efficient architecture for data collection and data exchange. In this paper we will introduce an intelligent transportation system with new network paradigm to collect important information from the road environment based on the vehicular sensor network (VSN). Data aggregate provides the drivers by valuable information in order to make the road safer and less congested. Our system framework consists of active vehicular sensor node, passive vehicular sensor node and sink node distributed according to the road segmentation for collecting data from active vehicular sensor node passing by, while active vehicular sensor node collect data from passive vehicular sensor node in their segment using multihop data harvesting. Our scheme aims to reduce broadcast storm and avoid collision. Finally, the simulation using the OPNET simulator shows the effectiveness of the proposed schema.**

***Keyword*- ITS, VSN, IVS, data harvesting, hybrid architecture, data aggregation.**

I. INTRODUCTION

Significant advances in manufacturing technology equipment and the advent of Micro-Electro-Mechanical Switches (MEMS) has opened the way for the construction of intelligent sensor nodes which are able to perform three major functions: sensing, processing and wireless communication. These wireless sensor nodes are characterized by their intelligence, their small size, low cost, battery powered, and easy to install and repair. These features open doors to deploy WSNs in the future for a wide range of applications because it greatly expands our ability to monitor and control the physical environment from remote locations.

An interesting field where the use of WSNs proves effectiveness is the field of Intelligent Vehicular Systems. An Intelligent Vehicular System (IVS) uses technological advances in computers and information technology to improve the efficiency of both new and existing vehicular systems.

Vehicular sensor networks (VSNs) is a technology where sensors are deployed in the road side and in the vehicles to sense various urban phenomenon’s and transmit information for vehicular traffic control and monitoring. VSNs have different characteristic from traditional sensor network (static network), interns of mobility, computational, power supply, memory storage and reliability. Moreover vehicular sensor network VSN has a much more dynamic topology as compared to the static WSN. It is often assumed that VSN will move continuously in a random fashion, thus making the whole network a very dynamic topology. This dynamic nature of VSN is reflected in the choice of other characteristic properties, such as routing, MAC level protocols and physical hardware, beside this, dynamic topology of vehicular sensor network VSN, communication links can often become unreliable [1]. The previous characteristics allow deploying vehicular wireless sensor network to design intelligent transportation system.

In this paper we are interested to design an optimal system architecture for such vehicular sensor network for vehicular traffic control and monitoring. several assumption have been made. First, we assume that vehicles communicate through a wireless interface, implementing a CSMA/CA MAC layer protocol that provides a RTS/CTS/DATA/ACK handshake sequence for each transmission. Vehicular sensor network adopt IEEE 802.11 as a cost efficient and widely deployed solution for network communication. IEEE 802.11p is a draft amendment to IEEE 802.11 standard to add wireless access in vehicular environment. It supports data exchange for vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) in the licensed ITS band of 5.9 GHz.

The number of sink nodes that are distributed beside the road is very small in proportion to the number of vehicular sensor nodes. So, we assume that sink node has a relatively fast processor and a large storage device and has enough energy resources. In addition, it has very large data base to store information from the vehicles. However, vehicular sensor node has lower storage space and low processor capability. It is assumed that each vehicle has unique ID to identify the vehicular node and included in each message sent.

Finally, we assume devices participating in vehicular networks are highly mobile with a speed up to 30 m/s. But, their mobility patterns are predictable due to the constrained movement imposed by the road system and constrained speed imposed by speed limits, traffic conditions and signals. In fact, the mobility of vehicular sensors poses challenges to the communication system. Mobility undermines the reliability of communication and also causes the topology to continuously change.

This paper focus on how to deploy WSN as an intelligent transportation system over effective and efficient architecture for data collection and data exchange allowing vehicle traffic monitoring and control. The rest of the paper is organized as follow. Section II provides a back ground and related work in vehicular sensor networks. Section V introduces the proposed scheme for data harvesting and dissemination. Section VI evaluates the simulation results conducted with OPNET. Finally, section VII concludes the paper.

II. RELATED WORK

Recently, there is a strong interest from researcher's in deploying WSNs in VSNs in many applications that involve constraints related to the traffic conditions such as traffic monitoring and control, traffic estimation and monitoring parking. Some research focus in moving vehicles to enable wireless sensor communication between roadside and vehicle or between vehicles. These applications aim to make roads safer and less congested in order to save the time for people. It's important to note that these applications encounter three types of communications [2]:

- Vehicle-to-Vehicle (V2V) communication: vehicles are equipped with sensors in order to exchange information that is crucial to avoid severe situations like traffic jam avoidance.
- Vehicle-to-Infrastructure (V2I) communications: information flow from vehicles to sensors installed on roadway infrastructure
- Hybrid communication: uses both V2V and V2I architecture

In [3] the author proposed a scheme based on the hybrid communication. Vehicles will send all their sensed data to infostations, where the data will be forwarded to corresponding infostation based on the infostations management area. Later, any vehicle requesting sensed data can request these infostation, which is more of an indirect form of vehicle to vehicle communication using relay nodes forming another type of data harvesting protocol. However, this technique requires installation of an infostation infrastructure, which can be very costly and complex.

With the use of wireless sensor network, multiple sensory devices can be networked together to share geographically distributed information. In [4], the system consists of two vehicle sensory nodes that are placed on each side of a two-lane road. These two wireless sensory nodes will send the collected data to the base node whenever a vehicle is detected. The road traffic monitoring system consists of a 3-tier structure. The system is made up of the end-node tier, base node tier and lastly the PC tier. All data that is captured by the end-nodes will be forwarded to the base node. The base node will then perform pre-processing before forwarding the message to the PC for analysis. Communication via the PC to the end nodes is also carried out through the services of the base node using a star topology. The end-node tier is responsible for vehicle detection and gathering all the data from its onboard sensors.

One of the biggest issues in realizing VSN is concerned with data harvesting which is a technique where sensors create data that summarize the characteristic of the data and send it to the target. In [5], the author proposed a novel multi-hop data harvesting (MDH) method for the V2I architecture. MDH have two scheme proposed for VSN. The MDH scheme using replicas (MDH-R) is proposed for requesting data from single sensor node, while data aggregation scheme is designed (MDH-RA) for cases when the request was made to a geo cast region. Many applications in VSN may require multi-hop data transmission to meet real-time constraints. The author see multi-hop data dissemination capabilities may become ideal for future researches in this area.

VSNs come out as new brand of vehicular networks, whose propose is the real time gathering and diffusion of information. In [2] the author used a Clustering Gathering Protocol (CGP) that is across layered protocol based on hierarchal and geographical data gathering, aggregation and dissemination. The goal of CGP is to gather from all node in the vehicular ad hoc network in order to offer different kind ITS services, it allows telecommunication/service providers to get valuable information about the road environment in a specific geographical area, using V2V network to minimize the high cost links usability and base station to gather information from the vehicles

III. OVER VIEW

Our system framework is consisting of static road side node (sink node), and mobile vehicular sensors. Road side nodes are distributed according to the road segmentation for collecting data from mobile vehicular sensors passing by and to exchange data about traffic condition. While mobile sensors on vehicles monitor the road condition and send this information to active mobile neighbors when they are close enough then to the road side sink node (see Figure 1). We focus on vehicular mobility, collaboration between mobile and static nodes, and information exchange among mobile cars. Mobile cars can gather latest information spreading on the map out of the reach of static node, whereas static node can gather information from more active cars coming across, where the connectivity between static and mobile nodes and also between mobile and mobile nodes are most likely meaningful and useful.

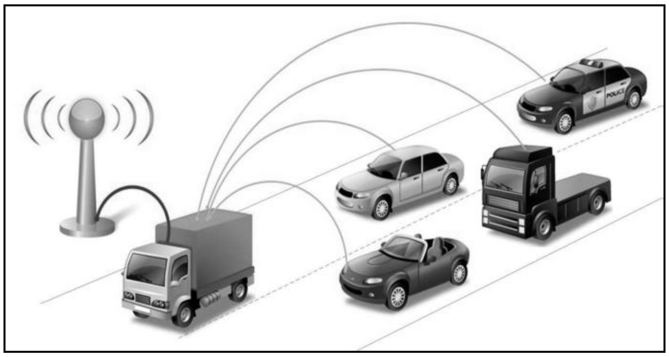


Figure 1: Vehicular Sensor Architecture

IV. ROAD SEGMENTATION

The roads are divided into small segments. On each road segment, there are two road side node (sink node) located at the both ends of the segment, as shown in Figure.2. Usually, the road side nodes are placed on the roadside with different distances $(d + i)$ based on the road environment to collect data from active cars passing by. So, cars can get the road condition before entering this segment; while vehicular mobile sensors, assisted by the mobility of the vehicles, can know the road information along their own path.

The road is divided into S virtual segments with the different length (Figure 2). In each segment an active node is selected to gather data from all nodes in its segment, aggregate them, and send the result to the sink node.

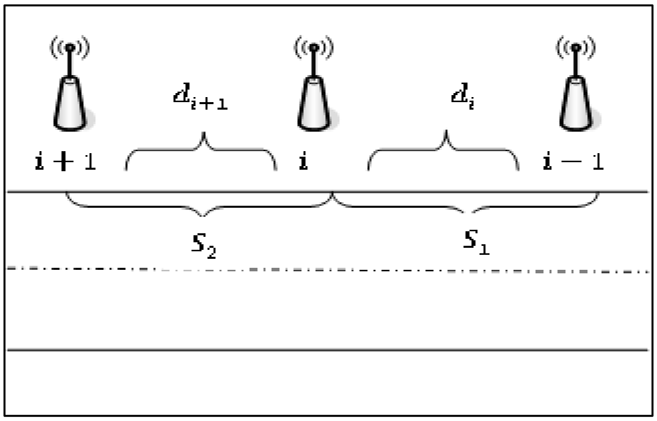


Figure 2: Road Segmentation

V. PROPOSED SCHEME

The proposed scheme consists of providing a feasible, efficient and robust vehicular sensor network framework to monitor road traffic and provide desired and reliable information for users, particularly for drivers in automobiles. In the context we decided to use active node based solution for the V2V dissemination. The scheme will be divided into three parts: Active vehicular sensor node selection phase, data harvesting/dissemination phase, and the data sharing phase.

ACTIVE VEHICULAR SENSOR NODE SELECTION PHASE

Every road segment has two such sink near the ends. Every vehicle enter the segment will send hello message to the sink node at the beginning of each segment containing the vehicle ID. Then, the sink node will store this information in the data base. Using this information, the sink node will create an active node based on two parameters as threshold; the maximum number of vehicles (passive nodes) detected in the road, and the elapsed time.

First, the sink node stores the data about each vehicle entering the segment until reaching the maximum number of vehicles. Second, the sink node will choose one of these car's randomly to be an active node by broadcasting control information as request including the ID of the vehicle. When collecting this request, the other vehicle nodes (passive nodes) identify the target node dedicated for forwarding their information. All other nodes must know the active node in their segment. To do so, the AVS will include its ID in the packet as new destination and then diffuse reply to the sink node which will be also received and processed by its neighbored vehicles.

The mobility of vehicular sensor network can affect the topology of the network. Therefore, we also use the elapsed delay to control when the sink sends request to create an active node exactly before the group of vehicle leave the wireless range of the sink. This time will be calculated using the equation below.

$$t = \frac{d}{v}$$

Where d is the distance that our wireless communication can support (IEEE 802.11), and v reflects the mobility of the vehicles which is the velocity of the vehicular sensor node. As a result, we need two counters in the sink node one for time and the other for the number of vehicles. So, in this way we guarantee that we create an active node for each group of vehicles.

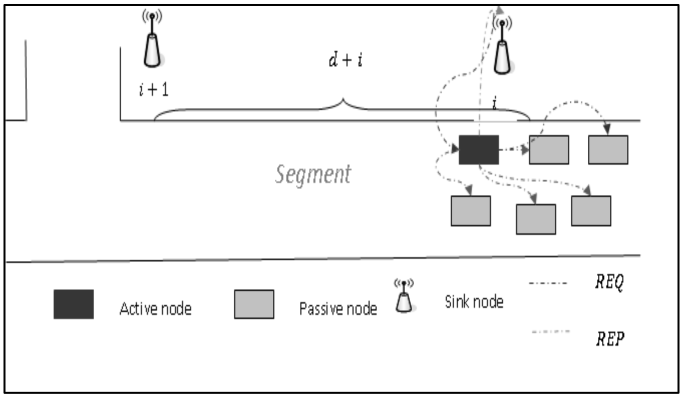


Figure 3: Active Node Selection

DATA HARVESTING AND DISSEMINATION PHASE

A road segment can be congested, or free. When it is congested, it can be either heavily congested or lightly congested. Moreover, the traffic condition is changing with time passing. So, a simple but effective method is needed to represent the road condition. For a better delivery ratio and to reduce broadcast storms, a message has to be relayed by a minimum number of intermediate active nodes to the destination. To do so, nodes are organized on a set of segments, in which one node or more (active node) gathers data in its segment and later sends this data to the sink node. Segment-based active node solutions provide less propagation delay and high delivery ratio with also bandwidth fairness. In [4] the authors use a distributed clustering algorithm to create a virtual backbone that allows only some nodes to broadcast messages and thus, to reduce significantly broadcast storms.

When the active node receives the data from the passive node originally holding the data, it will process it in store & forward fashion instead of sending directly to the sink node when the next sink node is far. Similarly, the passive node keeps the data in its memory during a parametric time, and waits for active node to be closest enough from it. An example is given in figure 4, where the node (F) cannot reach the active node. In this case, it will store its data till it reaches the active node or wait another active node.

All nodes in the segment unicast their sensed data to the active node, using a mechanism similar to DCF (Distributed Coordination Function, presented in IEEE 802.11).

- Each node wait a random bounded back-off time,
- At the end of the back-off time, a node send a request to send to the active node,
- The active node acknowledges the reception by sending a Clear to Send message,
- The node sends its data to the active node.

An example is shown in figure 4.

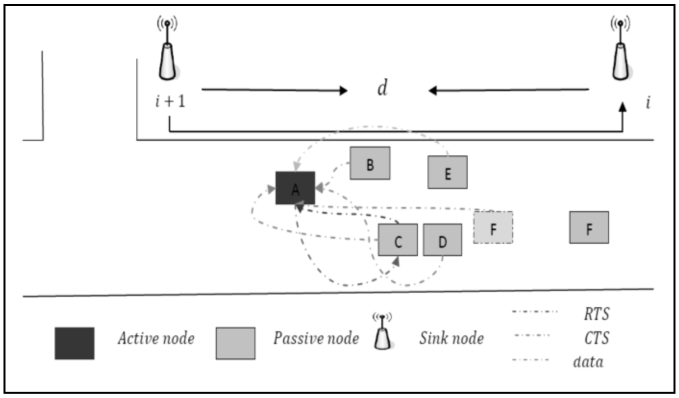


Figure 4: Data Harvesting Model

Passive vehicular sensor nodes will monitor the traffic condition by measuring the speed of the vehicle and send this data to active node in its segment. The active vehicular sensor node will store the data message and count the number of vehicles in its segment, because it has limit number of vehicles. This data will be forwarded to sink node $(i + 1)$ that is located at the end of this segment. In turn, the sink node $(i + 1)$ will send this data to the preceding sink node (i) that to update its data about traffic condition in this segment and floods it to new coming vehicles that wish to enter this segment.

When the vehicle is going to enter a new road segment, the sink node at the near end will communicate with this vehicle. So, the vehicle can know the road condition of this segment in advance. There is no needs to place more sink node in the middle of one segment, because even if the vehicle get information at the middle of a segment, drivers still cannot change their direction or change the route trip.

DATA SHARING PHASE

In some areas, data traffic may increase dramatically due to many vehicles requesting for data at the same time. In this case, there is a high probability that more than one vehicle is requested to be an active vehicular node from the sink node i. When the active node reach its limit from the passive node due to the congestion that involve many vehicle in the segment, in this situation the sink node send request to create new active node to provides fairness which is very important in a sensor network where every node has to send its data. It also reduces significantly broadcast storms and thus avoids collisions. another case that we can have new active node when we have two groups of cars and there is a time between them and one of them reach the end of wireless communication range of sink i but the maximum number of passive node still not complete. In this case, the sink node will request new active node from the coming group based on time factor in order to ensure that each group have active node to send information to it. It can be seen from figure 5, when there are many vehicles have data to be sent about the traffic condition which means congestion occurs on a specific segment. In this case, the active node sends to the sink node *i* message in order to create new active node.

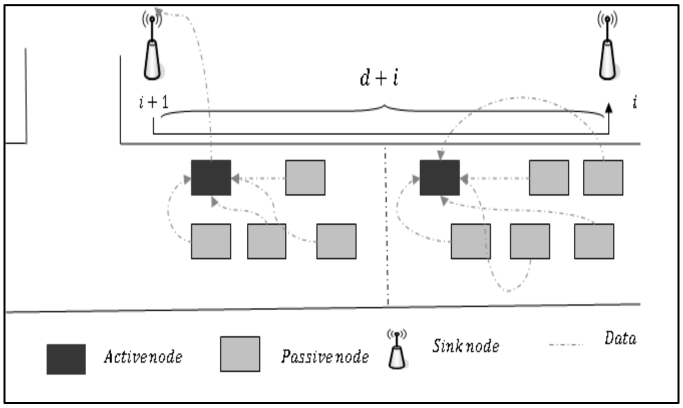


Figure 5: Sharing Phase

VI. SIMULATION

The network topology was built in OPNET Modeler with the following design considerations. There are three scenarios in which we compare the traffic behavior based on congestion and active node availability. For the first scenario, we have one active node sidelined with 30 passive nodes and the second scenario deals with congestion by increasing the traffic generation by dramatically doubling the passive nodes. Finally, the third scenario involves a mitigation step from the sink by requesting an additional active node to accommodate the situation for this congestion.

The environment for the research is set as, CSMA-CA with minimum back-off exponent set to 3 and having 4 as maximum number of back-off's with channel sensed every 0.1 seconds and operating in 2.4Ghz frequency band and the transmit power is set to 0.002, with ACK wait duration set to 0.05 seconds for the participating active and passive nodes.

First of all the traffic received at the active agent is of utmost importance to study the performance of the system. In figure 6 the traffic received is indicated as in the three scenarios. When there is no congestion in the network the average traffic received converges to 4300 b/s which decreases to 3500 b/s as soon as there is congestion in the network due to extra management and control information besides data traffic and finally having the additional active node will help in having more data traffic. Secondly, we will consider the delay incurred for transmission in the three scenarios, average values are considered to study the effect of congestion on the network. In this first scenario where we have just one active node the average delay is around 0.013 seconds where as it rises to 0.014 seconds when congestion occurs with the inclusion of additional passive nodes in the network.

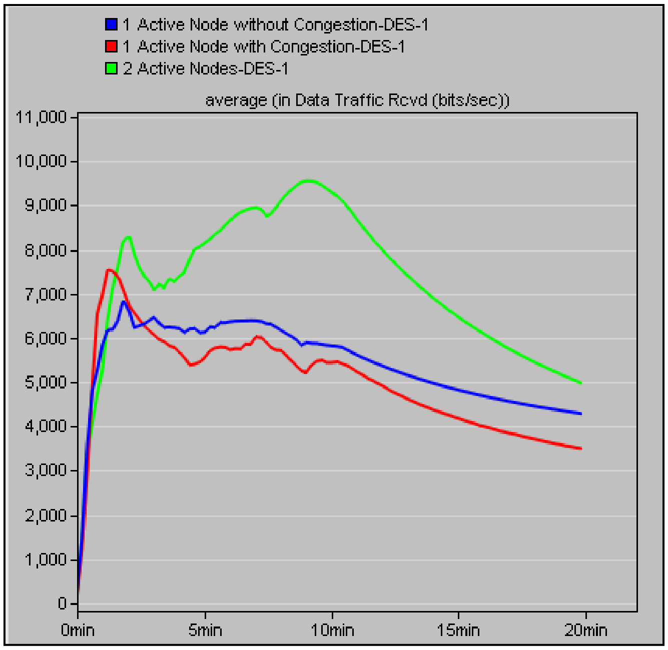


Figure 6: Data traffic received

After this the sink requests for an additional active node to corner out the increase in the delay and with the inclusion of this active node the load is distributed between the two active nodes hence decreasing the delay incurring at each node which is illustrated in figure 7.

Finally, the delay to access media is shown in figure 8 which reveals the fact that without congestion in the network the channel was accessed on an average of 0.0014 seconds but with the inclusion of congestion the delay rises to a value of 0.021 seconds approx. and with the second active node sharing the load reduces this value to 0.018 seconds which is in the desirable range.

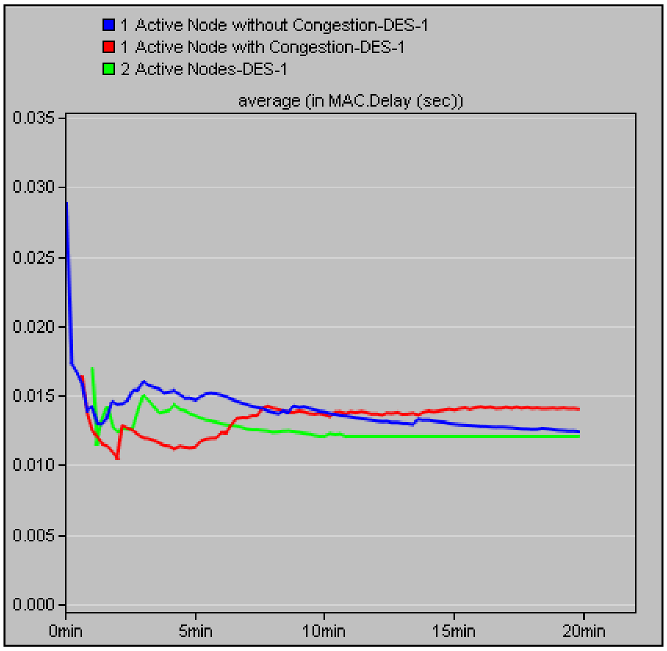


Figure 7: MAC delay

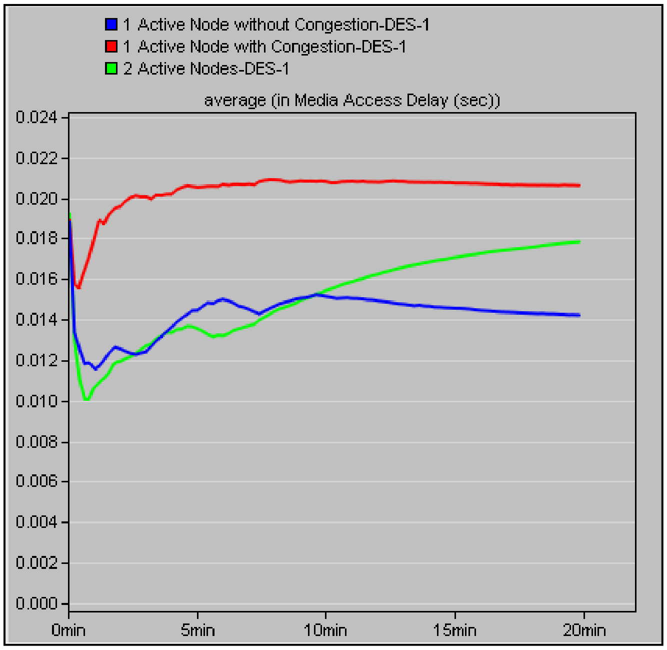


Figure 8: Media access delay

So, in all we can say that including extra active node is only required in scenarios where the congestion in the network increases and this will distribute the load among the participating active nodes which are used for harvesting of the information to the sink.

VII. CONCLUSION

In this paper, a scheme for data harvesting and data exchange based on active vehicular sensor node is proposed. We provide a collaborative hybrid method to deliver important information to particular drivers effectively. We use road side sink and vehicular sensor nodes to restore and exchange data, then we use OPNET simulator to study our novel scheme which illustrates that during the time of congestion in the network, it is better to have additional active node beside the old one and have many advantage as we see from the result.

ACKNOWLEDGMENT

The authors are very grateful to the Research Centre of the College of Computer and Information Sciences – King Saud University. For their continuous help and support in the development of this research.

REFERENCES

[1]. Munir, S.A.; Biao Ren; Weiwei Jiao; Bin Wang; Dongliang Xie; Man Ma , ”*Mobile Wireless Sensor Network: Architecture and Enabling Technologies for Ubiquitous Computing*”, Advanced, pp:113 – 120, 08 August 2007.

[2]. I. Salhi, M. O. Cherif, S. M. Senouci, “*A New Architecture for Data Collection in Vehicular Networks*”, IEEE International Conference on Communications ICC'09, pp.1–6, August 2009.

[3]. U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, A.Corradi. “*Efficient data harvesting in mobile sensor platforms*”, Proceedings of the 4th annual IEEE International Conference on Pervasive Computing and Communications Workshops PERCOMW’06, pp. 352, 2006.

[4]. H. Ng, S.L. Tan, J. G. Guzman, “*Road traffic monitoring using a wireless vehicle sensor network*”, International Symposium on Intelligent Signal Processing and Communications Systems, ISPACS’08. pp. 1–4, Feb. 2009.

[5]. K.W. Lim Y.-B. Ko, “*Multi-hop data harvesting in vehicular sensor networks*”, Communications, IET, Vol.4, no7, pp.768–775, April 2010.

Non Iterative Decoding of Low Density Parity Check Codes Using Artificial Neural Networks

D. Blackmer[1], F. Schwaner[1], and A. Abedi[1]
[1]Department of Electrical Engineering, University of Maine, Orono, ME, 04473 USA

Abstract[1]— *The decoding of Low Density Parity Check (LDPC) codes through the iterative process of belief propagation presents practical challenges for designers looking for real time performance in communication systems. Due to the geometric and finite pattern nature endemic in the construction of LDPC codes, this paper proposes the use of Multi Layer Perception (MLP) feed forward artificial neural networks to replace belief propagation to achieve constant decoding times while retaining performance levels comparable to more traditional decoding methods. Due to the back propagation training method used for neural networks, and the requirement of showing the network every possible input output sequence it will ever see, this paper also presents a novel method of approaching long block length codes far larger than is otherwise possible to train neural networks for with modern computer hardware.*

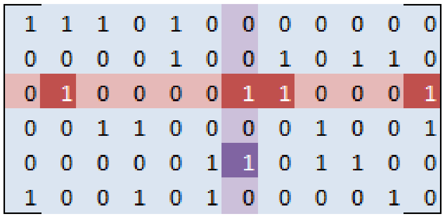
Keywords – Neural Networks, Low Density Parity Check, Decoding algorithms, channel coding, communication theory

1. INTRODUCTION

In the field of designing error correcting codes, two types of codes stand out as being able to achieve performance rates near the theoretical upper channel capacity limit laid out in the Shannon theorem [1]. In the field of convolution codes, turbo codes with proper iterative decoding based on belief propagation can come within fractions of a dB of the limit. [2] And in the field of linear block codes Low Density Parity Check (LDPC) codes with sufficient block length also can achieve comparable performance arbitrarily close to this theoretical upper limit on binary symmetric channels [3]. The limitation of both of these types of codes is that the decoding is an iterative process to which performance is frequently tied to the number of iterations in the decoding process. With sufficient iterations, the decoding algorithm can converge with maximum likelihood to what the original transmitted data was. The general rule tends to be, the more iterations, the closer to the Shannon limit the performance of each type of code can achieve. For real time communications this poses a problem, because for many applications users will not tolerate the latency required to iterate through a long code, and from a engineering point of view a powerful floating point processor with sufficient memory must be dedicated to the decoding process of the Sum Product Algorithm (SPA), where the computational complexity is a linear function of the number of 1's in the parity check matrix *H* [4]. Some research has been done with limited precision SPA, but the precision drops off for greater and greater quantization error [5]. For low signal powers associated with mobile devices which must contend with battery storage capacity and maximum efficiency requires that every clock cycle must be used to complete as much work as possible. A new method of efficiently decoding these signals at high, constant speeds has been proposed. The idea of using Artificial Neural Networks (ANN) for the purpose of decoding LDPC codes by itself is not a new idea, however one of the inherent limitations has always been the length of the codes being constrained by the length that is reasonable to train [6]. Using the inherent pattern recognition and generalization abilities of a properly trained neural network can enable constant time very high speed, non iterative LDPC decoding, with error performance levels on short codes approaching or even surpassing more traditional iterative belief propagation decoding methods.

2. LDPC A BRIEF REVIEW

To understand how errors are corrected with LDPC, first look at the highlighted row in *Figure 1*. As with all other rows in this particular parity check matrix there are exactly 4 bits. These four bits for every valid codeword will have known parity.



$$\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Figure 1: H Matrix for (12,6) LDPC code

$$mod_2\{X_2 + X_7 + X_8 + X_{12}\} = 0 \qquad (1)$$

As demonstrated in (1) the parity check matrix

[1] This work was sponsored by NSF IGERT Grant Number DGE-0504494 and NASA Grant Number MSGC-EP-11-05.

guarantees that the indexed values from the rows of the received codeword specified by the ones will have known parity. This is at the core of what the LDPC does to provide the decoder with a-priori information to allow error correction. To further this process and to allow for the exchange of information between rows, the highlighted column of *Figure 1* demonstrates that the two rows share one common check value. Using a tanner graph one can verify that the smallest cycle in this *H* matrix is of length 3. The objective is that each row has no more than one bit in common with any other row to reduce short cycles in the connectedness graph. The example in *Figure 1* only has two ones per column; however there is no specific number of 1's per row or column which makes a good code. Irregular LDPC codes have been shown to have better performance in certain situations [7]. The neural network approach proposed by this research allows us to deal with regular or irregular codes; however a new network must be trained for each new row with a different number of ones.

As the proposed method discussed here breaks apart the H space into individual rows, where each row has $\boldsymbol{n}$ ones**.** Now since each row must have even parity that means there are $\mathbf{2^{n-1}}$ possible permutations of these values. Thus in the case of a row with $\boldsymbol{n = 4}$ there are $\mathbf{2^3 = 8}$ valid codewords. Each sequence of which is separated by at least $d_{min} = 2$ values. Since the neural network only has to be trained with $\mathbf{2^{n-1}}$ sequences it simplifies and expedites the training process. But more than simplifying it, this approach makes the training possible. It can be easily seen that with modern computer memory limitations it would be impossible to train a network with a data of binary length 100 bits, since it would have $\mathbf{2^{100}}$ permutations. Even in binary form this would require $\mathbf{5.07 * 10^{21}}$ GB of ram to store all the permutations of the training vector. And due to the so called curse of dimensionality, require such a large network as to be completely impossible to train or operate with any modern systems.

3. MULTI LAYER PERCEPTIONS

Artificial Neural Networks (ANN) of the Multi Layer Perception (MLP) are a class of feed forward neural networks, meaning they have no recursive or feedback connections. As shown in Figure 2, they are constructed by interconnecting multiple summing blocks which each sum together some scaled version of the input. For the application of pattern recognition, the results of each summation gets multiplied by a nonlinear sigmoid activation function then feed to the next layer of perceptions where the same set of operations gets repeated. After this has propagated through several layers as shown in Figure 2, the signal reaches a probability based competitive layer which makes a decision about the likelihood that the input vector belongs to each possible class of valid outputs.

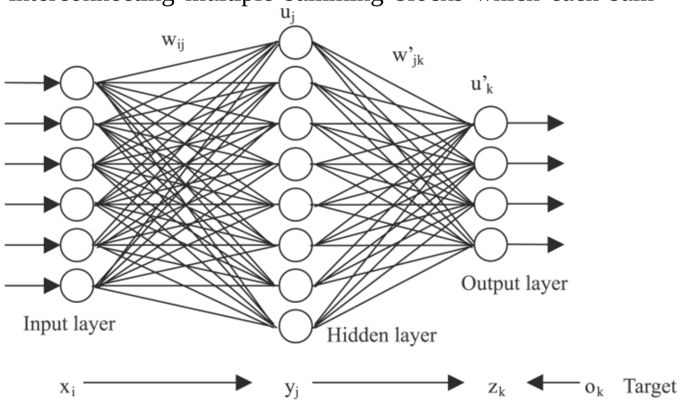


Figure 2: Neural Network Structure

These feedforward networks are ideally suited for pattern recognition [8], and have several major benefits which make them a great choice decoding signals. Their ability to properly classify inputs when presented with novel signal data means that even when corrupted with random noise, the neural networks can be trained to look past the signal errors and noise and find the underlying geometric relationship that defines the coded signal. The method of training the network, rather than having a static design provides another benefit for the purposes of codeword recognition. Being shown perfect versions of the signal, then having a gradient descent algorithm update each interconnecting weight in an attempt to iteratively find global function minima in the output space. Then being shown corrupted versions to allow the network to generalize itself to novel inputs better. This method allows the training to be done offline, iteratively approaching successively better and better network weight and bias configurations for network performance. Then when the network is online, the performance will be tied to the performance of the trained network. However MLP networks, unlike other algorithms used for pattern recognition, require no recursion or iteration to achieve similar performance levels online as other algorithms achieve. [9]

4. FPGA NEURAL NETWORK PATTERN RECOGNITION

One major benefit of this approach can be seen in the lack of precision required by neural networks to achieve good results. The exact mathematical precision for the sigmoid activation functions is not nearly as important as its shape, so fixed point lookup tables can be used to perform what could otherwise be a computationally complex nonlinear transformation [10]. It has been shown that when using 16-bit fixed point math with VHDL synthesis tools for FPGA's using the *uog_fixed_arith* library that performance of neural networks can be increased by 12x, and exhibit quantization error of only $2.411x10\ ^{-4}$ for the bipolar range of [-8 8]. [11] Also due to the manner in which the network is trained instead of programmed, there are many possible convergent points which will produce good results. This means that the network can be trained using fixed point weights which consist only of powers of two and sums of powers of two. This means that all the multiplications can be done with shift registers and additions rather than requiring dedicated multiplication hardware [12]. This results in a network which can be implemented in a FPGA in a massively parallel fashion taking up no extra clock cycles for a CPU to accomplish near real-time decoding. This highlights the true benefit of the neural network approach, the ability to do the training offline, then to implement the trained network as dedicated logic registers. The referenced papers and texts provide proof for the reader that neural networks can be implemented effectively in this way. It is outside of the scope of this paper to implement and demonstrate the specific performance of this method, however the reader should be aware that through the work of others achieving this extreme performance is possible using the neural network approach.

5. Implementation

Therefore this paper presents two different approaches to using neural networks for decoding. The first approach is for illustrative purposes and has been investigated before [13] for linear block codes. This technique is to train the network with all possible valid codewords. Being that this approach will only work for short codes, a second approach is proposed by this paper. The idea of training the network using the row based parity sequences endemic of valid LDPC codes, then allowing the belief about the state of certain bits to propagate through the network.

The procedure for the first process is:

1) *Offline: With a linear block code of M data bits and N parity bits. Train a neural network where the input is a* $(2^M \times (M+N))$ *channel output matrix, and the output is a* $(2^M \times M)$ *matrix of properly decoded values.*

2) *Online: Feed the network input with each of the received noise corrupted vectors from the channel, and the maximum likelihood decoded sequence will be produced on the output.*

There are two drawbacks to this design. The first being the training being prohibitive for long sequences as discussed before. The second is referred to as the curse of dimensionality. This basically means that the number of multiplications and additions performed by a feed forward network is given by:

$$(M + N + 2^M) * H$$

M = Data Bits
N = Parity Bits
H = Number of Hidden Nodes

As the number of data bits grows, the number of hidden nodes must grow proportionally for the decoding performance to stay constant. Accordingly the number of multiplications can grow to the point of inefficiency quite quickly.

As a hypothesized alternative which removes these limitations, this paper proposes the following alternative process:

1) *Offline: Determine the number of ones* X_i *in each of the i rows of the H matrix.*

2) *Train individual networks for each unique* X_i*.*
Ex: if a LDPC has 6 rows with 5 ones each and 2 rows with 4 ones each, two unique networks, one for 5 ones and one for 4 ones are necessary

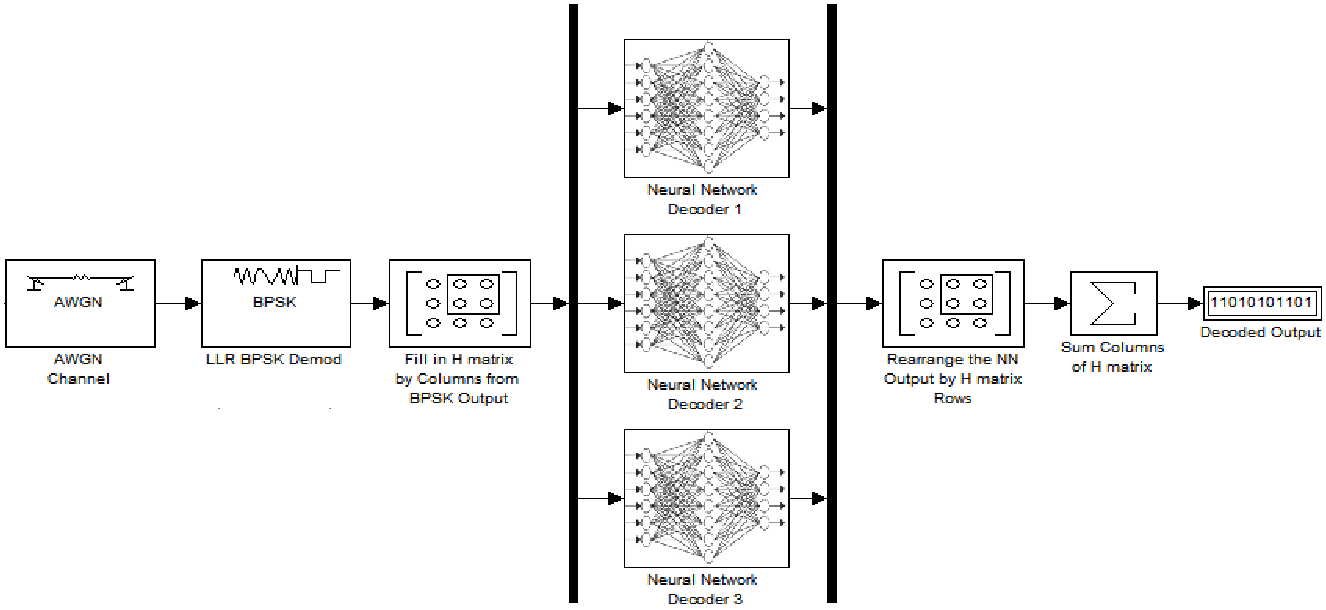


Figure 3: Full Proposed Decoder Arrangement

3) Online: Encode and transmit the LDPC vector through the channel as normal. On the receiving end rearrange the received vector from the channel into the shape of the encoded H space.

4) Feed each row of the received vector through the correctly sized neural network.

5) The output Y will now be the size $\mathbf{1}$ x $\mathbf{2^{X_i-1}}$ vector of likelihoods that the given input sequence belongs to each of the $\mathbf{2^{X_i-1}}$ possible sequences. Multiply the $\mathbf{1}$ x $\mathbf{2^{X_i-1}}$ likelihood output sequence by the $\mathbf{2^{X_i-1}}$ x $\mathbf{X_i}$ matrix of all possible valid sequences. This will generate the probabilistic $\mathbf{1}$ x $\mathbf{X_i}$ values to fill back into the X_i positions from the current row of the H matrix.

6) Sum each column to update the likelihood of each bit with the knowledge passed from each other common column bit.

The arrangement of these steps is shown in Figure *3*.

6. Results

In *Figure 4* it is shown that several different neural network structures haven nearly indistinguishable performance to the Sum Product Algorithm. Once the network can be shown every possible input / output sequence, the decoding performance will achieve maximum likelihood. However because of the limitations discussed earlier this approach only works for very short codes.

Figure 5 shows the performance of the proposed alternative method. While it can be seen that the BER performance isn't up to par with the SPA approach, this demonstrates neural networks can decode the DVB-S2 standard (64,800, 32,400) LDPC code, which would as discussed earlier be completely impossible with modern hardware without using this approach.

7. Conclusion

Despite the lack of complimentary BER performance for the method this paper has proposed compared to SPA decoding, the concept that a neural network can be used to decode a much longer code by identifying substructures which can be individually and independently decoded has been verified. Performance levels could be improved if the identified substructures were uncorrelated, since it is assumed that it's the correlation between various sub elements which prevents the performance from improving any more with a greater number of ones in each column. Further research is needed to improve the performance levels of this method. This paper has presented a starting framework from which to build a better performing neural network decoder which can approach substantially longer LDPC codes than was previously possible.

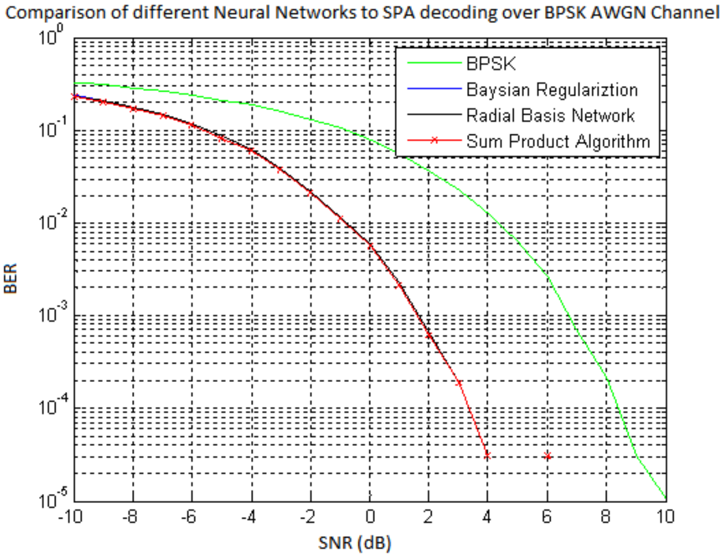


Figure 4: Comparison of Several Neural Network Types vs. SPA for a short (6,2) LDPC code

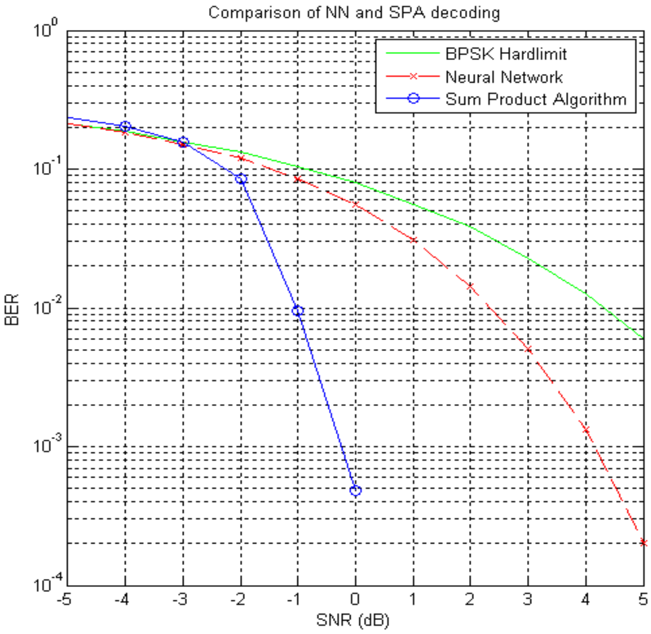


Figure 5: DVB-S2 (64,800, 32,400) LDPC Code with both methods

8. References

[1] Sae-Young Chung, G. David Forney, Jr., Thomas J. Richardson, and Rüdiger Urbanke, "*On the Design of Low-Density Parity-Check Codes within 0.0045 dB of the Shannon Limit*", IEEE Communication Letters, vol. 5, pp.58-60, Feb. 2001. ISSN 1089-7798.

[2] Berrou, C "Near Shannon limit error-correcting coding and decoding: Turbo-codes" Communications, 1993. ICC 93. Geneva. Technical Program, Conference Record,

IEEE International Conference on Volume 2, 23-26 May 1993 Page(s):1064 - 1070 vol.2.

[3] Gallager, R. G., “ Low Density Parity Check Codes”, Monograph, M.I.T. Press, 1963.

[4] Lun, Shu.; Costello, Daniel J.; *"Error Control Coding 2nd edition"*, Prentice Hall, 2004.

[5] Moberly, Raymond; O'Sullivan, Michael; Waheed, Khurram; *"LDPC Decoder with a Limited-Precision FPGA-based Floating-Point Multiplication Coprocessor"*. Proceedings of SPIE, the International Society for Optical Engineering. ISSN 0277-786X

[6] Krose, Ben.; Smagt, Patrick van der; *"An Introduction to Neural Networks"*, University of Amsterdam: The Netherlands 1996

[7] MacKay, David J. C.; Wilson, Simon T.; Davey, Mathew C.; *"Comparison of Construction of Irregular Gallager Codes"*. IEEE Transaction on Communications, July 1998

[8] Campos, P.G, “MLP networks for classification and prediction with rule extraction mechanism”, Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, Volume 2, 25-29 July 2004 Page(s):1387 - 1392 vol.2

[9] Benediktsson, Jon A.; Swain, Philip H.; Ersoy, Okan K.; *"Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data"*, IEEE Transactions on Geoscience and Remote Sensing, Vol 28. Nov 1991.

[10] Zhu, Jihan; Sutton, Peter; *"FPGA Implementation of Neural Networks - a Survey of a Decade of Progress"*, In Proc. 13th Ann. Conf. on Field Programmable Logic and Applications. pp. 1062-1066.

[11] Omondi, Amos R.; Rajapakse, Jagath C. *"FPGA Implementations of Neural Networks"*, The Netherlands: Springer, Page 46, 2006.

[12] Marchesi, M.; Orlandi, G.; Piazza, F.; Uncini, A. "*Fast Neural Networks Without Multipliers"*, Nerual Networks, IEEE Transactions on, 1993 , Page(s): 53 - 62

[13] Ja-LingWu.; Tseng, Yuen-Hsing.; Huang, Yuh-Ming; *"Neural Network Decoders for Linear Block Codes"*. International Journal of Computational Engineering, 2002.

Non-intrusive Movement Detection in CARA Pervasive Healthcare Application

Bingchuan Yuan
Department of Computer Science
University College Cork
Cork, Ireland

John Herbert
Department of Computer Science
University College Cork
Cork, Ireland

Abstract—**Pervasive healthcare promises to support the desire of many elderly for independent living and at-home care. This paper presents the CARA (Context Aware Real-time Assistant) system whose design goal is to provide a pervasive real-time intelligent at-home healthcare solution. While this is the goal it is recognized that current practice and current user requirements of both the subject (such as an elderly person) and caregiver (such as a medical consultant) may differ from our ultimate goal. Recognizing this, we have built a solution that supports scenarios of use of CARA other than as a fully automatic pervasive healthcare solution. In this paper we describe a scenario where the full wireless medical BAN (body area network) is used remotely under supervision, and supplementary continuous monitoring of the patient is done in a non-intrusive way via a movement sensor. This movement monitoring is integrated with the CARA system and can make use of CARA's intelligent analysis as well as its recording and playback facilities. While this scenario does not use the full capabilities of the CARA system, it provides a less disruptive introduction to the technology for an elderly person, and leads easily to incremental incorporation of the technology. The full CARA system has sensors to continuously measure physiological signals, and it can either store the data on a server or stream the data to a remote location in real-time. Implementation as a rich internet application means that it is available to a remote caregiver just using any web browser that has the standard Adobe Flash plug-in. Thus the CARA system can be accessed on any internet-connected PC or appropriate smart device. The results of experiments are presented that illustrate the effectiveness of the system in analyzing a patient's movement, and its performance over the internet.**

Keywords- Pervasive Healthcare, Data Review, Remote Monitoring, Movement Monitoring, Rich Internet Application

I. Introduction

Healthcare systems are challenged by a rapidly growing aging population and rising expenditure. According to the U.S. Census Bureau [1], the number of people over the age of 65 is expected to hit 70 million by 2030, having doubled since 2000. Healthcare expenditures in the United States were projected to rise to 15.9% of the GDP ($2.6 trillion) by 2010. This challenge calls for a major shift of healthcare from a traditional clinical setting to an at-home patient setting, which can reduce healthcare expenses through more efficient use of clinical resources, earlier detection of medical conditions and proactive management of wellness.

One proposed solution to the current crisis is pervasive healthcare[2]. The wide scale deployment of wireless networks will improve communication among patients, physicians, and other healthcare workers as well as enabling the delivery of accurate medical information anytime anywhere, thereby reducing errors and improving access. At the same time, advances in wireless devices such as intelligent mobile devices and wearable sensors have made possible a wide range of efficient and powerful medical applications. Our solution is based on the CARA pervasive healthcare architecture. The architecture supports various scenarios of use that are discussed later. The main components of the CARA system are:

1. Wearable Wireless Sensors.

A key component of the system is a BAN (Body Area Network, i.e. a portable electronic device capable of monitoring and communicating patient vital signs), and this includes medical sensors such as the ECG, SpO2 meter, and temperature sensor. The BAN plays a central role in health monitoring and in the emergency detection functionality of the system.

2. Non-intrusive Movement Sensor.

A smartphone with a gyroscope and accelerometer is used as a non-intrusive alternative to the BAN for movement monitoring. This can provide valuable information regarding an individual's functional ability and general level of daily living activities.

3. Remote Monitoring System.

This is responsible for remotely controlling the BAN and continuously measuring physiological signals of the elderly through the BAN and internet connection. A web camera is integrated to this application that may be used for monitoring and for interaction between the elderly and the caregiver. Furthermore, real-time data obtained from the BAN are recorded on the server for further reviewing and analyzing.

4. Data & Video Review System.

This is designed for the medical consultant or caregiver to review the data previously collected from the elderly in case s/he might be not available for real-time monitoring. This application not only can present the recorded data in graphical form but also allows the consultant to play the recorded video of the monitoring session along with synchronized playback of the associated real-time sensor data.

5. Healthcare Reasoning System.

This is implemented by a Windows Workflow Rule Engine, and applies simple medical rules, appropriate for the individual, to real-time data that is received from the vital sign sensors.

II. Incremental use of CARA

The CARA system can be used in different ways, varying from fully automatic real-time at-home monitoring of patient vital signs resulting in automated response, to use as a non-automatic assistant for remote real-time consultation by a specialist. While the fully automated system is the ultimate goal, it is recognized this may be too disruptive initially for both patient and caregiver. Healthcare is about more than just immediate application of the most advanced technology. Recognizing this, we have built a solution that focuses on an incremental introduction of CARA as a pervasive healthcare solution. This solution supports a scenario where the wireless BAN is used under supervision (using a two-way video link) during a real-time interactive session with a remote healthcare specialist. This use of CARA is over a short interval of time and is fully supervised with guidance for the patient in the wearing of the BAN and the use of the interactive system. It makes effective use of time for both the specialist and remote patient, and furthermore a facility in CARA to record both video and associated sensor data streams allows the session information to persist.

This restricted use of the CARA system is also important from a number of viewpoints. An inherent problem with all wearable wireless small sensors is noise of various kinds, and this results in data errors. The real-time visual monitoring along with the sensor recordings allows the consultant to disambiguate erroneous readings from significant readings. Furthermore, it avoids the medical, legal and social issues associated with introducing new models of healthcare, and instead is an alternative, less-disruptive approach. This use of the system provides an immediately practicable solution that respects current healthcare practice and the experience of both patient and caregiver, and leads to incremental incorporation of the technology.

While this use of the CARA system is useful, there may also be a need to continuously monitor a patient during normal daily activities. While ultimately this would be done be the wearing of a BAN, a less disruptive initial scenario is the monitoring of patient movement using the sensors available in a commercial smartphone. This is more immediately acceptable, and, by integration with the CARA system, it can make use of the sophisticated analysis and data management capabilities of that system.

III. Related work

Recent advances in sensor technology and wireless networks allow long-term unobtrusive health monitoring in a post-hospital residential environment. By integrating wearable physiological sensors and environmental sensors using wireless networks, a wireless health care system can provide continuous monitoring of patients' vital signs (e.g., blood pressure, pulse oxygen, etc) and their environments (e.g., temperature, humidity, etc), which makes pervasive computing in healthcare [3] become feasible.

As a result, sensor-based remote monitoring system becomes an active research area [4], [5], [6].

Bentsen et al. [7] develop the Open-Care Personal Communication Device (OCPCD) which is an extension of the Open-Care Service Engine infrastructure in the shape of a mobile wrist or neck-born device for monitoring users vital signs and activity-level while not at home. It also provides fall detection by integrating with an external fall detection module.

Sordo et al. [8] present the state of the art in intelligent pervasive healthcare applications and the corresponding enabling technologies. They discuss pervasive healthcare systems in either controlled environments (e.g., health care units or hospitals), or in sites where immediate health support is not possible (i.e. the patient’s home or an urban area)..

Sweta et al. [9] present an architectural framework of a system utilizing mobile technologies to enable continuous, wireless, electrocardiogram (ECG) monitoring of cardiac patients. The proposed system has the potential to improve patients' quality of life by allowing them to move around freely while undergoing continuous heart monitoring and to reduce healthcare costs associated with prolonged hospitalization, treatment and monitoring.

Capua et al. [10] present an original ECG measurement system based on web-service-oriented architecture to monitor the heart health of cardiac patients. The projected device is a smart patient-adaptive system able to provide personalized diagnoses by using personal data and clinical history of the monitored patient.

IV. CARA System overview

Advancements in internet technology have made possible innovations in the delivery of healthcare. Universal access and a networking infrastructure that can facilitate efficient and secure sharing of patient information and clinical data, make the internet an ideal tool for remote patient monitoring applications.

An overall architecture of the CARA healthcare system is shown in Figure 1. At the core of the system is the user, also referred to as the “subject” (in a research environment) and as the “patient” (in a clinical or therapeutic environment). The user’s vital signs are monitored by different kinds of sensors within a wireless BAN, and are amplified and converted to digital signals. All measurement data gathered by the base-station are then transferred to a gateway (often a PC or a smart phone). The communications links used between the BAN and the personal server will vary according to circumstances.

A real-time classification system for the types of human movement associated with the data acquired from 3-axis accelerometer and gyroscope sensors has been implemented on the gateway. It distinguishes between periods of activity and rest, recognizes the postural orientation of the wearer, detects events such as sitting/lying and falls. Monitoring of movement and different postures involved in the daily activities of older persons who are living alone may help to identify persons at risk of falling or who have fallen. Such ability may also allow a better assessment of activities of daily living and the effects of treatments, and progress of medical conditions.

The gateway connects over the internet to the CARA server which provides sensor data services. An Adobe Flash application running in the gateway publishes real-time sensor data along with live video streams to the CARA server. On the server side, data derived from the sensor data is stored in an implementation independent generic format (i.e. XML), and also kept in an embedded database. As a part of the system, the reasoning or rule engine is developed and deployed on the server-side as an intelligent agent; it is designed to help reactive decision-making on data alerts. The medical consultant logs into the flash application remotely and selects the appropriate patient. The application then provides the consultant with continually updating views of the real-time readings. Additionally, the consultant can record the session, and the system provides the facility to review the session by replaying from the server the recorded readings and synchronized video. The consultants can thus analyze the session and issue a clinical report containing their findings.

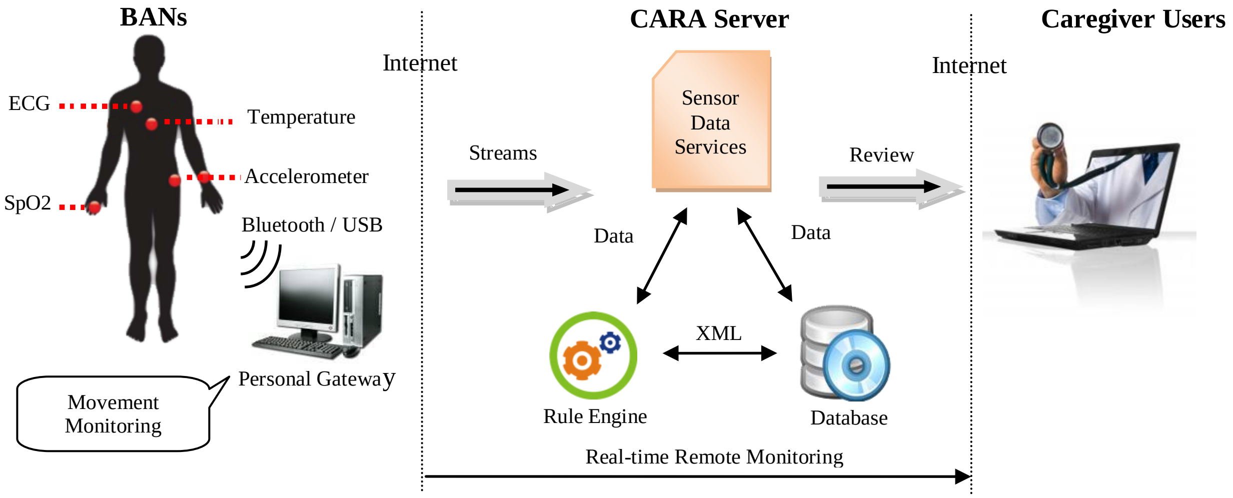


Figure 1. CARA system architecture

V. CARA System implementation

The current CARA system prototype provides remote physiological signal monitoring using a medical BAN with on-demand video recording services, along with a data and video review functionality to assist diagnosis. The remote monitoring is able to provide continuous real-time physiological signal monitoring over the internet, and it is also able to send alarms when an emergency is detected. The on-demand video monitoring can be used to provide a live two-way video link supporting a fully interactive session between caregiver and patient. Additional movement sensors in a smartphone provide additional non-intrusive monitoring capabilities and these sensor readings are also integrated into the CARA system. All the sensor readings and video records are stored in the database on the CARA server so that the caregiver can review the data anywhere anytime.

A. *Hardware Devices*

The wireless monitoring devices which are the basis of the body sensor network are developed by the Tyndall Institute of University College Cork. The sensor platform is a generic 25mm×25mm module that has been deployed in applications ranging from medical measurement to agriculture. Table I shows the sensors we use. The base-station module comprises an Atmel AVR Atmega 128 micro-controller and a Nordic RF2401 Transceiver. It not only automatically receives the incoming physiological signals from the wireless sensors using synchronization messages, but also transfers the physiological data through a USB connection to the personal gateway. A web camera can be attached to the client when the live video monitoring is required.

TABLE I. SENSOR TYPES AND THEIR SENDING RATES

Sensor type	Sending rate(Bytes/Sec)
Blood oxygen	200
ECG	500
Temperature	100

B. *Software Design*

The CARA system employs a multi-layered software infrastructure based on the features and functions at each of the network tiers. We use the Embedded C++ programming language to implement the physiological data sensing in the wearable sensors and the data transmission between mobile wearable sensors and the base-stations. The base-station is also developed in Embedded C++, and this software acts as the middle-ware of the CARA system. It receives sensor data from wireless wearable sensors and forwards them to the personal gateway. We use the windows operation system and Adobe Flex builder to develop the software application for the healthcare client. The client running on the personal gateway is able to collect the sensor data and transfer them to the CARA server in real-time. The CARA server is developed and deployed under the ASP.NET environment on the windows operating system. A built-in database and windows workflow rule engine provide the basis of the data management system on the server which provides data services for the users. Data services include real-time remote monitoring and data review functions.

VI. LOW-DISTURBANCE SCENARIOS AND THEIR IMPLEMANTION IN CARA

As outlined above, a number of scenarios are supported by CARA, and these allow the incremental use of the system and thereby encourage the adoption of the technology.

A. *Real-time Remote At-home Monitoring*

This scenario involves real-time at-home monitoring under remote supervision by a caregiver. Real-time sensor data and video streams are generated by CARA and sent to a remote caregiver, who might be any suitable healthcare worker, specialist or non-specialist. It involves use of a two-way video link whereby the patient and remote caregiver have direct views of each other. This is an important part of making this scenario low-disturbance and non-stressful, and thereby gaining acceptance for the technology.

The Remote Monitoring System is designed as a web application using RIA (Rich Internet Applications) technologies. In traditional web applications, there is a limit to the interactivity that can be added to a single page. This often leads to delays, during which time users may get tired of waiting and doctors may waste valuable consulting time. With RIA technologies, the client computer and the server can communicate without page refreshes. In this way, web applications can support more complex and diverse user interactivity within a single screen. This allows real time user interaction, essential for our system. Adobe Flex can be used to build RIAs based on Adobe Flash, and it is estimated that over 90% of web users now have the Flash Player installed on their computers. The client application is implemented using Adobe Flex Builder and Microsoft Visual Studio tools.

Once the wireless wearable devices are set up properly, the application starts receiving data from sensors through wireless communication. In the meantime, the sensor data is transmitted to the CARA server through Action Message Format (AMF) [11] protocol. The sensor data will also be published along with the video stream on the Flash Media Server (FMS) as metadata. (The Flash Media Server is a proprietary data and media server from Adobe Systems).Continuous monitoring is carried out by the caregiver remotely in real-time while the elderly is wearing the wireless monitoring sensors (See Figure 2).

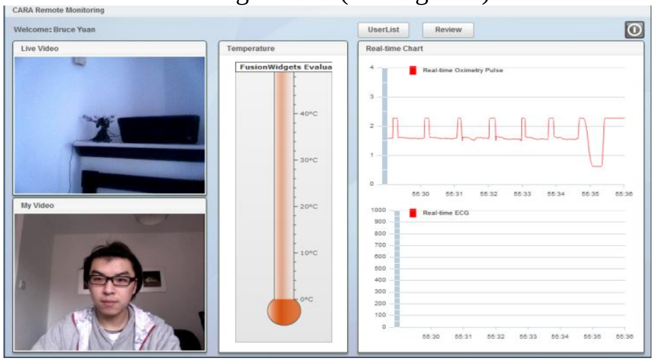


Figure 2. Remote Monitoring Integrate With Live Video Application.

B. *Remote Healthcare Data Review*

An important use of CARA for the caregiver is the ability to record and review the real-time patient monitoring session. It is very convenient for the caregiver to record a monitoring session and then review both the video stream and associated real-time vital sign data at any later time. This plays a part in encouraging use of the CARA system, and is an incremental stage that leads on to the automated use of the system, where long-term at-home real-time vital-sign data may be reviewed by health professionals.

The Data Review System supporting this scenario includes the sensor data review and video replay applications. The data review application allows the caregiver to analyze the full context of sensor readings in order to distinguish critical from non-critical situations. Clicking the review button on the main user interface brings up the data review interface and the caretaker is able to define parameters for reviewing the sensor data, for example: selecting patient profile; defining duration of data reviewing; setting data priority; choosing types of sensors. Once the caretaker sets up all these options, they can get the sensor data review chart shown in Figure 3. It shows the recorded sensor readings from the sensor database for the selected patient or a certain time period. The chart is implemented in Flash by using the amCharts API. It supports zooming and scrolling functions so that users can adjust the graph easily and analyze the data. A playback function is also integrated into the chart which enables the user to play the sensor data graph at any speed.

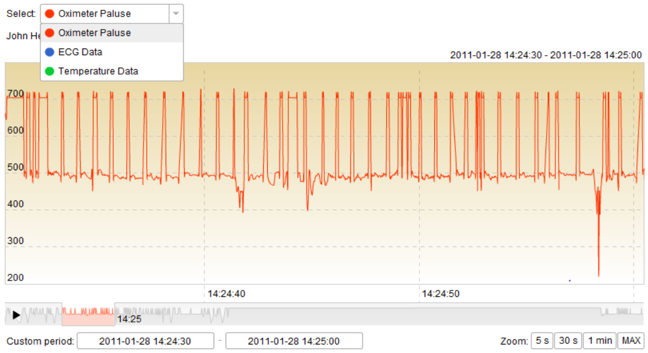


Figure 3. Sensor Data Review Chart

The video replay application is designed for the caregiver to review the recorded patient's live video along with the associated real-time sensor data. This function is developed within the real-time remote monitoring system. Whenever the patient's live video stream is published on FMS, it is also recorded as a flash video file on the server. To distinguish the video file and to synchronize with the sensor data, several correlations of the video must be recorded into the database as well (e.g. video start time, end time, patient's information). An added functionality of the Data Review System is the ability to annotate the BAN data streams to highlight readings and situations that demand attention. This will allow a caregiver to add their expertise to the system and in this way allow the Healthcare Reasoning System to be improved and refined as more data is added and reviewed.

C. Automated Movement Monitoring

This scenario is where the system is analyzing the real-time vital signs and applying rules that identify critical patient conditions. In the more advanced use of the CARA system this will involve analysis of the real-time vital signs when the patient is wearing the medical BAN continuously. For the low-disturbance scenario considered here, rather than wearing the BAN the patient just has to carry a smartphone, and the CARA components that implement this scenario are the Movement Monitoring System and the Healthcare Reasoning System.

The classification algorithm we used for movement monitoring is based on the previous study by Mathie et al. [12]. The determination of patient movement is based on the relative tilt of the body in space, called postural orientation. Movement classification is based on evaluating the subject's tilt angle. A tilt angle is 0 to 60 degrees, is classified as an *upright* position; an angle of 60 to 120 degrees is classified as a *lying* position; a tilt angle greater than 120 degrees is classified as an *inverted* position. Following Mathie [12] a tilt angle between 20 and 60 degrees is definitely *sitting*, whereas angles of 0 to 20 degrees may be either *sitting* or *standing*, and therefore sitting and standing may sometimes be incorrectly classified. The detection of falls is based on the signal magnitude vector (SVM), defined as the root mean square of the accelerations in the x, y and z dimensions (x_i, y_i, z_i):

$$SVM = \sqrt{x_i^2 + y_i^2 + z_i^2} \qquad (1)$$

A threshold is determined based on the accelerations in the x, y, and z axes, using simulation of falls and stumbles. If at least two consecutive peaks are detected above the defined threshold then a possible fall is determined. A classification of possible fall is assigned to the current time period, and the CARA server will be notified of a potential safety threat. The server will receive the fall event data together with the complete signal for 30s following the event, and this can then be transmitted to the medical consultant.

The Healthcare Reasoning System provides a general rule engine that can be tailored with different rules for different situations. It executes in real-time and can offer immediate notification of critical conditions. Some critical conditions may only be identified from correlating different sensor readings and trends in sensor readings accumulated over time.

VII. SYSTEM EVALUATION

A. Precision of Movement Monitoring System

An evaluation of the test results involved comparing the movements classified by the system with the actual movements of the subject. A correct classification is recorded if the classified movement actually occurred during the appropriate time interval; an incorrect classification is recorded if some particular movement results in an unexpected classification. In this way, the accuracy of the system in correctly classifying the movements of the tests was determined(see Table II).

TABLE II. RESULTS OF MOVEMENT EXPERIMENTS

Movement	Task	Tests	Correct	Accuracy (%)
Postural Orientation	Stand→Sit	40	35	87.5
	Sit→Stand	35	31	88.6
	Lying→Sit	28	26	92.9
	Sit→Lying	30	28	93.3
Fall	Fall (active)	22	21	95.2
	Fall (inactive)	23	22	95.7

B. Physical Performance

Two experiments were conducted to test the CARA system's physical performance. The first experiment is to evaluate signal quality between the wearable monitoring devices (WMD) and a basestation PC at different distances. We fixed the location of the basestation PC and tested wireless communication link quality at distances ranging from 1m to 15m from the WMD. We found that the closer WMD offered better transmission quality (see Table III). The signal to noise ratio (SNR) value was also affected by some obstructions such as doors or movement of the subject.

TABLE III. SNR VALUE OF THE WMD

WMD	Distance		
	1m	*7m*	*15m*
SNR(dB)	11	20	33

The second experiment is to evaluate the impact of the potential delay of the network. We tested our remote monitoring system through localhost, intranet and internet respectively, and the results indicating data transmission delay in milliseconds are shown in Figure 4. The delay caused by internet latency is unavoidable under the current approach. However, this delay does not significantly affect the working of this system.

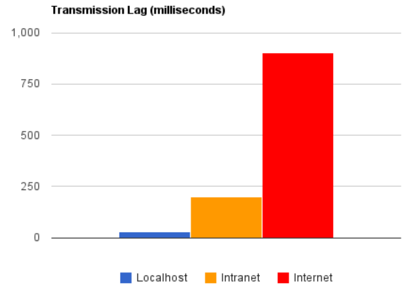


Figure 4. Data Transmission Lag

C. Security & Privacy

Security and privacy issues are taken into account in the CARA system as well. Password control allows only authorized user to log in to the CARA system. Authority management is integrated into the system to achieve the privacy control, which means different users can access different functions of the system according to their authorities. For example, the medical consultant can view the patient profile while others cannot. Further work will provide more precise control of various security functions in the CARA system.

VIII. CONCLUSIONS

The CARA pervasive healthcare system presented here provides an advanced technical solution for automated at-home healthcare. It is recognized, however, such a change from current practice may be unacceptable, and incremental introduction of the technology may be the best path to successful use of the technology. Following this approach, the system supports scenarios where the wireless BAN is used under remote supervision for a real-time interactive monitoring session with a caregiver, and a scenario where the BAN is not worn and continuous automated monitoring is just based on sensors in the smartphone carried by the patient. While not using the full capabilities of CARA, these scenarios provide a non-stressful introduction of the technology, and gain acceptance for more advanced scenarios such as non-interactive, at-home automated patient monitoring using the BAN.

Important aspects of the CARA system include: inter-visibility between patient and caregiver; real-time interactive medical consultation; and replay, review and annotation of the remote consultation by the medical professional, where the annotation of significant parts of the multi-modal monitored signals by the medical professional provides the basis for the improving automated intelligent analysis. A design goal of ubiquitous access has resulted in a web-based implementation that provides access and analysis capability on any internet-connected PC or appropriate smart device. The paper has provided an overview of the CARA system and scenarios of its use, and presented results of experiments using the system.

ACKNOWLEDGMENT

This project is funded by Science Foundation Ireland. We are also grateful to University College Cork and the MISL Research Group.

REFERENCES

[1] U. C. Bureau, "International Database. Accessed via Internet:http://www.census.gov/ipc/www/idb/," November 2008.

[2] U. Varshney, “Pervasive healthcare,” *Computer*, vol. 36, no. 12, pp. 138-140, 2003.

[3] J. E. Bardram, A. Mihailidis, and D. Wan, "Pervasive computing in healthcare", London: CRC, 2007.

[4] T. M. T. Gao, L. Selavo, M. Welsh, and M. Sarrafzadeh, "Participatory User Centered Design Techniques for a Large Scale Ad-Hoc Health Information System," *Proc. of HealthNet*, 2007.

[5] W. S. Q. Wang, X. Liu, Z. Zeng, C. Oh, B. K. A. li, M. Caccamo, C. A. Gunter, E. Gunter, J. Hou, K. Karahalios, and L. Sha, "I-Living: An Open System Architecture for Assisted Living," *Proc. of IEEE SMC*, 2006.

[6] H. N. F. Dabiri, and H. Hagopian, "Lightweight Medical BodyNets," *Proc. of BodyNets*, 2007.

[7] P. Bentsen, J. H. Nielsen, J. Nybo *et al.*, "Continuous healthcare." *Pervasive Computing Technologies for Healthcare*, 2010, pp. 1-2.

[8] M. Sordo, S. Vaidya, L. Jain *et al.*, "Intelligent Pervasive Healthcare Systems," *Advanced Computational Intelligence Paradigms in Healthcare*, *Studies in Computational Intelligence*, pp. 95-115: Springer Berlin / Heidelberg, 2008.

[9] S. Sneha, "A wireless ECG monitoring system for pervasive healthcare," 1/2007, U. Varshney, ed., *International Journal of Electronic Healthcare*, 2007, pp. 32-50.

[10] C. De Capua, "A Smart ECG Measurement System Based on Web-Service-Oriented Architecture for Telemedicine Applications," *A. Meduri, ed., Instrumentation and Measurement, IEEE Transactions*, 2010, pp. 2530-2538.

[11] S. Inc, "AMF3 Specification. Category: ActionScript Serialization," 2006.

[12] M. J. Mathie, and B. G. C. A. C. F. Coster, and N. H. Lovell, "Classification of basic daily movements using a triaxial accelerometer," *Med. Biol. Eng. Comput*, 2004, pp. 670-687.

FlexiSim: A user friendly Environment for Network Simulations

A. Harshit Agarwal[1], B. Antriksh Saxena[1], C. P. Venkata Krishna [1], and D. V.Saritha[1]
[1]School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India

Abstract - *NS2 is a discrete event simulator targeted at networking research. NS2 provides substantial support for simulation of routing protocols over wired and wireless (local and satellite) networks. NS2 uses TCL script for its front end and C++ as its back end [1,5]. NS2 is completely command based and one needs to create a TCL file for running any network simulation. Therefore, it's very tedious for a user to go through all the syntaxes and semantics of the required commands. This paper presents the development of a new graphical network simulator named FlexiSim which is based on NS2 [1]. Additionally, in this paper we discuss the design of FlexiSim and the implementation of a routing protocol through it. FlexiSim remarkably simplifies the procedure involved in running the simulations on NS2, by providing an interactive and flexible environment to the user therefore making it user friendly. No simulator existed for the code generation of NS2 prior to the design of FlexiSim. The most important advantage of FlexiSim is that it is a simple simulator that can be used for performing simulation exercises of moderate complexity under realistic network characteristics.*

Keywords: FlexiSim, NS2, Wireless Networks, Simulation, Simulator Development

1 Introduction

A network simulator is a software program that suggests the working of a computer network. In network simulators the network is modeled and it is generally used to analyze the performance. Using actual test bed by setting up routers, computers and data links may involve high cost and time. Simulators emulate the real world scenario therefore provide relatively inexpensive and fast way. A popular example of a network simulator is NS2 which is used in the simulation of routing protocols and it is heavily used in ad-hoc networking research. NS2 is an open source model with online documentation. However, modeling is a very complex and time consuming task in NS2 since it has no GUI [1].

In this paper, we report the development of a new NS2 based simulator, FlexiSim, which is built to analyze different wireless protocols in a more flexible way. One of the major problems with simulating wireless networks on NS2 is the complexity involved in using NS2 [1]. It is extremely challenging for the user to first go through the complete working of NS2 [1] then write the protocol and topology. FlexiSim was developed with extensibility in mind. In this paper, we discuss, as well, the issues involved in the implementation of AOR-GLU [2] using FlexiSim. This serves to demonstrate a case of the use of FlexiSim to perform wireless simulations.

A variety of network simulators (e.g., GloMoSim [3] and OMNeT++ [4]) is already available and is widely used by the industry and the academia for various purposes. However, use of these simulators by a novice may turn out to be a tough job for him because of the perplexity involved. Since working with NS2 is not an easy task, a user has to go through all the commands and the available functionalities before simulating even a simple network topology. So the primary purpose of FlexiSim is to simplify this process, make it faster, flexible and more robust. Instead of memorizing all the syntaxes and semantics of the commands the user just needs to have a basic overview of the topology and the protocols, the TCL code [5] used to implement this simulation will be produced by FlexiSim. FlexiSim incorporates its own code verifying mechanism and checks for the validity of the input given by the user, thereby preventing the generation of erroneous code. It consists of various menu options through which the user can create new nodes, agents, applications, links etc and configure them as well. The application automatically generates the TCL code [5] for the input given by the user. When this code is run in this application, FlexiSim automatically calls NS2 which generates the trace (.tr) file [6] and .nam file [6]. These generated files are saved in the project folder and can be used later for further analysis. Once .nam file is generated the FlexiSim opens up the NAM (Network Animator) [6] and runs the file on it, so that user can visualize the simulation. Tracegraph, a free network trace analyzer for NS2 can be launched from FlexiSim with the help of which user can make scrupulous study of the results of the simulation. For FlexiSim to work, it is necessary that the target machine should be installed with ns-2.34 [1] and nam-1.14[6]. At present FlexiSim 1.0 is designed only for wired and wireless topologies (local).

FlexiSim is developed on a Linux platform. Most of development was carried out on Ubuntu 10.04 [7], but the software was so designed that it could be run on all variants of Linux and UNIX systems. NetBeans 6.9.1 [8] development environment was used for rapid and organized development. NetBeans offered unparallel support for development and debugging in Java [11], the language used for development. FlexiSim is a standalone Java based desktop application which provides a GUI for NS-2 [1].

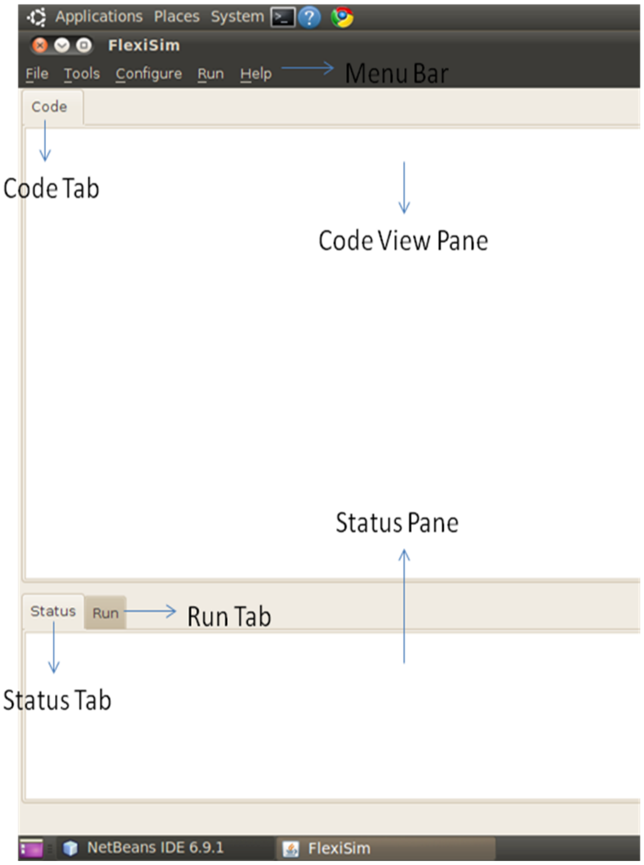


Figure 1. FlexiSim screenshot

2 Motivation

Before starting the development of FlexiSim we analyzed various available network simulators as mentioned above. Among them NS2 is the most commonly used and trusted simulator. Compared to its contemporaries, NS2 provides simulation for almost all types of wired and wireless networks [12]. But there is a major usability issue with NS2 as it involves the use of two languages one is C++ for modelling the behaviour of simulation nodes and other one is oTcl script that controls the simulation and create network topologies.

FlexiSim was designed keeping in view the difficulty in employing NS2 for simulations. Firstly, before using NS2, user has to comprehend its working and he should have an overview of its architecture. Secondly, he has to learn TCL for creating network topologies and C++ for creating protocols [1,5,6].

In spite of these issues, NS2 is widely used for simulations because of its accuracy in simulating wireless networks. Therefore FlexiSim was designed to use NS2 at the back end. At the front end it provides user with a rich GUI where even a naïve user can perform simulations and get the same results as he would get by using NS2 [1].

3 Development of FlexiSim

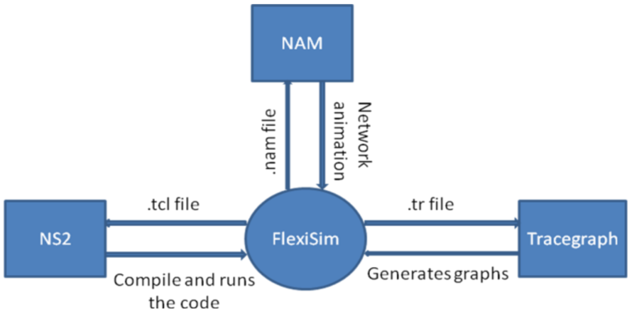


Figure 2. Interactions of FlexiSim with other programs

As mentioned above FlexiSim is developed with extensibility and scalability in mind. It was developed using Java [11] on Ubuntu 10.04 [7]. The design choices and the programming structure were also kept flexible and extensible to make the simulator easy to use. The simulator has been designed keeping in mind the problems the end users might face while writing the TCL code.

3.1 Platform Overview

The simulator was developed using Java on Ubuntu 10.04, but it is targeted for all the operating systems. The main reason behind choosing a Linux based platform is because of the NS2.34, NAM, and various trace analyzers, all of these are compatible with Linux [1,6,7,9].

An object oriented approach has been followed to develop this software. The cause behind using Java is it being completely object oriented and platform independence. Because of the platform independence offered by Java later versions of FlexiSim can be made compatible with windows too. The whole project was developed on NetBeans IDE 6.9.1. NetBeans provides a simplified development of the Java Swing Desktop application. It also has a Java Swing GUI builder (formerly known as "Project Matisse") which provides an interactive and rapid way of creating GUIs [8].

3.2 Code Organization

For the development of FlexiSim through Netbeans IDE [8] the code is distributed into various packages. The packages were formed on the basis of the functionality that they provide. Each of them holds a set of classes pertinent to it. For Example, there is a package named FlexiSim.node. This package contains all the classes related to nodes. Likewise there are packages for other entities too.

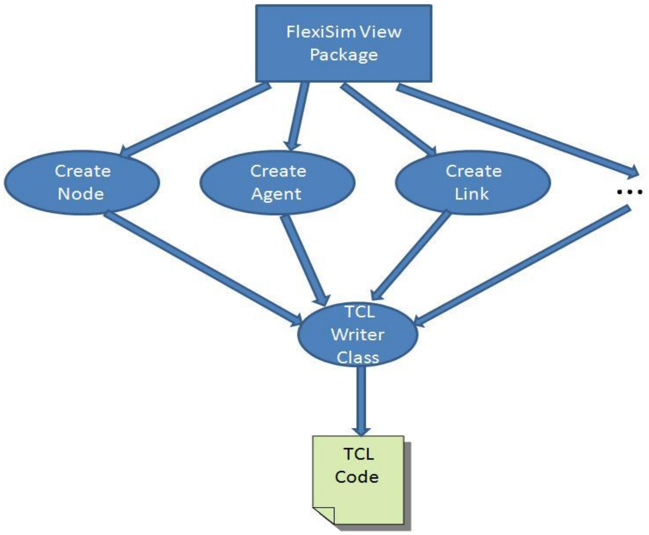


Figure 3. FlexiSim Package

As shown in Figure 2 there is a package called the FlexiSim View Package. This package contains the classes of all the entities that can be either created or configured through FlexiSim. For example, there is a class for creating nodes, creating agents and many more. All these classes call the object of TCL Writer class for writing the respective TCL code. In the TCL Writer class there are various methods for creation or configuration of each entity. As the user creates or configures any property of an entity the corresponding method in the TCL Writer class is called for generating the TCL code [5].

Suppose the user is creating a node. When the user clicks the Create Node button, an object of the Create Node class in FlexiSim View package would be called. This object would take in the parameters associated with node and then it would call the Node Write method in the TCL Writer class. This method would write the TCL code with the provided parameters into the TCL file. Similarly it works for other entities too [5].

4 Simulation Development through FlexiSim

It is very easy to develop a network simulation through FlexiSim. In order to maintain the brevity of the paper a simple wired simulation is explained. In general, to create a wired topology the following steps are to be followed.

1) Create a new project by specifying the project name and the project location. Also select the type of simulation to be created as wired.

2) After the project has been created nodes are created. While creating nodes there are two options, either the user can create a single node at a time or multiple nodes can be created at one go. While creating nodes the user needs to specify the node names.

3) Then links are created between the nodes that were made. The link type can be simplex or duplex link. Also properties of the link like bandwidth, delay and queue type need to be specified.

4) Agents can be created at different nodes. While creating agents the agent name, type and the node on which it is to be created is specified.

As an agent is created a new dialog automatically pops up regarding the application that is to be created. For application the name, type and its start and finish time are specified. Certain properties (if any) of application like packet size and bandwidth in case of CBR (Constant Bit Rate) are also specified.

5) After the agents and the application are created, the agents can be connected. Whatever agents that were created are shown in the drop down box, and the user can choose the ones which are to be connected.

6) Subsequently the events are scheduled. In event scheduling there are options for specifying the up-link, down-link time of the links as well as the detach time of the agents.

7) The various properties of the nodes such as routing protocol can be configured according to the requirements of the topology.

8) The above steps successfully create a topology. The next step is to run the simulation. In the run dialog the running time of the simulation needs to be specified.

As the simulation is run, FlexiSim automatically calls the NS2 which in turn generates the trace (.tr) file and the .nam file in the project folder which was created in Step 1. FlexiSim then calls NAM which is used for the visualisation of the simulation. Also the tracegraph can be called to analyse the performance of the simulation [6][9].

5 Case Study

In this section, we show how FlexiSim was used to simulate a routing protocol such as AOR-GLU [2]. Intentionally many implementation details were provided, while maintaining the conciseness of the paper, so that the readers could use these ideas from the simulation study and simulate experiments using FlexiSim for solving their own problems.

A typical addition of any new protocol in NS2, through FlexiSim, starts with modifying few base files in the NS2 installation folder.

5.1 Modification of NS2 Files

AOR-GLU [2] is an AODV [10] based protocol. To implement AOR-GLU [2] protocol in NS2 the existing AODV [10] protocol files are to be modified.

Apart from modifying basic protocol files to implement any protocol in NS2 [1] we need to modify NS2 [1] base files. The selection of the files and the modifications to be done in them are totally dependent upon the protocol being implemented. There is no standard as such suffice to which we can make the changes in the files. AOR-GLU [2] is a routing protocol for MANET and generally for implementing ad hoc based routing protocols some of the NS2 [1] files are required to be changed.

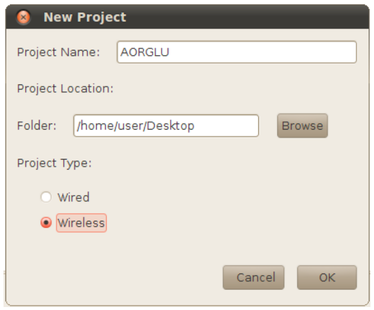
Once the protocol files and the above mentioned files are modified, NS is rebuilt to incur all the changes. To simulate the protocol a topology is created so that its performance can be analyzed. This analysis can be further used to compare the newly implemented protocol with the existing ones.

In the simulation 100 nodes were randomly distributed in an area of 500m X 400m. The node starting positions, destinations, and travel speeds were set random. Simulation duration was set to 150 seconds.

5.2 Simulation Sequence of Protocol Through Flexisim

Use The implementation of AOR-GLU [2] protocol through FlexiSim is explained below in steps.

1) For creating any new simulation or creating or modifying a protocol, through FlexiSim, firstly it's required to create a new project. While creating a new project the user is prompted to fill the project name, choose its location and select the project type (i.e. wireless or wired). In this case we chose the project type as wireless.


New Project

Project Name: AORGLU

Project Location:

Folder: /home/user/Desktop

Project Type:

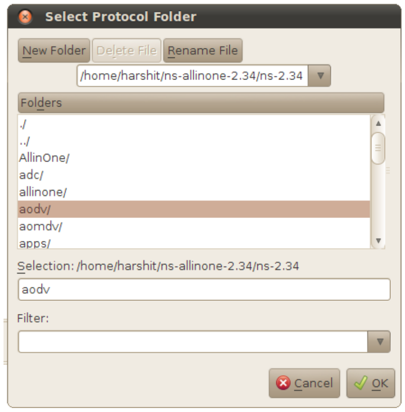
○ Wired

● Wireless

Figure 4. New Project

2) After the project was successfully created, the Modify Existing Protocol (because AOR-GLU [2] is implemented by modifying the AODV [10] protocol) option, present in the Tools menu, was selected. This launched a directory chooser through which the location of the protocol to be modified was selected. By default the path of the directory chooser is set to that of the NS2 installation folder.

After the protocol directory was chosen (in this case AODV) FlexiSim opened all the .cc and .h files, present inside the protocol folder, in different tabs. Along with these files the base NS files, which were to be modified as mentioned above, were also opened.


Select Protocol Folder

/home/harshit/ns-allinone-2.34/ns-2.34

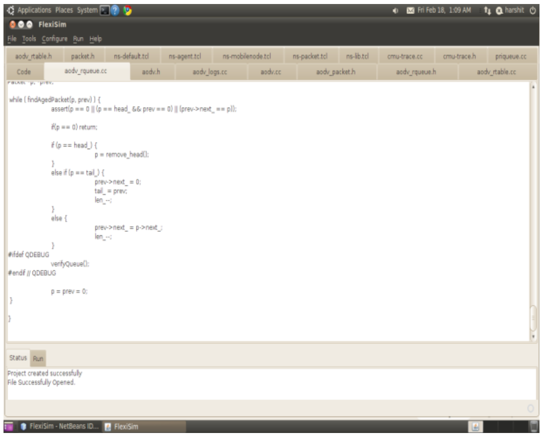
Folders
./
../
AllinOne/
adc/
allinone/
aodv/
aomdv/
apps/

Selection: /home/harshit/ns-allinone-2.34/ns-2.34

aodv

Filter:

Figure 5. Select Protocol Folder



```
Packet *p, *prev;

while ( findAgedPacket(p, prev) ) {
        assert(p == 0 || (p == head_ && prev == 0) || (prev->next_ == p));

        if(p == 0) return;

        if (p == head_) {
                p = remove_head();
        }
        else if (p == tail_) {
                prev->next_ = 0;
                tail_ = prev;
                len_--;
        }
        else {
                prev->next_ = p->next_;
                len_--;
        }
#ifdef QDEBUG
        verifyQueue();
#endif // QDEBUG

        p = prev = 0;
}

}
```

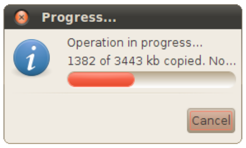
```
Project created successfully
File Successfully Opened.
```

Figure 6. Opened Tabs

3) The files opened by FlexiSim in various editable tabs were modified as per the requirements of the AOR-GLU [2] protocol.

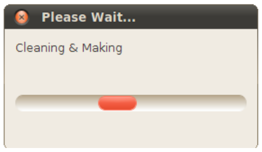
4) The project was saved so that the modifications are written in the protocol files. All the changes done in these files were performed in the copy of the original protocol which was present in the project folder.

5) The MakeNS option was selected from the Run menu. This option cleans and then rebuilds NS in order to incorporate the modifications in NS2. During this procedure a backup of the existing NS is also created in the project folder and the original protocol files and NS base files are replaced by the modified ones [1].



Operation in progress...
1382 of 3443 kb copied. No...

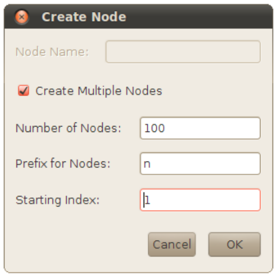
Figure 7. Backup creation in progress



Cleaning & Making

Figure 8. Clean and make operation in progress

6) A topology with 100 nodes was created by selecting the New Node(s) option from Tools menu.

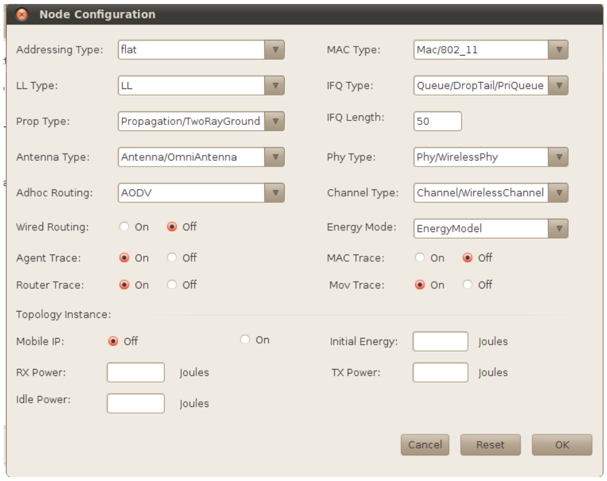


Node Name:
☑ Create Multiple Nodes
Number of Nodes: 100
Prefix for Nodes: n
Starting Index: 1

Figure 9. Multiple node creation

7) Configuration of the node was set using Node option from Configure menu. The following node configuration was set.

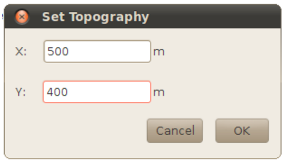
```
Channel/WirelessChannel     ;# channel type
Propagation/TwoRayGround    ;# radio-propagation
model
Phy/WirelessPhy             ;# network interface type
Mac/802_11                  ;# MAC type
Queue/DropTail/PriQueue     ;# interface queue type
LL                          ;# link layer type
Antenna/OmniAntenna         ;# antenna model
50                          ;# max packet in ifq
100                         ;# number of mobilenodes
AODV                        ;# routing protocol
ON                          ;# agentTrace
ON                          ;# routerTrace
OFF                         ;# macTrace
ON                          ;# movementTrace
```



Addressing Type: flat
LL Type: LL
Prop Type: Propagation/TwoRayGround
Antenna Type: Antenna/OmniAntenna
Adhoc Routing: AODV
Wired Routing: ○ On ● Off
Agent Trace: ● On ○ Off
Router Trace: ● On ○ Off
MAC Type: Mac/802_11
IFQ Type: Queue/DropTail/PriQueue
IFQ Length: 50
Phy Type: Phy/WirelessPhy
Channel Type: Channel/WirelessChannel
Energy Mode: EnergyModel
MAC Trace: ○ On ● Off
Mov Trace: ● On ○ Off
Topology Instance:
Mobile IP: ● Off ○ On
Initial Energy: Joules
RX Power: Joules
TX Power: Joules
Idle Power: Joules

Figure 10. Node configuration

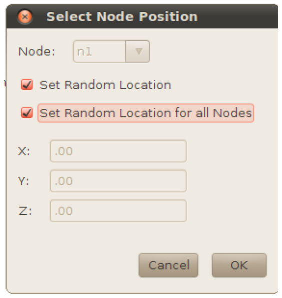
Configuration of the node was set using Node option from Configure menu. The following node configuration was set.



X: 500 m
Y: 400 m

Figure 11. Set topography

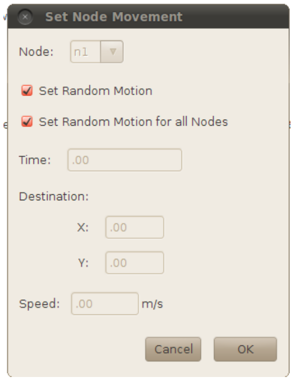
8) The position of the nodes was set randomly so as to distribute the nodes throughout the topography.



Node: n1
☑ Set Random Location
☑ Set Random Location for all Nodes
X: .00
Y: .00
Z: .00

Figure 12. Node position

9) The movement of the nodes was set with random speed and destination.



Node: n1

☑ Set Random Motion

☑ Set Random Motion for all Nodes

Time: .00

Destination:

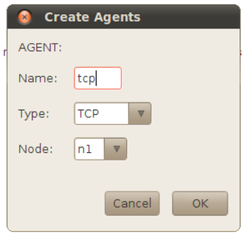
X: .00

Y: .00

Speed: .00 m/s

Figure 13. Node movement

10) Two agents, TCP and TCPSink, were created on two different nodes and a connection was established between the two. A FTP application was defined on the node with TCP agent where the start time was set as 10s. Later these two agents were connected.



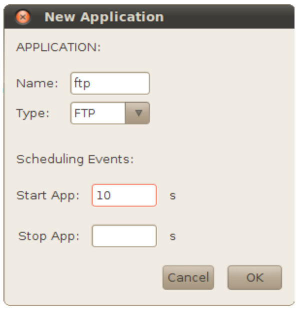
AGENT:

Name: tcp

Type: TCP

Node: n1

Figure 14. Creation of TCP agent



APPLICATION:

Name: ftp

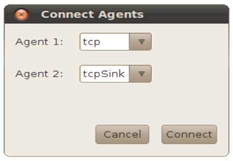
Type: FTP

Scheduling Events:

Start App: 10 s

Stop App: s

Figure 15. Creation of new FTP application

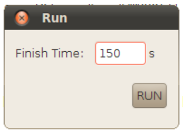


Agent 1: tcp

Agent 2: tcpSink

Figure 16. Connection of TCP and TCPSink agents

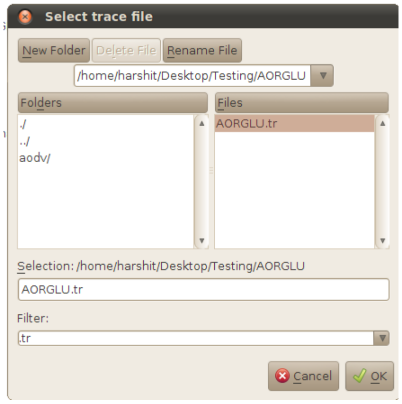
11) Once the creation of topology was complete, the simulation was run by selecting the Run option from the Run menu. The finish time was specified as 150s.



Finish Time: 150 s

Figure 17. Run

12) To further analyze the protocol the open trgraph was selected from the Run menu. The trace file for the above simulation was selected and subsequently the tracegraph [9] was launched.



/home/harshit/Desktop/Testing/AORGLU

Folders: ./ ../ aodv/

Files: AORGLU.tr

Selection: /home/harshit/Desktop/Testing/AORGLU

AORGLU.tr

Filter: .tr

Figure 18. Selection of trace file for tracegraph

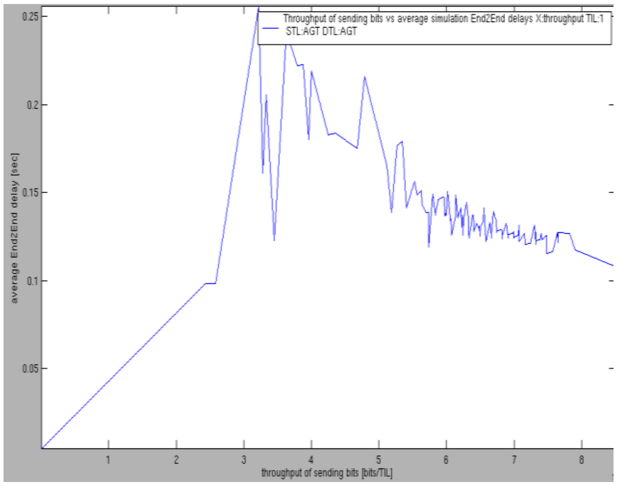


Figure 19. Tracegraph of the above simulation between throughput and average End2End delay

6 Conclusions

This paper gives a detailed explanation on design and implementation of a new NS2 based simulator, FlexiSim. FlexiSim is one of its kinds as it offers the user a flexible interface for creating topologies, creating new protocols in NS2 as well as modifying the existing protocols. FlexiSim not only makes easier for the user to run simulations on NS2 but also provides them with a faster and effective way of doing this.

In the future we plan to automate the modifications that are needed in creating a new protocol and also make the TCL file code pane editable. For the better evaluation of protocols we plan to include the feature of comparison between the different protocols with the help of suitable trace analyzer FlexiSim is available on request on http://www.flexisim.org.

This website provides FlexiSim along with its manual and tutorials.

7 References

[1] The Network Simulator – ns-2, 2002. http://www.isi.edu/nsnam/ns/.

[2] Chia-Chang Hsu; Chin-Laung Lei; “A Geographic Scheme with Location Update for Ad Hoc Routing” in Systems and Networks Communications, 2009. ICSNC '09. Fourth International Conference on vol., no., pp. 43-48, 20-25 sept,2009.

[3] GloMoSim. http://pcl.cs.ucla.edu/projects/glomosim/obtaining_glomosim.html

[4] OMNeT++, http://www.omnetpp.org/

[5] TCL code, http://www.tcl.tk/

[6] NAM: Network Animator. http://www.isi.edu/nsnam/nam/

[7] Ubuntu. http://www.ubuntu.com/

[8] Netbeans IDE http://netbeans.org/

[9] Tracegraph. http://www.angelfire.com/al4/esorkor/

[10] AA C. E. Perkins and E. M. Royer. “The ad hoc on-demand distance vector protocol,” in Ad Hoc Networking, C. E. Perkins, ed. Addison-Wesley, 2001, pp. 173-219.

[11] Java, http://www.java.com/

[12] AA Mekni, M, Moulin, B., “A survey on Sensor Webs Simulation Tools”, Sensor Technologies and Application, 2008, SENSORCOMM ’08. Second International Conference on , vol, no, pp. 574-579, 25-31 Aug 2008.

Software Update Methods for WSNs

Hay-young Jung[1], Yeungmoon Kwon[1], Byoungchul Ahn[1], and Bruce Kim[2]
[1]Dept. of Computer Engineering, Yeungnam University, Gyungsan, Korea
[2]Dept. of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL, USA

Abstract - *With the rapid advance of semiconductor technology, it is possible to use high performance sensor nodes to many application areas. Today's sensor nodes perform many functions at the same time and contain complex control programs. During the life time of sensor nodes, it is required to be reprogrammed their operations because of environment changes, control program bugs or update programs. To upgrade the control software of sensor nodes in WSNs remotely, this paper compares three update methods, which are the Many-Neighbors method, the Longest-Distance method and the Random selection method. Their performance is measured by upgrade times, the number of relay nodes, energy consumptions and error rates according to packet sizes. By the simulation results, the Many-Neighbors method shows 5 to 10 percent better performance than others. 160 bytes or 192 bytes of packet size are fast update time and low error rates.*

Keywords: WSN, data relay, firmware upgrade

1 Introduction

The performance and function of a node in Wireless Sensor Networks was very limited a few years ago. After sensor nodes were used until their batteries were exhausted, they were discarded without reusing or reprogramming. Typically sensor nodes for WSNs have been developed to reduce their costs and power consumption. With the recent rapid advance of semiconductor technology it is possible to implement high performance sensor nodes with performing many functions. Therefore they have very complex control programs without increasing its hardware cost. Also they can be reused again for different application areas.

It is necessary for sensor nodes to reprogram because of environment changes, control program bugs or program updates. In this case, a lot of costs and efforts are wasted if they are recollected or replaced them with new nodes. Therefore, it is required to reprogram their software remotely as far as they are reusable.

In this paper, software upgrade methods for sensor nodes are discussed and analyzed them to update efficiently.

2 Related works

Many researches for WSNs are studied for low power consumption or data collection methods from nodes to their sink node efficiently. For efficient data transmissions, Intanagonwiwat *et al*. suggests direct-diffusion method[1]. But this research is focused on data aggregation and data transmission path. So it is not suitable for software updates since the data flow direction is reversed. For lower power consumption, Wei Ye *et al*. suggests S-MAC. S-MAC is an MAC level protocol using sleep cycle and clustering[2]. This protocol use periodic sleep and data bandwidth is very low. It can't be applied for software data transfer since the data size of control software is much larger than that of the normal data.

Since control software contains execution code for a processor of sensor nodes, it is very important to maintain reliable data transfer. A method for reliable data transfer in WSN is developed for 1:1 communication such as S-TCP[3] and RMTS[4]. But 1:1 communication methods are inefficient to upgrade many nodes of WSNs. If these methods are used for re-programing sensor nodes of WSNs, each node must be updated first and retransmit software to another node one by one. Therefore it is necessary to develop an efficient upgrade method for sensor nodes with fast upgrade time and small data retransmission.

The direction to upgrade control software is the opposite direction of normal data transfer[5][6]. It is necessary to study for large data transfer from one node to many nodes efficiently. There are some researches about upgrades for sensor nodes. But they are focused on system management, not an upgrade itself [7].

3 System Model

All sensor nodes of WSNs are assumed to be the same model with the memory size and the same processor. It means that all sensor nodes use the same software version. And a distance between two nodes is the same and the location of nodes is fixed.

.

3.1 Data relay

When a sink node starts to transfer its software data to others, all nodes stop their sensing operation and switch to the software upgrade mode. The sensing operation mode is operating in low power consumption and each node transmits

its sensed data to the sink node. In case of software data transmission, data size is very large and must be transmitted very fast and continuously. If a node detects software data transmission protocol, it should switch the normal sensing mode to the software upgrade mode and prepare for software upgrade. When the node finishes receiving all software data, it requests lost packets to its source node. After lost data are received again, the node reprograms its own flash memory and restarts its operation again. After reprogramming, the node may relay software data to other nodes. But it is not necessary for all nodes to participate in relaying software to another node in WSNs. Only a few nodes relay software data to other nodes. It is very important to choose relay nodes because of overall performance.

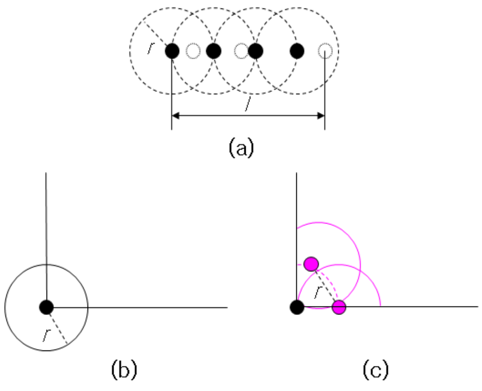
Fig. 1 (a) shows a data relay model when nodes are placed in-line. Black nodes in Fig. 1 (a) represent nodes that participate in relaying software data. And the most left black node is the start node. In Fig. 1, "*r*" means radio radius. All relay nodes are placed in multiples of radio radius of the location. The other nodes only receive software data and reprogram themselves. So the number of relay nodes, *N*, is calculated by Equation (1).

$$N = \left\lceil \frac{l}{r} \right\rceil \qquad (1)$$

r is a radio radius of a node

l is a distance between the start node and the last node

If nodes are placed in line, the number of relay nodes is proportional to the distance of the start node and the last node. But in real WSNs, each node is located on 2-dimensional plain rather than in line. Fig. 1(b) and Fig. 1(c) show software data relay nodes that are located on 2-dimensional plain at each step. If a node is located on the boundary of radio radius, it is the ideal software update. So the total number of relay nodes, N, is calculated by Equation (2).



(a)

(b)

(c)

Fig. 1. Data relay model

$$N = \begin{cases} 1, & (l <= r) \\ 1 + \sum_{n=1}^{n=\lfloor d \rfloor} \left\lceil \frac{2\pi \cdot n \cdot r}{r} \cdot \frac{1}{4} \right\rceil = 1 + \sum_{n=1}^{n=\lfloor d \rfloor} \left\lceil \frac{\pi \cdot n}{2} \right\rceil & (l > r) \end{cases} \qquad (2)$$

r is a radio radius of a node

l is distance between start node and last node

d = l/r

The minimum number of relay nodes is calculated by Equation (2). To upgrade all nodes in the field at least *N* nodes should participate in relay operation. When the start node is located in the corner of WSNs, Equation (2) is acceptable. If the start node is located in the center of WSNs, the number of relay nodes, *N*, is increased up to four times.

3.2 Power consumption and time to upgrade

The power consumption of relay nodes is calculated by the sum of receiving data and transmitting data, and is calculated as below.

$$J_{nr} = J_r \cdot file\_size\,(1+e)$$
$$J_{ns} = J_s \cdot file\_size\,(1+e) + J_{nr} \qquad (3)$$

J_r *is the energe for datas end*

J_s *is the energe for datas send*

J_{nr} *is the energy consumped by receiving node*

J_{ns} *is the energy consumped by relaying node*

e is trasnmissi on error rate

file_size is the size of firmware data

In Equation (3), J_{nr} and J_{ns} are the power consumption of receiving data and transmitting data to other nodes.

Some nodes located in duplicated radio area are received a few multiple times of data size of software data. Therefore, the total energy consumption of all nodes is *J* in Equation (4).

$$J = (J_s + J_r \frac{N_t}{Area\_of\_field} \cdot \pi \cdot r^2 + e \cdot J_s) file\_size \cdot N \qquad (4)$$

N_t *is total number of nodes in a filed*

N is a number of relaying nodes

In Equation (4), all parameters are fixed except *e* and *N*. It is very important to reduce transmission error and the number of relaying nodes.

The time to upgrade all nodes of WSNs depends upon the step count of relay. The time is calculated by Equation (5).

$$T = \lceil d \rceil \cdot (t_s + e \cdot t_s + t_u) \qquad (5)$$

d is relay step count (l/r)

t_s *is firmware file sned time (file_size / bandwidth)*

t_u *is update time*

e is transmission error rate

From Equation (5), the update time depends upon relay step count, *d*, and transmission error rate, *e*. If all relay nodes send software data to other nodes at the same time, they make a lot of collision or interference by radio signals and the transmission error rate is increased. It is necessary to prevent all nodes from relaying data at the same time in each step. To reduce energy consumption and upgrade time, it is very important to select a relay node at each step.

4 Relay procedure and simulation

In this section, a relay procedure to transmit software data is described. Simulation results are discussed.

4.1 Relay procedure

It is very important to reduce the number of relay nodes and prevent neighbor nodes from retransmitting data at the same time. As soon as a relay node transmits software data, it should select the next relay node to propagate the software update operation.

If the next relay node is located on the boundary line of radio radius of the present relay node, it is the most effective next relay node. But in real WSNs, there are a few nodes on the boundary line of radio radius. It is hard to recognize if the nodes are located on boundary line of radio radius. Selection methods of relay nodes are carefully considered.

The software transmission protocol has two steps. The first step is pre-transmission step. A relay node should know the status of all neighbor nodes before it starts sending software upgrade data. All neighbor nodes broadcast their status to the relay node periodically in the first step. The status data includes software version information, the number of neighbor nodes and the node status.

The next step is to receive software data and update it. Before the relay node starts to send the software data, neighbor nodes respond by sending a “*receive-start*” message. And neighbor nodes start to receive the software data and store it to their memory. After data transmission, neighbor nodes must request the relay node to retransmit lost packets. After they finish receiving the lost packets, they reprogram themselves and send back “*reprogram-done*” message to the relay node. If a time-out occurs or it receives messages from all participated neighbor nodes, the relay node should select the next relay node from neighbor nodes and sends “*relay-start*” message.

To select a next relay node, three selection methods are evaluated.

(a) Longest-Distance - A node is located at the longest distance from the relay node.

(b) Many-Neighbors - A node has more neighbor nodes than others.

(c) Randomly selected node - A node is selected randomly by the relay node

4.2 Simulation environments

NS-2 network simulator is used for the simulation. 100 nodes are deployed in the field uniformly. The distance between two nodes is 40*m* and each node has a 60*m* radio radius. Radio bandwidth is 256Kbps and firmware file size is 128 Kbytes. The energy consumption is 75.9*mW* to send data and 62.7*mW* to receive data. The time to upgrade a node is 1.5*sec*. The start node is located in the left bottom corner. Three selection methods are simulated to compare their performances.

4.3 Performance metrics

In order to evaluate the proposed models, following metrics are used.

(1) The number of relay nodes - Depending on the selection method of relay nodes, the number of relay nodes might be varied. It is the most important factor.

(b) Energy consumption - Since sensor nodes in WSN are operated by battery power, the energy consumption is important factor with total upgrade time.

(c) Total upgrade time to all nodes - This factor shows the effectiveness of software upgrade.

(d) Data loss rate - This factor shows the performance of each selection method

4.4 Simulation results

The number of nodes that participated in relay is calculated by Equation (2). Sensor nodes in WSNs are deployed in rectangular. It is very difficult to calculate exact number of relay nodes. By Equation (2), the number of relay nodes about 37. Fig. 2 illustrates the number of relay nodes that participated in relay. In all case, the number of relay nodes is between 33 ~ 48. When a node is selected as a relay node by Many-Neighbors, the total number of relay nodes is the smallest.

The energy consumed by each node is calculated with Equation (3). If there is no lost packet, error rate, *e*, is zero. Ignoring control and status packets, the energy consumed by each data receiving node, J_{nr}, is 257*mJ*. And the energy consumed by each relay node, J_{ns}, is the sum of receiving data and transmitting data. The energy consumed by relay nodes, J_{ns}, is calculated as 568*mJ*. The total energy consumption of WSNs is calculated by Equation (4) and is about 103.9*J*. Fig. 3 shows total energy consumption to

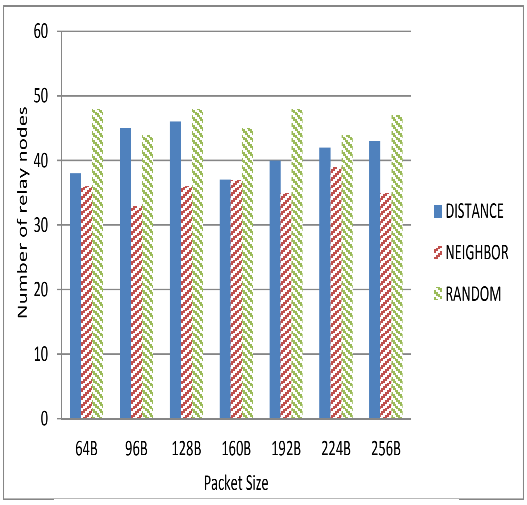


Fig. 2. Number of relay nodes

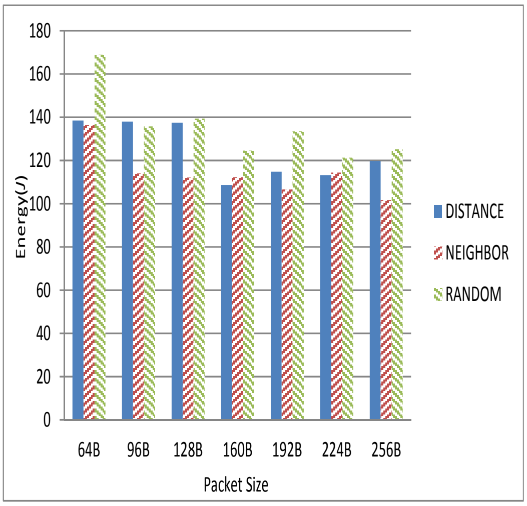


Fig. 3. Consumption of energy

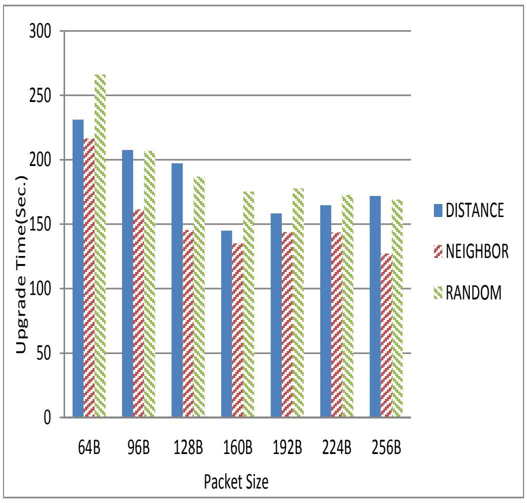


Fig. 4. Upgrade time

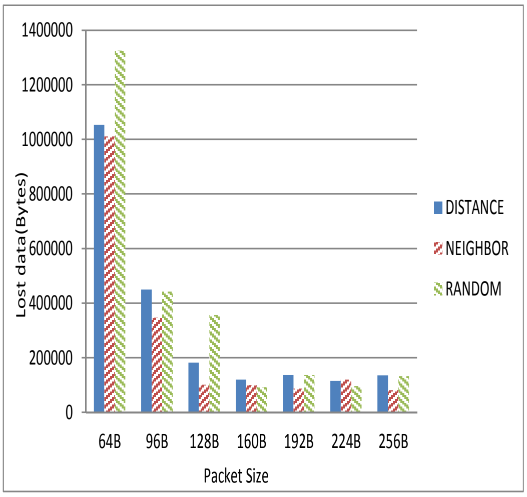


Fig. 5. Lost data size

upgrade all nodes. When the size of the data packet is 160 bytes, the energy consumption is low.

Total time to upgrade firmware is estimated as 51.2 sec by Equation (5). This value is the minimum value. But in the real world, some additional time might be added such as error recover time, message wait time and so on. Fig. 4 shows the time to upgrade firmware to all nodes.

Fig. 5 shows the size of lost or missing data. When packet sizes are between 160 bytes and 256 bytes, data loss is very small. When the packet sizes are smaller than 160 bytes, nodes transmit data frequently and they generate so many collisions. And the collision affects energy consumption and total upgrade time.

5 Conclusions

In this paper, three software upgrade methods are proposed by relaying data to sensor nodes in WSNs. The proposed methods are analyzed for the number of nodes to relay firmware data, energy consumption and upgrade time. These factors are simulated and measured by the several packet sizes. The performance evaluations are measured by update times, the number of relay nodes, energy consumptions and error rates by packet sizes. By the simulation results, the Many-Neighbor method shows 5 to 10 percent better performance than other methods. 160 bytes or 192 bytes of packet size shows fast update time and low error rates.

Through simulation results, the selection method of a relay node is very important and also the data packet size affects overall performance to update software.

6 References

[1] C. Intanagonwiwat, R. Govindan, and D. Estrin "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," Proc. of the 6th annual international conference on Mobile computing and networking(Mobicom '00), pp.56-67, 2000

[2] Wei Ye, J. Heidemann, and D. Estrin, "Sensor-MAC (S-MAC): Medium Access Control for Wireless Sensor Networks," Proc. of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), vol.3, pp.1567-1576, 2002.

[3] Y. G. Iyer, S. Gandham, and S. Venkatesan, “STCP: A Generic Transport Layer Protocol for Wireless Sensor Networks,” Proc. of 14th International Conference on Computer Communications and Networks, pp.449-454, 2005.

[4] F. Stann, and J. Heidemann, "RMST: Reliable data transport in sensor networks," Proc. of the First IEEE. 2003 IEEE International Workshop on Sensor Network Protocols and Applications, pp. 102 -112, 2003.

[5] W. Chen, P. Chen, W. Lee, and C. Huang, "Design and Implementation of a Real Time Video Surveillance System with Wireless Sensor Networks," Proc. of Vehicular Technology Conference, 2008, pp.218-222. 2008.

[6] Honggang Wang , Dongming Peng , Wei Wang , Hamid Sharif , Hsiao-Hwa Chen , “Image transmissions with security enhancement based on region and path diversity in wireless sensor networks”, IEEE Transactions on Wireless Communications, vol. 8, no. 2, pp.757-765, 2009.

[7] C-C. Han, R. Kumar, R. Shea, M. Srivastavam, “Sensor Network Software Update Management: a Survey,” Intl. Journal of Network Management, no. 15, No. 4, John Wiley & Sons, pp. 283-294, 2005.

Based on Boundary Location Technology to Separate an Image Semi-automatically

Peng Ge-gang, Li Xin-yu, Song Ying, Xiang Li-sheng, Shen Qing
(Talkweb Information System Co.Ltd., Changsha, Hunan 410205 China)

Abstract**—**the key audiences of mobile animation are those high school students, working youth population, etc. Their pursuit of new fashion, fresh and willing hands makes them like to change style frequently to get a pleasurable and a satisfactory time. This paper presents a semi-automatic way to achieve image fragment and synthesis for the creation of mobile cartoon characters. Based on such a technical sketch, a cartoon server may provide a software support to help user satisfy their willing on DIY and create a plenty of animation characters, then transmitted them from mobile to mobile.

To deal with this task, the first and important step is locating a good boundary-line around the target area and separating it automatically, so as to get a sub-image. The second step is using a digital logical operation to synthesize the sub-image into a target image. At last, achieve a "grafting" effect. Moreover, apply pseudo-color on the synthesis' to get a rather new fancy visual view.

*Index Terms***—**Mobile Cartoon, Digital Image Processing

I. INTRODUCTION

Mobile phone animation products currently include MMS, four cell phones comics, animation clips, and turn-based mobile games, etc. They are different from TV cartoon audiovisual works in three salient features [1,2]: First, it does not require large contiguous time. This feature just to meet those younger people living in the modern city, fast pace of life, large pressure of work, wish to enrich cultural life but have no continuous time to deal with. Secondary, The key audiences of these products are those high school students, working youth population, etc. Their pursuit of fashion, fresh and willing hands makes them very keen on change them self and others' style to make funny. Third, those high-end entertainment products such as mobile video and mobile game have highly requirements on mobile's hardware performance, such as processor speed, memory size, color resolution, screen size, etc. and they are more expansive. On the other hand, MMS, four cell phones comics, animation clips, etc. are in lower hardware performance needs, of cause in lower cost, so it is prefer by most mobile users.

This paper provides a semi-automatic way to realize image fragment and synthesis to create mobile cartoon characters. Based on this sketch, a cartoon server may provide a group of tools to help younger to satisfy his/her willing and to create many cartoon characters[3]. It is clear that, this is a new way on mobile users' own DIY to create a plenty of funny characters, and transmit them from here to there to heighten interesting on mobile cartoon.

To deal with this task, the first and important step is to realize a good boundary location. Then separate it automatically to get a sub-image, inside the boundary. And by add a digital logic operations merging the sub-image with another image and getting a "grafting" effect. Moreover, apply pseudo-color on the synthesis' to get a rather new fancy visual view.

II. TECHNICAL COURSE

A. Implement Boundary Location and setup marks

A.1 prepare

(a) Establish a temporary file F to store the gray-value $g(x, y)$ of each pixel of the original color image. The formula is:

$$I = 0.3B + 0.59G + 0.11R$$

Here R, G, B is the red, green and blue components of the original color pixel (x, y).

† This research was supported in part by National Technical Supporting Project Foundation of China (2007BAH14B04)

(b) Establish a mark file M to store the mark $m(x, y)$ of each pixel for the original color image. And set every mark to "255" (as white) beforehand.

(c) Establish an array A (the size is 256) for gray histogram to record the number of pixels that has the same gray value. For example, if there are 3564 pixels have a same gray value 134, so there is $A[134] = 3564$.

A.2. Calculate the best value as threshold T_h to find out a boundary

(a) Count all pixels in array A and named as Sum. $Sum = \sum_{i=0}^{255} A_i$, here A_i is the number of pixels that has same gray value i.

(b) With formula (1) to calculate the mean M_{t0}, M_{t1} and weight W_{t0}, W_{t1} for different range 0~t and t+1~255 (here variable t may change from 1 to 254),

$$M_{t0} = \frac{\sum_{i=0}^{t} iA_i}{\sum_{i=0}^{t} A_i}, \quad M_{t1} = \frac{\sum_{i=t+1}^{255} iA_i}{\sum_{i=t+1}^{255} A_i}$$

$$W_{t0} = \frac{\sum_{i=0}^{t} A_i}{Sum}, \quad W_{t1} = \frac{\sum_{i=t+1}^{255} A_i}{Sum} \qquad (1)$$

(c) By select different variable t and compare the variance δ_t on different range (0~t and t+1~255) to get a maximal δ_t. This maximal variance is the best value for threshold.

In this way, first, with formula (2) to calculate δ_t

$$\delta_t = W_{t0} W_{t1} (M_1 - M_0)^2 \qquad (2)$$

then, with formula (3) to get a threshold T_h when δ_t reach maximum.

$$T_h = t, when \nabla_t = \max(\delta_t), t = 1,2,...254 \qquad (3)$$

Here, the threshold T_h is a value calculated from program and user may modify this value to fit his own needs and get shorter or longer boundary-line.

A.3. **Compare the gray value of every pixel with T_h in entire image**. When one of the gray values is smaller than T_h the program changes the mark of relevant point in file M to "0" (as black).

A.4 **Scan the entire map again to find all boundary points.** When program discovers a "black" spot is around with at least one "white" adjacent point, then this point is marked as a "boundary point". With similar to the threshold T_h, the "white" adjacent points is a modifiable parameter too.

A.5 connect all boundary points to form a boundary-line

B. Manually assist to demarcate a border

When part of the boundary can not be set by program automatically or attempt to strip a body but it is not inside the border, in such case, manually assist is needed to demarcate a clear and closed border.

For example, if the background of an image is in light gray and the cheek is in light yellow, so it is difficult for program to set a valid boundary between the cheek and background. Another example may take a person wears a dark jacket, although there are good boundary between the neck and coat, but if the needs for creation do not want include a necklace, in this case may require to locate a new boundary manually above the necklace and abandon the original closed border.

C. Separate out the target area (inside the border) automatically

There are two stages to finish a separate operation. The first stage is marking a square area just surround the target sub-image. It includes three steps as below:

Step1. Scan the entire image from the very beginning line of whole image. When meet a line has boundary point firstly, this line is marked as a TOP line of the cut-off sub-image. Scan line by line again till the BOTTOM line of the sub-image, which means no longer a boundary point below this line.

Step2. Scan a line from left to right. When meet a border point on any column firstly, records it as a left border point (briefly named as LBP) of this line. Scan columns one by one again to find the right boundary point (briefly RBP).

Step3. Turn to next line and do the same operate as step2 does, to find out all the LBPs and RBPs of these lines.

The second stage is used to cut-off the target sub-image. It includes three steps as below:

Step1. Set all lines, which are beyond of the boundary-line, before the top and behind the bottom line, to null (as "white").

Step2. Set all columns which are beyond of the boundary-line (left of LBP and right of RBP) to null (as "white").

Step3. Find out the most LBP and RBP (named as LBPest and RBPest). Move up and left according to TOP line and LBPest column. Cut off the image behind the BOTTOM line and RBPest column.

Till now a target area is cut off from the original image (see Fig 2 below).

D. Paste to a target image

Our goal here is paste the cut-off sub-image to another image to achieve a "grafting" effect. There are 3 steps includes in this stage:

Step1: Define out a target area in the target image.

Step2：Resize the cut-off sub-image as large as the defined out area above.

Step3：Do merge operating with a digital logical OR (formula 4) to force the cut-off sub-image cover the defined out area.

$$h(x,y)=\begin{cases} null & a(x,y)=null\ ,\quad b(x,y)=null \\ a(x,y) & a(x,y)\neq null\ ,\quad b(x,y)=null \\ b(x,y) & a(x,y)=null\ ,\quad b(x,y)\neq null \\ a(x,y) & a(x,y)\neq nill\ ,\quad b(x,y)\neq null \end{cases} \qquad (4)$$

E. Pseudo -color rendering

This is an additional stage of this scheme. It is based on region growing algorithm. The goal is make the merged image a rather new fancy visual view.

Step1: Select a point inside the image as a seed. Usually a seed is one of the points inside the area which wants to be overspread to a larger area from the seed.

Step2： Set a stack and push the seed into the stack.

Step3：Check the gray value of the 8 neighbors of the seed. If the gray difference (between one of the neighbors with the seed) less than the pre-setting threshold, mark this neighbor as a new growing point and turn to step4. Otherwise, it is consider there is no new growing point here and turn step4 too. The range of pre-setting threshold is 0~255 and it may adjust by user in the program.

Step4：If the stack is empty turn to step5, otherwise pop out a new point as a new seed and turn step3.

Step5：Do render to the just grown area. The selected colorized color may chose from color table by user in this program.

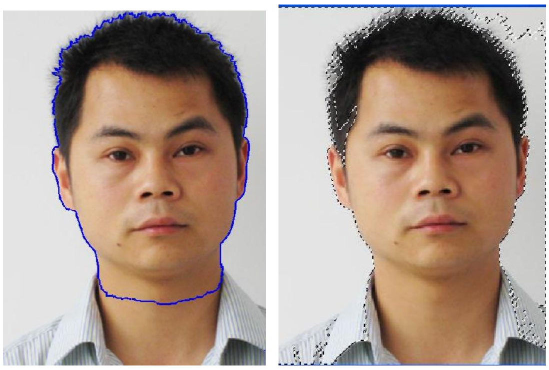
III CONCLUSIONS

This scheme is designed to make the image separation and synthesis (similar to Magic Wand in Adobe PhotoShop) in lesser manual intervention and more arbitrary. We regret to say that no human intervention in image separation is unlikely to be done. Because the computer is difficult to predict the separation object of the operator’s mind. And in some shooting conditions obtained images, it is difficult to fully automatic separation.

In addition, although we can not say which kind of border positioning technology is used by PhotoShop's Magic Wand, but we can happily report: As in our program, a user can modify the threshold T_h and the number of "neighbors", so that the boundary location was more controllable, see Figure 1, it is some what better than PhotoShop's Magic Wand. And the results of sub-image cut off and pseudo-color rendering shown in Figure 2, 3.

If you have any question or need the source code you are welcome to contact us with E-mail.

The inadequacies of the paper want to be criticized and corrected.



(a) This project (b) Adobe Photoshop

Fig 1 the effect compare on boundary location

Fig 2 a separated sub-image

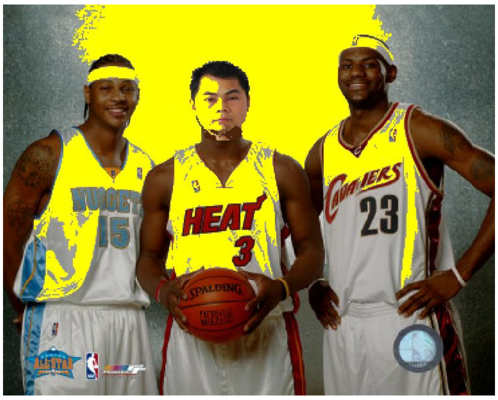


Fig 3 a non-commercial results on image merging and pseudo rending

IV ACKNOWLEDGMENT

We would like to thank deeply to the National Technical Supporting Project Foundation of China and the Ministry of Culture of China for their support on this project

References

1. Peng Ge-gang, Li Xinyu, Song Ying, Xiang Lisheng, Shen Qing, A Facial Organs Positioning Method Based on Grayness Mutation，ICWN’10.

2. Li Xinyu, Song Ying, Xiang Lisheng, Shen Qing, A Mobile Cartoon Creating Scheme Based on the Materials Reuse, ICWN ’09.

3. Peng Ge-gang, Li Xin- yu, Song Ying, Xiang Li-sheng, Shen Qing, Li Ren-fa, Research and Implementation of Mobile Phone Comic and Animation Assisted Production and Creative Platform， ICWN ’10

Authors

Peng Gegang: Peng is a post-doctor in computer applictation technologyHe is a senior architectural designer and interested in systemarchitecture, mobile communication, etc. Email: penggegang@hotmail.com

Li Xinyu： Li is the general manager of Talkweb Information System Corp. He is a senior architectural designer and received a BS degree in Computer and Information Engineering from HoHai University, Nanjing in 1988. He is interested in system architecture, mobile communication, etc.

Song Ying: Song is the vice general manager of Talkweb Information System Corp. He received a BS degree in Chemistry from China University of Geosciences,Wuhan in 1989. He is interested inmobile communication,multimedia, etc.

Xiang Lisheng: Xiang is a vice technical supervisor in Talkweb Information System corp. He received a BS degree in computer science from Hunan University in

1996.He is interested in MM, animation, etc. E-Mail: 13974838381@hnmcc.com

Shen Qing: Shen received a BS degree in electronic instrument from Changchun Geologic Institute. He was a professor in the Institute of Computer Science at National University of Defense Technology, China. Now he is the first chief technical consultant of Hunan Talkweb Information System corp. He is interested in pattern recognition, AI, MM, animation, etc. E-Mail:sq1950224@yahoo.com.cn

SESSION

THEORY + INTERESTING RESULTS

Chair(s)

TBA

MANET Localization Using Non-Circular Overlapping Range Maps

Harish Muralidhara
Dept. of Electrical and
Computer Engineering
University of Wyoming
Laramie, Wyoming - 82072
Email: hmuralid@uwyo.edu

Robert F. Kubichek
Dept. of Electrical and
Computer Engineering
University of Wyoming
Laramie, Wyoming - 82072
Email: kubichek@uwyo.edu

***Abstract*—This paper analyzes errors resulting from using free-space and Hata propagation models for localization of MANETs, in an urban environment using overlapping range maps. We conclude that when range predictions exceed actual transmission range, significant loss of resolution can occur. A modified localization technique is therefore proposed that replaces the free space model or the Hata model with COST 231 Walfisch Ikegami model, which better accounts for local terrain and lead to more accurate position estimates. Also, compared to sophisticated models like 3D raytracing, Walfisch Ikegami model is very less computationally intensive and thus can be used in ADHOC networks where resources are constrained.**

I. Introduction

A Mobile AD-HOC Network (MANET) is made up of nodes having wireless capabilities. One future scenario where MANETs could be used is in the battlefield, where soldiers would randomly deploy these wireless nodes as they moved into a new urban or suburban area. The nodes are small self-contained units with limited mobility that would organize themselves to form a network for data communication. Accurate position information is required for mobile nodes to establish the most efficient communication paths possible. For example, a mobile node that is aware of its position might be able to move away from a nearby building to achieve lower communication path loss with other nodes in the network. This is critical due to their limited battery capacity and small antenna size.

In rural areas, localization might be possible using GPS, but in urban scenarios GPS signals are often severely affected by narrow streets and tall buildings. Thus, usage of non GPS localization is highly favorable.

Several such localization techniques have been described in the literature [1], [2], [3], [4]. One methodology is the range circle overlap method, where a flying 'anchor node' periodically transmits its location, and ground stations estimate their location by finding the overlap of received range circles (see *Fig. 1*).

We begin this paper by analyzing how simple propagation models such as the free-space model (FSM) can lead to localization errors. Then we describe a modified technique that uses more accurate yet computationally less intensive propagation model such as Walfish Ikegami model (WIM) [5], [6], [7].

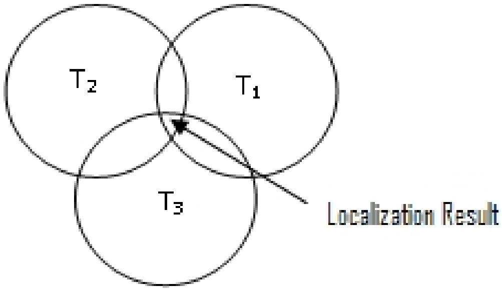


Fig. 1. Range circle overlap method for MANET localization. Figure shows 'range circles' from three transmitter locations.

II. Localization errors due to inaccurate path loss estimates

FSM is a good choice for estimating path loss between a high-flying transmitter and the ground node since no obstacles exist to block propagation. Ground nodes located in urban areas, however, would encounter signal blockage due to nearby buildings. In particular, path loss estimates from FSM and HM would become increasingly unreliable as the flying anchor node moves closer to the horizon. In this section, we investigate specific types of localization errors that can occur when using FSM or HM range circles.

First, consider two hypothetical range maps as shown in *Fig. 2*. 'True' represents the actual maximum range from a transmitter located at the center, and is illustrated by an ellipse. This shows better propagation in the N-S direction than in the E-W direction, which could be due to local changes in terrain. 'Model' represents a circular range map such as predicted by simple HM or FSM. Two types of errors are possible.

1) The probability P_m given by (1) is the probability the true receiver location is not included within the model range map so that the final location estimate is wrong. This type of error is often called 'probability of miss'

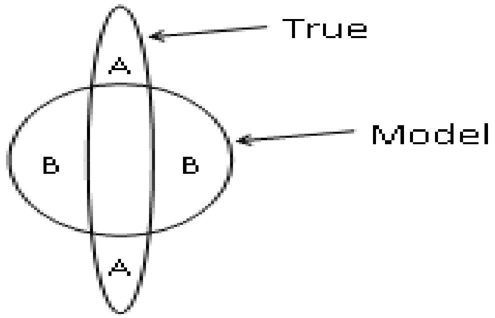


Fig. 2. Hypothetical range maps showing actual range ('True' curve) and predicted range ('Model' curve).

since the true location has been missed.

$$P_m = \frac{Area_{(Region-A)}}{Area_{(True)}}, \tag{1}$$

2) The probability P_{FA} given by (2) is the probability of 'False Alarm'. Although True represents the region where the receiver must be located, B is considered by the algorithm as a possible location. This is also referred to as a false positive. In practice, this is a less serious error since it represents a looser bound on location than we would have if we had access to the True range ellipse. For example, suppose the final receiver location is determined to be within a 1 sq-km region, and if $P_{FA} = 0.3$, it means that the search region could be reduced by 30% to 0.7 sq-km if a better propagation model were used.

$$P_{FA} = \frac{Area_{(Region-B)}}{Area_{(Model)}}, \tag{2}$$

The effect of false alarm probability with respect to building heights from the use of simpler models was carried out and the result is as plotted in *Fig. 3*. 'True' model results are based on WIM, and 'Model' is taken to be HM or FSM. As the building height increases, greater attenuation causes the True range circle to shrink and region B to increase in size. The Model circle does not depend on building height, so its area remains constant. We can see that with increasing building height P_{FA} increases. When the building height is $30m$ we can see that the total search area can be reduced by approximately 90% if WIM was used instead of FSM and by 55% if WIM was used instead of HM. We can also see that the urban version of the HM provides significantly better results than the free-space model which is obvious. This shows that the FSM has poor accuracy when there are ground obstacles and thus it should not be used for range based localization methods. Furthermore, it shows that the Hata model is useful in rural areas where building heights are less than $30m$, and then its accuracy begins to degrade.

Further analysis of the uncertainties from curves obtained from HM in an urban area was carried out. This applies to FSM as well. Consider the case when two range maps have been received by the ground node. Usually in urban areas with tall buildings, simpler propagation models like HM predict a larger range than the actual range as shown in *Fig. 4*. The inner circle represents the true range, and the outer circle could be HM or FSM range predictions. For ease of analysis, circular propagation patterns are assumed in this analysis.

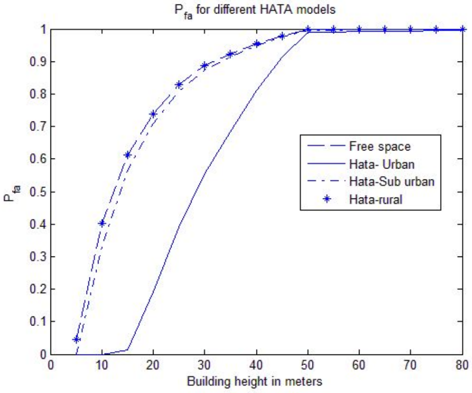


Fig. 3. False-alarm probability vs. building height for 50 m high transmitter for free-space and three versions of the Hata model compared to WIM. (Hata rural and suburban curves overlap for this case).

Here, the ground receiver is positioned in the center at the extreme range of each transmitter location. This occurs when the two 'true' range maps just touch each other. If the localization algorithm used actual range maps, there would be zero uncertainty in location since there is only a single point of overlap of the inner two circles. However, since the location algorithm is using the larger predicted range maps, the receiver position is only known to lie within the overlap area of the large circles, as shown by the shaded portion. This region of uncertainty could conceivably be reduced by increasing the distance 'd' between transmitters. Unfortunately, for this case, increasing d any further will put the receiver out of range of both transmitters. Therefore, the shaded region represents the smallest possible uncertainty region. As the ground receiver acquires additional range maps from the anchor node,

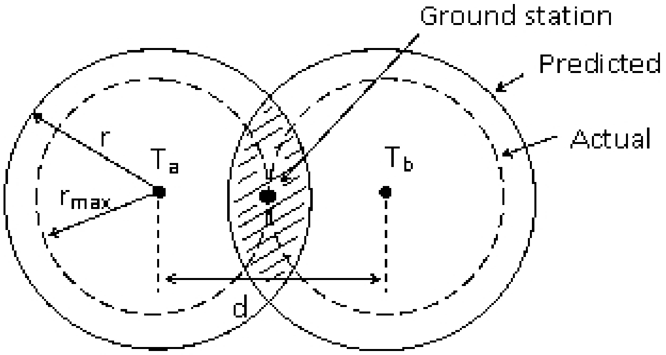


Fig. 4. Actual range circle (inner circle) and predicted range circle (outer circle). Receiver is located at maximum possible range from T_a and T_b. Overlap region represents smallest possible uncertainty for estimated receiver location.

the overlap region will decrease giving improved position estimates. However, only limited improvement is possible when range predictions are excessive. Mathematically it can be shown that the best uncertainty region cannot be less than $A_{min} = \pi(r - r_{max})^2$. This says that the localization technique should use more accurate propagation models.

Fig. 5, shows the area of uncertainty as a function of average building height. Here we assume that WIM provides true results, and analyze the effects of model error due to using the HM for range predictions. In rural areas with very low average building height, WIM predicts a longer range than Hata, resulting in the possibility that the receiver is located outside the predicted range circle (i.e., a miss). As building height increases, this type of error diminishes until average building height is about *25 m*. As building height increases above *25 m*, actual transmission range decreases below the Hata predictions resulting in increased probability of false alarm.

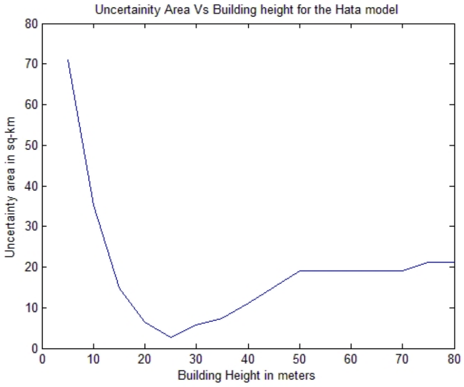


Fig. 5. Uncertainty area using Hata propagation model and two overlapping range maps. Simulation assumes transmitter height is 50 m and ground station height is 3 m.

III. Modified Range Curve Overlap Method

In this section, we investigate a localization method based on the Walfisch Ikegami model, which includes local terrain parameters such as average building height, street width, and street orientation. The resulting range contours are no longer circular, and depend on the anchor node's location and nearby terrain.

In this scheme, WIM is used to predict transmission range in all different directions from the known anchor node location. In addition to periodically broadcasting its location, the anchor node now also transmits the non-circular range map. The ground node computes its location by ANDing all the received range maps. The result is illustrated in *Fig. 6*, which shows the potential benefit of using WIM range maps in terms of more accurate position estimates in urban areas.

IV. Simulation Methodology

In order to estimate path loss based on local terrain, a modified version of WIM is used. A 10 km square terrain model is represented on an $N \times N$ grid. Each cell is described by local parameters including: average building height, average building width, and street orientation. Working outward from the transmitter location, total path loss is estimated as the sum of loss to the previous cell and loss in the current cell. Although modified WIM results are not as accurate as more sophisticated ray-tracing techniques, low computational cost for WIM make it feasible to implement on low-power computer that might be available on a drone aircraft.

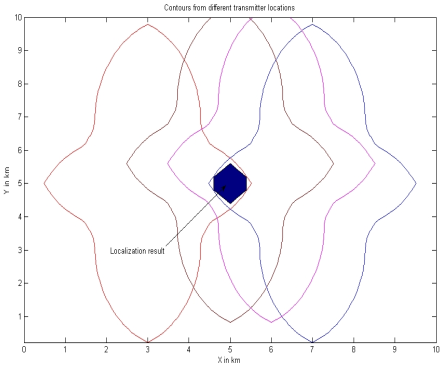


Fig. 6. Overlapping WIM-based range maps from four transmitter locations, with estimated ground station location shaded in center.

A simple example is shown to illustrate the range map method using WIM and a low-flying anchor node. This example represents a 10 km by 10 km urban region (using a 200x200 grid), dominated by small buildings (average height is 15 m) and roads running N-S, E-W. A second region (green block), interior to the first region, consists of taller buildings (average height is 40 m) with roads running at a 30 degree angle to the horizontal. The drone flies a path starting at the lower left, and arcs in a clockwise direction over the tall building zone as shown in *Fig. 7*. The transmitter which operates at $922MHz$ and at an altitude of $50m$, just above the tall buildings is used. The receiver antenna is assumed to be 3 meters high, corresponding approximately to the mobile node concept. The drone transmits its position and range maps at approximately 1 km intervals. We can see the range contours from different transmitter locations in *Fig. 8*. As the transmitter moves through different types of terrain we can see the change in the shape and size of the range curve. Attenuation is large in the regions of tall buildings, which is evident from the smaller size of the range curve. Also it can be seen that the orientation of the curve changes with the change in orientation of the street.

Ground nodes compute their location by ANDing all received range maps. The localization result is usually a random shaped region as shown in *Fig. 6*.It is useful to measure how good the localization estimate is for any given receiver location. Since location uncertainty is related to the area of the overlap region, one possibility is to compute an Equivalent Circle Radius (ECR) of the random shaped region is calculated by using (3). That is, find the radius of a circle that has the

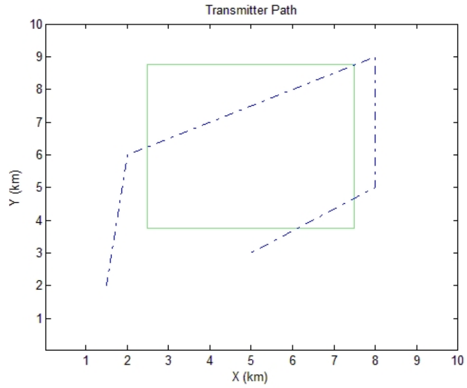


Fig. 7. Simple urban model with central area (depicted by green square) of tall buildings.Anchor node path is indicated by a dashed line.

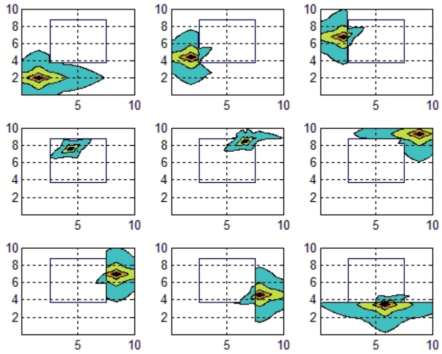


Fig. 8. Snapshots of non circular loss contours showing that attenuation in taller building parts of the city is much greater than that for parts of the city with shorter buildings. Also seen is lower attenuation along the direction of the streets, which is different for the two regions. Maximum loss shown is 140 dB.

same area as the resultant localization region. Thus, smaller ECR which translates to smaller localization region implies greater precision in the result. Continuing with the above example, the ECR is computed for each x,y location on the grid, and displayed in *Fig. 9*. Bright red colors indicate lower ECR values which indicates greater precision in localization result and cooler blue regions indicate higher ECR values which correspond to less precision in localization result. Thus in this case we can see that even in the region of tall buildings good precision in the localization result can be obtained.

$$ECR(x, y) = \sqrt{\frac{A_{(Result)}}{\pi}}, \tag{3}$$

V. Conclusion

Analysis of position errors resulting from using FSM or Hata are investigated, with the conclusion that when range predictions exceed actual transmission range, significant loss of resolution can occur in location estimates. A modified localization technique is proposed that replaces FSM and HM with WIM, that better accounts for local terrain and lead to more accurate position estimates while being less computationally intensive.

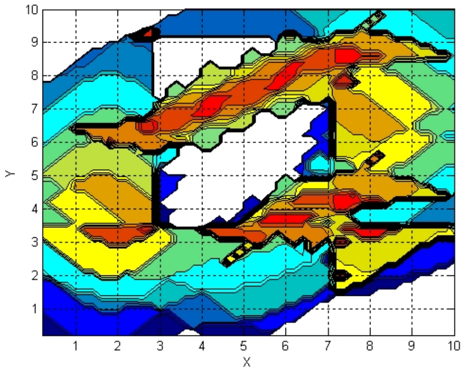


Fig. 9. Equivalent Circle Radius, ECR(x,y). Bright colors represent higher precision localization result(smaller ECR), while cooler colors correspond to less precision.

References

[1] D.P. Denson, Suresh Muknahallipatna and Margareta Stefanovic, "Performance of mobile ad hoc networks with random and predetermined mobility patterns using opnet: Preliminary results," in *Proceedings of IEEE conference on Local Computer Networks (LCN)*, Oct. 2007, pp. 781–783.

[2] D.P Denson, Suresh Muknahallipatna and Margareta Stefanovic, "A preliminary model and performance analysis for mobile group formation," in *Proceedings of the OPNET 2007*, Aug. 2007.

[3] H. B. Hüseyin Akcan, Vassil Kriakov and A. Delis, "GPS-Free Node Localization in Mobile Wireless Sensor Networks," in *Proceedings of MobiDE 2006*, Jun. 2006.

[4] D. E. Nirupama Bulusu, John Heidemann, "GPS-less Low Cost Outdoor Localization For Very Small Devices," *IEEE Personal Commun. Mag.*, Oct. 2000.

[5] J. Walfisch and H. L. Bertoni, "A Theoretical Model of UHF Propagation in Urban Environments," *IEEE Trans. Antennas Propag.*, pp. 1788–1796, Dec. 1988.

[6] L. Correia, "A view of the COST 231-Bertoni-Ikegami model," in *Proceedings of the EUcap 2009*, Mar. 2009.

[7] www.ee.bilkent.edu.tr/~microwave/programs/wireless/prop/costWI.htm.

[8] T. S. Rappaport, *Wireless Communications- Principles and Practice*, 2nd ed. Prentice Hall, 2002.

[9] J. Lee and L. Miller, *CDMA Systems Engineering Handbook*. Boston Artech House, 1998.

[10] S. planning team, "Investigation of Modified Hata Propagation Models," Australian Communications Authority, Tech. Rep., 2001.

[11] Dongsoo Har, Alix M.Watson and Anthony G. Chadney, "Comment on Diffraction Loss of Rooftop-to-Street in COST 231-Walfisch-Ikegami Model," *IEEE Trans. Veh. Technol.*, pp. 1451–1452, Sep. 1999.

[12] E. O. Rozal and E. G. Pelaes, "Statistical Adjustment of Walfisch-Ikegami Model based in Urban Propagation Measurements ," in *Proceedings of the International Microwave and Optoelectronics Conference*, Oct. 2007, pp. 584–588.

A Dynamic Hybrid Service Overlay Network for Service Compositions

Yousif Al Ridhawi[1], Gajaruban Kandavanam[1], and Ahmed Karmouch[1]
[1]Department of Electrical and Computer Engineering, University of Ottawa, Ottawa, Ontario, Canada

Abstract - *This paper presents a dynamic hybrid Service Overlay Network (SON) for service composition for multimedia delivery in mobile networks. The overlay considers the nomadic nature of mobile nodes in the decision for node placement within the overlay as well as the types of services provided and expected Quality of Service (QoS) levels. Dynamic re-organization of the overlay reflects changes in stability and QoS provided by the overlay nodes, thus promoting nodes providing high QoS towards the center and demoting unstable nodes towards the edge of the network. This paper also presents a mechanism to efficiently search for services based on the type of service and the required QoS to meet client's expectations.*

Keywords: hybrid overlay, SSON, QoS, service composition

1 Introduction

Delivering customized media by providing advanced services through service composition and forming Service Specific Overlay Networks (SSONs) requires dynamic construction of service-specific overlays without prior knowledge of the underlying physical network. These media services include dropping frames to meet QoS constraints, caching media content before being viewed and converting media formats to meet the clients' needs. The process becomes more challenging in mobile networks where nodes are not fixed, and may arrive and depart from the network without warning thus increasing the topology's instability.

Establishing SSONs involves discovering network-side nodes that support required media processing capabilities, deciding which of the discovered nodes should be included, and configuring overlay nodes [1]. Using a registry-like solution for information has many limitations, such as scalability, difficulty and cost involved in keeping an up-to-date registry in a dynamic environment, and single point of failure. The idea behind SSONs emerged from SONs [2]; application-layer logical networks formed by links between service nodes with matching inputs and outputs. The initial proposal faced many problems with how complex service interactions should be dealt with in mobile environments. In the Ambient Networks architecture [3], SMART (Smart Multimedia Routing and Transport Architecture) [4], was proposed for guiding media flow through specialized network nodes to make use of their ability to cache, transcode, synchronize multimedia data or any other service through the use of overlay networks in an Overlay Control Space (OCS). OCS is responsible for selecting the necessary media processing nodes and establishing a service-specific end-to-end overlay network for media delivery.

SSONs are composed of three types of nodes: Media Ports (MP), Media Servers (MS), and Media Clients (MC). MPs are nodes supporting special services such as caching, synchronizing and media transformation. MCs are the actual clients that request media services. MSs are sources of media flows from which MCs receive the desired media services. SSONs are built to provide composite services based on technical, QoS, and Quality of Experience (QoE) requirements of MCs for basic services originating from MSs. This is done through a process consisting of a set of sub-processes between the MC and the MS. Each sub-process is performed by a service provided by an MP. This composition can be specified during design time or discovered at runtime. The former has many limitations especially in mobile environments, where service nodes are in constant movement and would face recovery and reconfiguration difficulties when there are node failures or visible degradation in the QoS. The latter approach however can maintain high levels of QoS during execution of composite services and has the ability to replace component services for changes in the requirements.

The contribution of this paper is to present a multi-layered approach to building SSONs based on predefined QoS. This is done by establishing a lower hybrid overlay network connecting MP nodes. In this proposed overlay, nodes of similar services are clustered in well-known regions. Additionally, the nodes are organized such that those with higher QoS are located closer to the center of the MP space (hybrid overlay) while those of lower quality are dispersed further away from the center. This method improves response time, reduces overhead and number of overlay hop counts required to provide the requested services. This is done in a way the network's resiliency to dynamic mobile topologies is improved. In such a network, overlay nodes, and their respective underlay mobile service nodes are constantly in motion with unpredicted arrivals and departures.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 presents an overview of a proposed hybrid MP overlay network. Section 4 describes a score-based overlay management method. Section 5 our circular MP search for service compositions. Section 6 discusses our simulation and system evaluation. Finally, section 7 provides our conclusion and future work.

2 Related Work

Gou et al. [5] presented a hybrid Peer-to-Peer (P2P) Overlay network for highly efficient searches by dividing the overlay into structured super peers connected through a ring-like structure, and standard unstructured peers forming an optimal tree connected to a super peer. Queries in the hybrid structure have a maximum number of hops after which the query is dropped. A major drawback of this method is the inverse relationship between the overlay's re-configurability and resource availability seen in the breadth and depth of the optimal tree each of these super nodes forms. In real networks, resource limitations in super peers require placing limitations on the number of direct regular peers connected as well as how deep trees can grow. Consequently, a failure of a high-level peer in a tree may require costly reconfigurations. Additionally, the proposed work did not illustrate how failures should be handled to minimize information loss, or message overhead for network reconfiguration and stabilization.

Han et al. [6] presented a solution to achieve a sustained throughput and fast data dissemination rate from any data source with a goal of minimizing overhead even under frequent node failures through a Churn-Resilient Protocol (CRP). The Chord-like structure proposed dynamic fingers among nodes satisfying certain proximity properties. From the CRP overlay, a shared spanning tree is extracted with a minimized latency. In the first hops of the tree, scope-flooding is used and tree traversing in the remaining hops.

The authors correctly indicated tree-based dissemination methods are vulnerable to node failures. Hence they proposed the use of churn resilient overlay nodes to achieve good fault resilience. New nodes form finger links to other nodes using proximity measured through latency and capacity thus placing high-capacity nodes at high positions in the delivery tree. The purpose of choosing a Chord-like structure is to overcome the vulnerability of tree structures to node failures. Any node failure can be temporarily bypassed through links neighboring nodes have to other Chord nodes in the overlay. Hence if any node fails with probability $e^{-a(a>1)}$, the structure is connected with a probability higher than $1\text{-}n^{1\text{-}a}$- A link between any two nodes is given a network proximity weight measured by the link's latency and capacity difference between the two nodes. Although the solution improved traditional Chord networks, it's reliance on structured P2P overlay meant data placement and topologies are tightly controlled. Although structured P2P solutions have efficient data access and reduced number of search hops, processing fuzzy queries is challenging. In addition, the stabilization costs associated with the arrival and departure of nodes exceeds the original cost of Chord.

Chord was presented as a distributed lookup protocol that efficiently located nodes with particular data [7]. This was implemented through DHTs that associated a key with each data item. Our proposed lower overlay utilizes Chord structures for their scalable join and leave process, ability to provide an organized structure for fast search of nodes within predetermined regions of the overlay, and ability to provide correct lookup results even during recovery.

To reduce search time for available services in P2P systems, Li et al. [8] illustrated a super-chunk-based fast search network (SURFNet) for prompt Video on Demand (VoD) delivery. Their design had some limitations; need for timely content location and delivery, and limited cache size. Therefore, multimedia files were divided into chunks and grouped into super-chunks. Peers formed a structured overlay based on the super-chunks they held. This structure consisted of an AVL tree layer of stable peers and a holder chain layer of nodes holding the same super-chunk. All remainder peers formed an unstructured overlay that periodically exchanged chunk-level data availability information with neighbor nodes.

Makaya et al. [9] focused on user-generated contents and user-generated services in Next-Generation SONs (NGSON). The objective was to support efficient delivery of services while being end-user-centric by allowing network operators and service providers to have service management, control, creation, composition, and execution. However, the approach was highly centralized requiring an NGSON controller to act as an overlay manager. The controller formed a primary point of contact for end-users and service nodes responsible for service composition, orchestrating, chaining, service node discovery and monitoring, overlay management for self-organization and sessions and path establishment between service nodes. However, the approach faced the problem of a central point of failure, and lack of scalability. Our work focuses on finding a decentralized solution for service composition that maximizes the benefits and minimizes the limitations found in centralized and decentralized solutions.

3 Media Port Hybrid Lower Overlay

We focus our work on dividing the Overlay Network layer in the SMART SSON architecture into two layers; a lower overlay and the SSONs built over it. The lower overlay constitutes a hybrid overlay structure whose node organization and links reflect the types of services provided by the nodes, their stability, and QoS associated with services provided by them. The lower overlay network is composed of all MP nodes present that can be utilized during formation of a service composition between the MS and the MC. MPs are chosen to form SSONs reflecting the MC's needs in terms of QoS and expected QoE. By using a lower overlay network, we are able to organize MPs prior to any service requests from MCs for fast MP search and SSON construction. Nodes are assumed to be mobile, hence failures are expected.

Promoting use of hybrid overlay stems from the following:

- Searching for component services to form a composite service according to an expected level of QoS may require a large number of search messages exchange between nodes.
- When forming SSONs, it is beneficial to find MPs in the same path from MS to MC, provided required QoS for the service is available and the MPs have low hopcount reflecting low number of hops in the network level thus reducing delay.

3.1 Justifying the use of the Hybrid Overlay

Our proposed hybrid structure forming our lower overlay network is used to overcome limitations seen in both structured and unstructured overlay networks. Unstructured overlay networks are easily constructed with arbitrary overlay links. However, there is notable limitation when a peer is interested in acquiring some data. In this case, the query needs to be flooded in the network to find a node holding the desired data. The less popular the service clients are interested in acquiring, the less likely it will be found in a nearby node to the requestor. Consequently, the number of messages exchanged would consume significant bandwidth in a wireless network besides the operational overhead. In addition, since no correlation exists between service requested and corresponding MP, it is highly unlikely that a search is successful. Structured overlay networks overcome many of the limitations of unstructured networks by maintaining DHTs and distributing content in the network. The disadvantage of structured overlays is the large number of messages that need to be exchanged to stabilize the network, maintain existing links, and heal links to failing nodes that drop from the network unexpectedly. This rigidity is highly unfavorable in dynamic mobile environments where nodes leave the network without warning and new nodes are constantly arriving.

Hence, we propose the use of a hybrid overlay structure that maximizes the advantages of both structured and unstructured overlays and minimizes their disadvantages. Our overlay consists of several hierarchical layers of Chord-like rings and unstructured leaf connections. Each node in the network can be one of two types; a *major node* or a *minor node*. A Major node is any node capable of joining one of the Chord structures present within the network. These nodes should have the following characteristics:

- Have a long life span and are highly likely to remain in the network. This is directly related to the node's arrival time.
- Have adequate level of available resources to support links to nodes present within the same Chord structure as theirs, be the managing node for further sub-Chord structures and be the parent for other leaf (Minor) nodes.
- Have popular services frequently requested during service compositions.

Major nodes are divided into two types;

- *Normal Major Node (NM)*: a major node that may have any number of child leaf nodes, Chord nodes, and neighboring nodes. NM nodes do not link directly to higher Chord nodes.
- *Link Major Node (LM)*:a major node having the same characteristics as NM nodes with the addition of a link to a higher level Chord node.

A *minor node (leaf)* is any node linked to only a single node that is its parent and shares the following characteristics:

- Newly arrived nodes that have not built stability credentials. We assume they are short-lasting service nodes.
- Nodes that do not have the resources to support links to other neighboring nodes to be part of a Chord ring.
- Have unpopular services rarely used in compositions.

The formation of the hybrid overlay network is performed in three stages: initialization, sub-Chord formation and extension, and promotion-based stabilization and overlay management. Initialization stage occurs when the first '*m*' nodes start arriving forming the central Chord structure. Nodes continue to join Chord until C_{Thresh} threshold is reached; a network dependent variable related to the expected number of nodes in the network, available bandwidth and other criteria. Arriving nodes send a join request message to a nearby node chosen randomly and await a response message that includes the node's successors and finger links. If a response is received then the node joins the Chord Structure. Otherwise, the process is repeated and the join message is sent to another Chord node until an acceptance is received. Once a C_{Thresh} number of nodes have joined, the central Chord network is formed and an arriving node must request a join from one of existing Chord nodes. Receiving an acceptance, the node is added as a *minor node* to the responding node.

4 Score-based Overlay Management

To maintain a stable Chord network, nodes exchange messages to monitor departure of neighboring nodes and fix broken links. To maintain the dynamic form of our hybrid overlay we have added a new metric that nodes must inform their neighbors of, reflecting node's stability and supported QoS. The score for each node can be evaluated as follows:

$$NodeScore = [w_{arriv}(T_{curr} - T_{arriv}) + w_{serv}\sum_{i=1}^{n} P_i (\sum_{j=1}^{m} w_{Qual_j} Q_j)] \qquad (1)$$

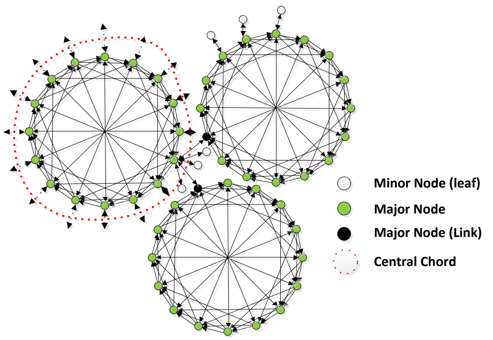


Figure 1. Proposed Hybrid Lower Overlay Network.

Where, w_{arriv} is the weight given to the node's total time within the system since the last failure. The longer a node remains in the system, the higher is its score, recognizing it as a stable node. Each MP can provide up to 'n' services ranging in P_i popularity levels reflecting the frequency with which MCs request these services. Additionally, each service can have a number of QoS criteria Q_j, such as delay and accuracy. Each of these is given a weight reflecting its importance, given by w_{Qualj}. Each Chord ring in the overlay has a minimum threshold, $MinThresh_i$. When a node's score falls below $MinThresh_i$, it is demoted to a lower Chord ring or becomes a *minor node* to neighboring Chord node. Similarly, if a *minor node*'s score exceeds $MinThresh_i$ then it can be promoted by a *major node* to join the Chord ring. Additionally, if a MP's score in any Chord ring exceeds a maximum threshold $MaxThresh_i$ set for that ring, then the node can be promoted to a higher level ring. In the event that no higher level Chord ring exists, the node will begin forming a new higher level Chord ring to become the new center of the lower overlay network. Fig. 1 illustrates a simplified view of hybrid lower overlay network showing the presence of two main types of nodes in the system; *major nodes* and *minor nodes*.

4.1 Overlay Join

When an MP arrives in the network, a join request must be sent to one of the *major nodes*. The receiving node evaluates the score of the arriving node to determine if an acceptance should be sent. If the joining node's score falls below the structure's *MinThresh*, then the request is rejected. Prior to rejection, the receiving node determines whether a lower Chord structure with a lower threshold exists. If such a structure exists, the request is forwarded to the *link major node* of that lower Chord. The *link node* may decide to accept the join request, forward it to one of its neighbor nodes or pass it to a lower Chord structure. If there is no lower level node to forward the request to, the receiving node may accept incoming node as a *minor node* or send a join rejection. If the node's score falls between the Chord structure's *Max*- and *Min-threshold*, then the node may either accept the joining node as one of its *minor nodes* or allow it to join as a neighbor within the same Chord structure. Finally if the node's score exceeds the structure's *MaxThresh*, the join request may be accepted and the node added as a *minor node* or as a neighbor in the Chord structure. Additionally, the request may be forwarded to a higher leveled Chord structure closer to the center of the MP space. The join process must be done in a time frame, $t_{response}$. Each node can only sustain a connection to a certain number of neighboring *major* and *minor nodes* depending on the available resources and QoS level that permits maintaining a high node score. Once the maximum number of nodes is reached, join requests received are rejected, forwarded to some other node or accepted after making modifications to its existing links to *minor nodes*.

4.2 Minor Nodes Reorganization

One possible approach to accepting new node links or reducing resource drainage on a *major node* is through *minor node* reorganization. The process is triggered by a *major node* contacting its *minor nodes* that are not become part of a lower level Chord structure and determining the node with the highest score. A *FormNewPartition* call is sent to that node informing it to start forming a new Chord structure. This node will become a *major link node* linking the parent node with the new lower level Chord structure and will maintain its status as a *minor node* to the parent node. A *JoinNewPartition* call message is sent to all other *minor nodes* informing them of joining the new Chord structure. The message includes the identity of the node they must contact; i.e. the sender node of the *FormNewParition* call message. The next step involves sending a join acceptance message to the join request sender node. The *major node* continues to use this same process of grouping its single *minor nodes* into lower Chord structures until all *minor nodes* have become *major link nodes* and is unable to accept any new join requests.

4.3 Handling Node Failures

4.3.1 Failure of Minor Nodes

Since each *minor node* is only linked to its parent *major node*, failure of *minor nodes* has no negative effect on the overlay structure. Affected *major node* simply removes any reference in its routing table to the failed node and can now accept an additional join request from an arriving MP.

4.3.2 Failure of Major Link Nodes

Failure of a *link node* disconnects its personal Chord structure from the parent *major node* and breaks the uniform shape of the Chord structure. Reconnecting the network has two phases. If the child node is a *major link node*, then failure of the parent node means separation of the lower level Chord structure and all attached sub-structures from the higher Chord structure that contained the parent node. In the first phase, the Chord structure detects failure of one of its nodes. All affected Chord nodes reconfigure their links to restructure the Chord network and form all links necessary to stabilize the network. In the second phase, from the level of the highest detached Chord structure, one node must be chosen as the new Major Link node. This node will connect with the same major node the old link node was connected. The process of choosing a new *link node* is done by a search process performed to find the node with the highest score. During lifetime of a Chord network, nodes periodically exchange stabilization messages to update successor, predecessor, and finger table links. We utilize these messages and expand them to include updates for node scores. Each node part of the optimal Chord structure maintains a table that includes node scores for its immediate successor, predecessor, and two finger links. Therefore, to discover the node with the highest score it would take at most *log n* messages. This process starts at any random node in the reconnected Chord structure. The node evaluates from the four nodes it is connected to as well as itself to find which of these nodes has the highest score. A message is sent to the node's successor containing identity of the chosen node and the identity of the rejected ones. This node then performs the same process and passes the message to its immediate successor. This process continues until all nodes within the Chord structure or one of the nodes connected to it is informed. The node with the highest score is then informed of being chosen as the new *major link node*. The chosen node then sends a join request message to the previous *link node*'s parent to reconnect the detached Chord and its sub-structures.

5 Circular Media Port Space

In a work presented in [10], search for MPs to form a composition was performed based on three assumptions: The location of the MC is known, the location of the MS is known, and the search for MPs is performed in the direction emanating from the MC towards the MS only. Using this, any node receiving a query will forward the query to local nodes only if it is within *α* degree. The angle *α* represents the maximum search scope for possible component services. This value depends on the locations of MC and MS. This approach, in general, avoids sending queries to regions of the network where answers are not likely to be found. However, in wireless mobile networks, the dynamic movement of nodes at the network layer may result in an unequal distribution of MPs at the overlay level. We may end up with high density regions where large numbers of diverse MPs exist while other regions may lack any MPs or have them sparsely located. In the latter, it is highly likely that α must be continuously increased until all component services are found.

In our work, we overcome the limitations of the previous solution using two methods. The first is by providing a hybrid lower overlay structure discussed earlier. The second is by dynamic node promotion and demotion according to Eqn. 1. These two methods result in a circular MP space that is further re-organized based on the types of services provided. As a result, nodes arriving in the network must first determine the overlay region they must join. We assume the following:

- The center of the MP space is a known region that can be located by all nodes joining the network.
- Angles *α* and *β* representing the start and end borders, respectively, of each service-type sector are also well known by all nodes joining the network. *α* is measured from the absolute 0 degree angle pointing to a relative direction accepted by all nodes in MP space.

Using these assumptions, any search for a specific type of service forming a composition can be done by searching within specified regions in the MP space. A simple example may involve a video format conversion service that converts from MP4 to AVI. The MP space sector providing this service is to be located. For e.g., starting at *α=47°* away from the absolute *0°* angle of the MP space, with a total angular range of *β=35°*. This region represents the total space within which this type of service may exist regardless of the QoS required to form the composition and meet the MC's QoE or the stability of the node. Stability in our case refers to the length of time the node has been present in the network and thus the length of time this node is likely to remain in the future. However, to further reduce the physical search area for potential component services we have indicated that the dynamic promotion and demotion of overlay nodes in the lower overlay network results in movement of nodes with higher scores towards the center of the network. Therefore, the nodes with higher stability, QoS, available resources, etc. are generally located closer to the center of the MP space.

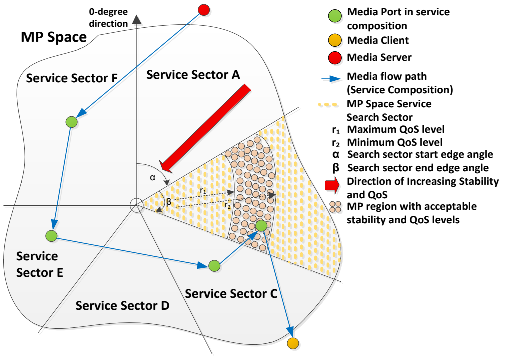

0-degree direction
MP Space
Service Sector F
Service Sector A
α
β
r1
r2
Service Sector E
Service Sector D
Service Sector C
Media Port in service composition
Media Client
Media Server
Media flow path (Service Composition)
MP Space Service Search Sector
r1 Maximum QoS level
r2 Minimum QoS level
α Search sector start edge angle
β Search sector end edge angle
Direction of Increasing Stability and QoS
MP region with acceptable stability and QoS levels

Figure 2. Potential MP search area in circle-based sectors.

We can utilize this overlay organization to search for nodes that meet the MC's service quality. This is done by finding a QoS upper and lower thresholds which when accumulated for each component service in the composition would maintain an acceptable level of QoS for the MC. In Fig. 2, *r_2* represents the maximum distance away from the network's center where a node having the minimum acceptable level of stability and QoS can be chosen to perform the component service required for the composition. *r_1* represents the distance away from the center of the network where nodes with the maximum level of stability and QoS exist. These nodes meet the minimum QoS requirements of the MC's service composition. Including nodes in the region with a radius less than *r_1* is possible; however, it would mean a lower level of resource utilization for the chosen node. From the above, we can conclude that the overlay's geographical area searchable for a possible MP in a composition is given by the following equation:

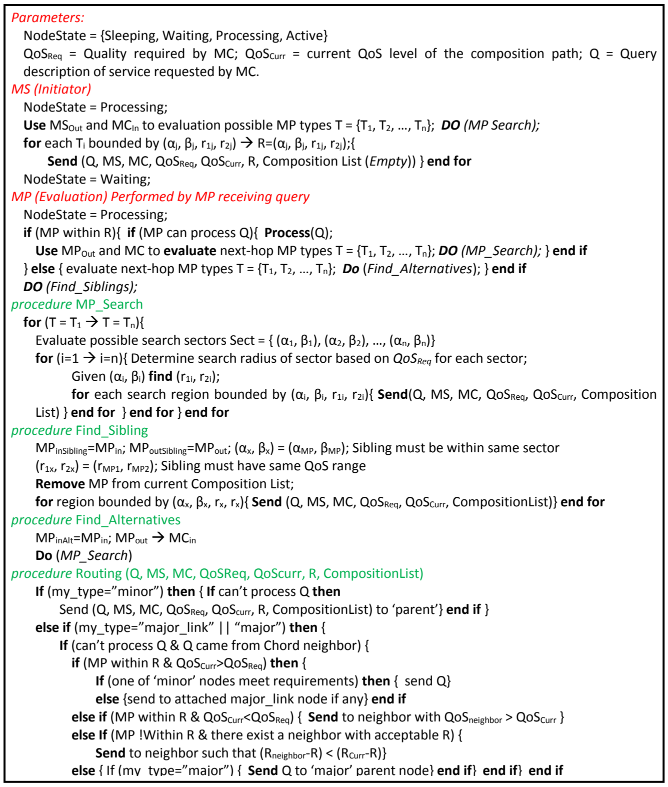
$$SearchArea = \frac{\pi\beta}{360}(r_2^2 - r_1^2) \qquad (2)$$

Where *β* represent the component service's sector angle, *r_2* represents the area's border with lowest acceptable level of QoS, and *r_1* represents the area's border with highest expected QoS level. Any node receiving a query can then be added to a composition if following conditions from Eqn. 2 can be met:

$$(\alpha \le \theta \le (\alpha + \beta)) \wedge (D_{Low} \le D_{MP} \le D_{High}) \qquad (3)$$

Where *Θ* represents the MP's angle relative to the absolute 0° angle known by all nodes and D_{Low} represents the distance of any point on the arch *r_1* distance away from the center of the MP space given by the following formula:

$$D_{Low} = \sqrt{(x_l - x_c)^2 + (y_l - y_c)^2} \tag{4}$$



```
Parameters:
  NodeState = {Sleeping, Waiting, Processing, Active}
  $QoS_{Req}$ = Quality required by MC; $QoS_{Curr}$ = current QoS level of the composition path; Q = Query
  description of service requested by MC.
MS (Initiator)
  NodeState = Processing;
  Use $MS_{Out}$ and $MC_{In}$ to evaluation possible MP types T = {$T_1$, $T_2$, …, $T_n$};  DO (MP Search);
  for each $T_i$ bounded by ($\alpha_j$, $\beta_j$, $r_{1j}$, $r_{2j}$) → R=($\alpha_j$, $\beta_j$, $r_{1j}$, $r_{2j}$);{
      Send (Q, MS, MC, $QoS_{Req}$, $QoS_{Curr}$, R, Composition List (Empty)) } end for
  NodeState = Waiting;
MP (Evaluation) Performed by MP receiving query
  NodeState = Processing;
  if (MP within R){  if (MP can process Q){  Process(Q);
    Use $MP_{Out}$ and MC to evaluate next-hop MP types T = {$T_1$, $T_2$, …, $T_n$}; DO (MP_Search); } end if
  } else { evaluate next-hop MP types T = {$T_1$, $T_2$, …, $T_n$};  Do (Find_Alternatives); } end if
  DO (Find_Siblings);
procedure MP_Search
  for (T = $T_1$ → T = $T_n$){
    Evaluate possible search sectors Sect = { ($\alpha_1$, $\beta_1$), ($\alpha_2$, $\beta_2$), …, ($\alpha_n$, $\beta_n$)}
    for (i=1 → i=n){ Determine search radius of sector based on $QoS_{Req}$ for each sector;
          Given ($\alpha_i$, $\beta_i$) find ($r_{1i}$, $r_{2i}$);
          for each search region bounded by ($\alpha_i$, $\beta_i$, $r_{1i}$, $r_{2i}$){ Send(Q, MS, MC, $QoS_{Req}$, $QoS_{Curr}$, Composition
    List) } end for  } end for } end for
procedure Find_Sibling
    $MP_{inSibling}$=$MP_{in}$; $MP_{outSibling}$=$MP_{out}$; ($\alpha_x$, $\beta_x$) = ($\alpha_{MP}$, $\beta_{MP}$); Sibling must be within same sector
    ($r_{1x}$, $r_{2x}$) = ($r_{MP1}$, $r_{MP2}$); Sibling must have same QoS range
    Remove MP from current Composition List;
    for region bounded by ($\alpha_x$, $\beta_x$, $r_x$, $r_x$){ Send (Q, MS, MC, $QoS_{Req}$, $QoS_{Curr}$, CompositionList)} end for
procedure Find_Alternatives
    $MP_{inAlt}$=$MP_{in}$; $MP_{out}$ → $MC_{in}$
    Do (MP_Search)
procedure Routing (Q, MS, MC, QoSReq, QoScurr, R, CompositionList)
    If (my_type=”minor”) then { If can’t process Q then
        Send (Q, MS, MC, $QoS_{Req}$, $QoS_{curr}$, R, CompositionList) to ‘parent’} end if }
    else if (my_type=”major_link” || “major”) then {
        If (can’t process Q & Q came from Chord neighbor) {
          if (MP within R & $QoS_{Curr}$>$QoS_{Req}$) then {
              If (one of ‘minor’ nodes meet requirements) then {  send Q}
              else {send to attached major_link node if any} end if
          else if (MP within R & $QoS_{Curr}$<$QoS_{Req}$) {  Send to neighbor with $QoS_{neighbor}$ > $QoS_{Curr}$ }
          else If (MP !Within R & there exist a neighbor with acceptable R) {
              Send to neighbor such that ($R_{neighbor}$-R) < ($R_{Curr}$-R)}
          else { If (my_type=”major”) {  Send Q to ‘major’ parent node} end if}  end if}  end if
```

Algorithm1. Path formation.

Where x_l and y_l are the distances on the x- and y-axis of any point on the lower arch, and where xc and yc are the (0,0) point representing the MP space's center. The same can be applied to D_{High} by replacing x_l and y_l with x_h and y_h respectively. Using the above approach, we can rapidly decrease the search time for component services to form our composition and reduce the number of MPs where queries are sent. Using our hierarchical dynamic hybrid lower overlay, the service sector angle of search is reduced to an area guaranteed to contain the type of service required for a given composition. In addition, our use of node scores for node positioning in the overlay, the relative distance of required MPs can be utilized to further decrease number of potential nodes in a composition. Refer to algorithm 1 for the path formation process from the MS to the MC.

6 Performance Evaluation

In this section, we present the evaluations we did on the proposed lower overlay network structure using the C++-based OMNET++ [11] discrete event network simulator. This simulator permits the design of simulation models utilizing hierarchical architecture of a system module and its various sub-modules communicating via messages, gates, and links.

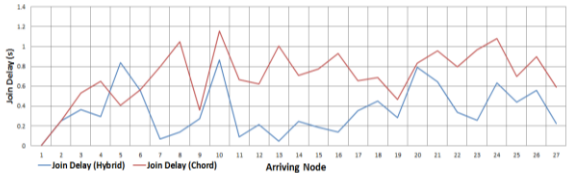
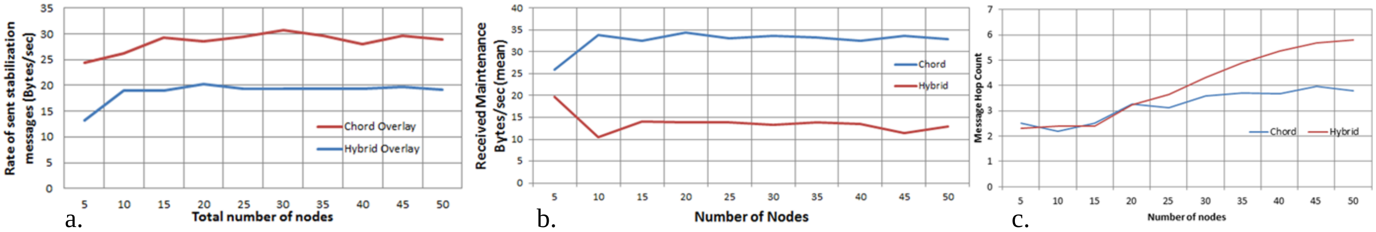


Figure 3. Join delay experienced by arriving nodes.

An open-source overlay and P2P network simulator OverSim [12] for OMNET++ was used to model Chord overlay networks and we extended the simulator to perform our proposed lower-overlay hybrid network simulation for the MP space. In our tests, nodes were placed in a 2-dimensional Euclidean space where delay between any two nodes is assumed proportional to the distance between them. In addition, a lifetime based churn model using Exponential distribution was used to model arrival and departure of nodes in the network. A sample application program of simple request-response type messages was used utilizing a full-recursive KBR protocol. KBR encapsulates messages and forwards them to the next hop according to the local routing table of a node closest to the node with the destination key. The message is then forwarded recursively back to originator.

In the simulation, control messages' packet sizes depended on the type of message exchanged, given the base message length of 32 bytes for join request messages. Each node was permitted a maximum of two join retries with an interval of 10 seconds before a join is determined as a failure. To maintain updated KBR. application test messages were set to a size of 100 bytes. Messages were generated with an interval of 60 seconds and a failure latency of 10 seconds. Chord stabilization messages were set with the following intervals: Chord finger reconfiguration with 120 seconds, predecessor update with 5 seconds, and stabilization with 20 seconds. Implementing the proposed hybrid overlay required modification to the routing protocol used by each node, and introducing new messages for *major* and *minor node* link stabilization and update, as well as link reconfiguration to form new low-level Chord structures and/or *minor nodes*.

In the first scenario, two simulation tests determined the average delay experienced by each node to join the Chord network compared with the hybrid overlay. An initially empty network was used with 10 sec. node inter-arrival rate. Results shown in Fig. 3 illustrate a visible difference in delay



a.

b.

c.

Figure 4. a. Mean rate of stabilization messages, b. Maintenance bytes received, c. Message hop count experienced by nodes in the two networks. Joining a Chord structure with *N* nodes and *K* keys requires *O(K/N)* keys to change hands and may result in some delay before a join is completed. The same applies in the hybrid overlay, however, the presence of the leaf-like unstructured extension in the proposed overlay results in a significant decrease in the time required for a join since only the major node is contacted to complete the join process. Fig. 4a presents the comparison of the mean rate stabilization messages sent between nodes in the two networks. In Chord, with increasing number of nodes, the number of messages needed to maintain links to the successor, predecessor, and finger links increases until it plateaus to approximately 30 Bytes/sec. With the same number of nodes, the same stabilization messages are exchanged in the hybrid overlay. However, reduced number of nodes, present within each Chord structure and the distribution of other nodes in the leaf positions indicates a decrease in the number of stabilization messages sent. Even when considering the fact that our hybrid overlay introduces two stabilization messages, *Parent_Live_Stabilization* and *Leaf_Live_Stabilization* exchanged between *minor*- and their controlling *major- nodes* respectively, the mean number of stabilization messages sent by nodes in the overlay remains well below Chord networks.

Fig. 4b illustrates the expected high rate of maintenance messages received by each Chord node to fix broken links, while the hybrid structure illustrates a reduction of approximately 62% in rate maintenance messages arrival. Given the fact that *minor nodes* only receive messages from their *major nodes*, the low rate associated with hybrid overlay is understandable. One tradeoff noticed with the hybrid overlay is the number of hops a message must travel to reach its destination. Fig. 4c illustrates presence of a slight increase in the number of hops over that of the Chord network. This is expected, because with use of hybrid network, the worst case scenario exists when the source and destination are *minor nodes* at opposite ends of the overlay in a lower-level Chord structure. Such a limitation does not exist in a Chord network; however the benefits illustrated above of the hybrid overlay outweigh this slight increase in hop count.

7 Conclusion and Future Work

In this paper, we presented a dynamic hybrid service overlay network for MP distribution in mobile networks. We illustrated the dynamic characteristics of the overlay taking node scores based on stability, services provided, and QoS into consideration and service search method using a circular MP space for fast and accurate service compositions. Simulation tests have shown significant improvements of join delays, improvements in the mean rate of stabilization messages, and the mean rate of maintenance messages sent in our hybrid overlay in comparison with Chord networks. We are currently extending our study to provide a fast, context-aware, healable, and reconfigurable service composition method based on bounded approximations of fuzzy sets over our service hybrid overlay.

8 References

[1] E. Asmare, S. Schmid, M. Brunner, “Setup and Maintenance of Overlay Networks for Multimedia Services in Mobile Environments,” in Proceedings of International Federation for Information, pp. 82-95, 2005.

[2] Z. Duan, Z. Zhang, Y.T. Hou, “Service overlay networks: SLAs, QoS, and bandwidth provisioning,” in IEEE/ACM Transactions on Networking, vol. 11, issue 6, pp. 870-882, 2003.

[3] N. Niebert, A. Schider, H. Abramowicz, G. Malmgren, J. Sachs, U. Horn, C. Prehofer, H. Karl, “Ambient Networks: An Archictecture for Communication Networks Beyond 3G,” IEEE Wireless Communications, vol.11, no.2, pp.14-22, 2004.

[4] S. Schmid, S. Herborn, J. Rey, “SMART: Intelligent Multimedia Routing and Adaptation based on Service Specific Overlay Networks,” in Proceedings of Ubiquitous services and applications, EURESCOM’05, pp. 69-77. 2005.

[5] H. Guo, J. Liu, Z. Wang, “Frequency-Aware Indexing for Peer-to-Peer On-Demand Video Streaming,” in IEEE International Conference on Communications, pp. 1-5, 2010.

[6] H. Han, J. He, C. Zuo, “A Hybrid P2P Overlay Network for Hight Efficient Search,” in 2nd IEEE International Conference on Information and Financial Engineering, pp. 241-245, 2010.

[7] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, H. Balakrishnan, “Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications,” in the ACM Special Interest Group on Data Communication, pp.1-12, 2001.

[8] Z. Li, G. Xie, K. Hwang, Z. Li, “Churn-Resilient Protocol for Massive Data Dissimenation in P2P Networks,” in IEEE Transactions on Parallel and Distributed Systems, pp. 1-8, 2009.

[9] I. Al-Oqily, A. Karmouch, “SORD: A Fault-Resilient Service Overlay for MediaPort Resource Discovery,” in IEEE Transactions on Parallel and Distributed Systems, vol. 20, no. 8, pp. 1112-1125, 2009.

[10] A. Varga, “OMNeT++ User Manual, Version 4.1,” [online report]http:// omnetpp.org/doc/omnetpp41/Manual.pdf

[11] D. Wang, C. Yeo, “Superchunk based Fast Search in P2P-VoD System,” in IEEE Conference on Global Telecommunications, pp. 1-6, 2009.

[12] I. Baumgart, B. Heep, S. Krause, “OverSim: A Flexible Overlay Network Simulation Framework,” in Proceedings of 10th IEEE Global Internet Symposium, p. 79-84, May 2007.

Delay-constrained dimensioning of WiMAX cellular networks carrying multi-profile real-time traffic

Sébastien Doirieux and Bruno Baynat
LIP6 - UPMC Paris Universitas - CNRS, Paris - France
firstname.lastname@lip6.fr

***Abstract*—This paper tackles the challenging task of developing accurate easy-to-use analytical models for performance evaluation of WiMAX networks. The need for precise fast-computing tools is of primary importance to face complex dimensioning issues of WiMAX cells. Here, we present generic Markovian models developed for the rtPS service class defined in the WiMAX standard. This service class is dedicated to real-time applications with constraints on both their bit rates and the latency of their packets (i.e., streaming). Our analytical modeling is performed in two steps. First, we focus on the connection level only considering the bit rate constraint. Then, we extend the model to the packet level to formulate the packet delay QoS parameter. The resulting models can account for multiple traffic profiles at both levels while always keeping an instantaneous resolution. The models are compared with simulations that show their accuracy. Finally, we show how our models can be used to solve dimensioning issues while avoiding days of simulations.**

Keywords: analytical modeling, WiMAX, cell dimensioning, real-time traffic

I. Introduction

One of the main candidate for 4G is WiMAX (World-wide Interoperability for Microwave Access), a broadband wireless access technology based on IEEE standard 802.16. The first operative version of IEEE 802.16 is 802.16-2004 (fixed/nomadic WiMAX) [1]. It was followed by a ratification of amendment IEEE 802.16e (mobile WiMAX) in 2005 [2]. A new standard, 802.16m, is currently under definition to provide even higher efficiency.

A great number of services such as voice, video and web are to be offered by WiMAX networks. To this aim, several service classes have been defined in the standard corresponding to specific QoS needs. The mains are UGS (Unsolicited Granted Service), BE (Best Effort) and rtPS (real-time Polling Service). UGS is designed for applications with constant bit rates and consists in a circuit use of the resource. Modeling UGS is thus easy enough as many classic models are available [9]. As for BE, it carries elastic traffic generated by web applications. In [8] we proposed models for performance evaluation of BE traffic in WiMAX cells. Finally, rtPS supports real-time applications with variable bit rates for which delay is an essential QoS requirement (streaming). As such, rtPS connections need guarantees on both their throughputs and the latency of their packets. In this paper, we focus on rtPS and provide a novel modeling approach to evaluate the packet delay.

A few WiMAX networks are already deployed but most operators are still under trial phases. As deployment is coming, the need arises for manufacturers and operators to have fast and efficient tools for network design able to account for the specific QoS needs of rtPS connections. Literature on this topic is constituted of two sets of papers: i) packet-level simulations that precisely implement system details; ii) analytical models that derive performance metrics at user-level. We focus here on the latter, as simulations require too much computation time and are thus impractical for dimensioning.

Authors of [6] provided a discrete-time model for performance analysis of UGS and rtPS traffic. They expressed through a complicated analysis the average delay of the packets, but, variations of the radio channel were not taken into account. Niyato and Hossain [12] formulated the bandwidth allocation of multiple services with different QoS requirements by using linear programming. They also proposed performance analysis, first at connection level, then, at packet level. However, the computation of the performance parameters relied on multi-dimensional Markovian models that require numerical resolutions, and thus, prevent in-depth dimensioning. In [7], an uplink scheduling policy for rtPS and ertPS services is designed using an analytical framework. The authors focused on cases where QoS needs are always respected and thus did not consider the impact of QoS degradations on the performances. Finally, several call admission controls (CAC) based on analytical models have also been proposed [10], [5]. Those papers focused on developing CACs to guarantee the QoS of each connection. As such, they too did not account for QoS degradation cases.

In this paper, we present novel analytical models for streaming traffic that take into account frame structure, precise slot sharing-based scheduling and channel quality variation of WiMAX systems. Our modeling is performed in two steps, each one corresponding to a level of traffic characterization. First, we focus on the connection level only considering the bit rate constraint. Then, we extend our models to the packet level to formulate the packet latency QoS parameter. The resulting models can account for multiple profiles for both traffic levels while keeping an instantaneous resolution.

In Section II, modeling assumptions including specific WiMAX network details are provided. Sections III and IV present our analytical models for the connection and packet levels respectively. In Section V, we validate our models through comparisons with simulations. Lastly, a dimensioning example using our models is provided in Section VI.

II. Modeling Assumptions

Our analytical models stand on several assumptions related to the system, the channel and the traffic. Most of these assumptions have already been discussed and validated in our study of BE traffic [8]. Wherever required, related details of WiMAX systems are specified and various notations are introduced.

A. System

A WiMAX time division duplex (TDD) frame comprises of slots that are the smallest unit of resource and which occupies space both in time and frequency domain. A part of the frame is used for overhead and the rest for user data. The duration T_F of this TDD frame is equal to 5 ms [2].

1) We consider a single WiMAX cell and focus on the downlink part which is a critical portion of asymmetric data traffic. Yet our models can also be used for the dimensioning of the uplink part in a similar way.
2) We assume that there is a mean number of slots available for data transmission in the downlink part of each TDD frame denoted by $\bar{N}_S$. This number is a mean value because the size of the downlink part can vary with the amount of overhead and with the downlink to uplink bandwidth ratio which can be adjusted dynamically over time.
3) We consider that only C_{max} simultaneous rtPS connections can be accepted in the cell. Also, since the buffers of the base station are limited, we assume that each connection cannot have more than P_{max} packets waiting to be transmitted. As such, any more arriving packet of a given connection will be dropped as long as at least one packet of this connection has not been transmitted.

B. Channel

One of the important features of IEEE 802.16e is link adaptation: different modulation and coding schemes (MCS) allows a dynamic adaptation of the transmission to the radio conditions. As the number of data subcarriers per slot is always the same, the number of bits carried by a slot for a given MCS is constant. The selection of appropriate MCS is carried out according to the value of signal to interference plus noise ratio (SINR). In case of outage, i.e., if the SINR is too low, no data can be transmitted without error. We denote the radio channel states as: MCS_k, $1 \leq k \leq K$, where K is the number of MCS. By extension, MCS_0 represents the outage state. The number of bits transmitted per slot by a mobile using MCS_k is denoted by m_k. For the particular case of outage, $m_0 = 0$.

WiMAX being a broadband technology, the radio link quality in these networks is highly variable. As such, the MCS used by a given mobile can change very often.

4) We assume that each mobile sends a feedback channel estimation on a frame by frame basis, and thus, the base station can change its MCS every frame. Since we consider that all mobiles have the same radio capacity, we associate a probability p_k with each coding scheme MCS_k, and assume that, at each time-step T_F, any mobile has a probability p_k to use MCS_k. Table I presents examples of MCS and their associated probabilities.

As a result, our analytical model only depends upon stationary probabilities of using the different MCS. This approach has been validated through extensive simulations considering radio channels with memory [8].

C. Traffic

The traffic model is based on the following assumptions.

5) We assume that there is a fixed number N of mobiles sharing the available bandwidth of the cell.

Note that operators find finite population models more suitable for the dimensioning of a cell since they can estimate the number of users they will have to serve in a cell.

6) Each mobile is assumed to generate an infinite length ON/OFF traffic. An ON period corresponds to an active rtPS connection while an OFF period to an idle time. Both are characterized by their duration. A connection in ON period generates packets according to a Poisson process. Obviously, a connection in OFF period does not generate any packets.
7) We assume that both ON and OFF durations are exponentially distributed. We denote by $\bar{t}_{on}$ the average duration of ON periods (in seconds) and by $\bar{t}_{off}$ the average duration of OFF periods (in seconds).

Memoryless traffic distributions are strong assumptions that have been validated by numerous theoretical results. Several works on insensitivity (see, e.g., [3], [4]) have shown (for systems fairly similar to the one we are studying) that the average performance parameters are insensitive to the distribution of ON and OFF periods. Thus, memoryless distributions are the most convenient choices to model the traffic.

8) We consider that the lengths of the packets are exponentially distributed and denote by $\bar{L}_p$ the average length of the packets (in bits).

However, our packet level modeling stays robust when considering fixed length packets as shown in Section V.

III. Connection Level Modeling

This section provides models considering the rtPS traffic of a WiMAX cell at connection level. They are based on the Engset model [9], however, their performance parameters are adapted to the specifics of WiMAX and our channel model.

A. Mono-Traffic Model

We associate to each rtPS connection a reserved bit rate called Guaranteed Bit Rate (GBR). In a first phase, no distinctions between users are made: all mobiles are considered statistically identical. As such, we assume that the N users are generating infinite-length ON/OFF traffics with the same traffic profile $(GBR, \bar{t}_{on}, \bar{t}_{off})$.

The number of slots needed at each frame by a rtPS connection to achieve its GBR varies with the MCS it uses.

In order to prevent the losses caused by outage periods, we assume that a rtPS connection is granted a slightly greater bit rate than its GBR, called Delivered Bit Rate (DBR):

$$DBR = \frac{GBR}{1 - p_0}. \tag{1}$$

If at a given frame there is enough available slots, an active connection using MCS_k receives g_k slots:

$$g_k = \frac{DBR \; T_F}{m_k}. \tag{2}$$

Obviously, no slots are allocated to a mobile in outage so $g_0 = 0$. However, if there is not enough resource in the frame, all active connections are evenly degraded.

We model this system by a continuous-time Markov chain (CTMC) where each state c, represents the total number of concurrent connections, regardless of the coding scheme they use. The maximum number of rtPS connections accepted being $C = \min(N, C_{max})$, this CTMC is thus made of $C+1$ states. This results in the famous Engset model [9] where a mobile initiates a connection with a rate $\lambda = \frac{1}{\bar{t}_{off}}$ and terminates it with a rate $\mu = \frac{1}{\bar{t}_{on}}$. By defining the traffic intensity parameter $\rho = \frac{\lambda}{\mu}$, the steady state probabilities $\pi(c)$ of having c current connections are thus derived as:

$$\pi(c) = \frac{\rho^c}{c!} \frac{N!}{(N-c)!} \pi(0), \tag{3}$$

with $\pi(0)$ obtained by normalization. The customary performance parameters are obtained as usual from the steady state probabilities. Only $\bar{X}$, the instantaneous throughput per connection and $\bar{U}$, the average utilization of the frame need to be adapted to WiMAX specifics and our channel model.

The available resource being limited, a mobile does not always achieve its GBR if C is too big. Thus, we derive $\bar{X}$, the instantaneous throughput obtained by a mobile:

$$\bar{X} = \sum_{c=1}^{C} \frac{\pi(c)}{1 - \pi(0)} \sum_{\substack{(c_0, ..., c_K) = (0, ..., 0) | \\ c_0 + ... + c_K = c \\ c_0 \neq c}}^{(c, ..., c)} \binom{c}{c_0, ..., c_K}$$

$$\cdot \left(\prod_{k=0}^{K} p_k^{c_k} \right) \frac{\bar{N}_S}{\max\left(\sum_{k=1}^{K} c_k g_k, \bar{N}_S \right)} GBR, \tag{4}$$

where $\left(\frac{\bar{N}_S}{\max\left(\sum_{k=1}^{K} c_k g_k, \bar{N}_S \right)} GBR \right)$ corresponds to the throughput achieved by a mobile when the c connections are distributed in $(c_0, ..., c_K)$. Finally, $\bar{U}$, the average utilization of the TDD frame by rtPS connections is expressed as:

$$\bar{U} = \sum_{c=1}^{C} \pi(c) \sum_{\substack{(c_0, ..., c_K) = (0, ..., 0) | \\ c_0 + ... + c_K = c \\ c_0 \neq c}}^{(c, ..., c)} \binom{c}{c_0, ..., c_K}$$

$$\cdot \left(\prod_{k=0}^{K} p_k^{c_k} \right) \frac{\sum_{k=1}^{K} c_k g_k}{\max\left(\sum_{k=1}^{K} c_k g_k, \bar{N}_S \right)}. \tag{5}$$

B. Multi-Traffic Extension

We now relax the assumption that all users have the same traffic profile. To do so, we distribute the mobiles among R traffic profiles defined by $(GBR_r, \bar{t}^r_{on}, \bar{t}^r_{off})$. Thus, the mobiles of a given profile r generate an infinite-length ON/OFF traffic, with a guaranteed bit rate of GBR_r bits per second, an average ON duration of $\bar{t}^r_{on}$ seconds and an average OFF duration of $\bar{t}^r_{off}$ seconds. We consider fixed numbers N_r of mobiles belonging to each profile in the cell.

Similarly to the mono-traffic model, we define DBR_r, the bit rate demanded by a profile-r mobile:

$$DBR_r = \frac{GBR_r}{1 - p_0}. \tag{6}$$

Note that if at a given frame, there is not enough slots to satisfy all active connections, they are then degraded proportionally to their respective DBR_r.

To model this system, we use the multi-class extension of the Engset model [11] where a profile-r mobile initiates and terminates a connection with respective rates $\lambda_r = \frac{1}{\bar{t}^r_{off}}$ and $\mu_r = \frac{1}{\bar{t}^r_{off}}$. Each states of the associated CTMC is characterized by a specific R-tuple $(c_1, ..., c_R)$ where c_r is the number of active connections of profile r. We assume that C_{max}, the limit on the maximum number of concurrent rtPS connections, is observed regardless of their profiles and we define $C_r = \min(N_r, C_{max})$, the maximum number of simultaneous profile-r connections. By defining the traffic intensity parameters $\rho_r = \frac{\lambda_r}{\mu_r}$, the steady state probabilities $\pi(c_1, ..., c_R)$ of having $(c_1, ..., c_R)$ concurrent rtPS connections can be expressed as:

$$\pi(c_1, ..., c_R) = \left(\prod_{r=1}^{R} \frac{\rho_r^{c_r}}{c_r!} \frac{N_r!}{(N_r - c_r)!} \right) \pi(0, ..., 0), \tag{7}$$

with $\pi(0, ..., 0)$ obtained by normalization. By first defining $\bar{g}_r(c_r)$, the number of slots needed by c_r profile-r connections to achieve their DBR_r:

$$\bar{g}_r(c_r) = c_r \sum_{k=1}^{K} p_k \frac{DBR_r \; T_F}{m_k}, \tag{8}$$

and $\bar{g}(v_1, ..., v_R)$, the mean number of slots needed by $(v_1, ..., v_R)$ connections:

$$\bar{g}(c_1, ..., c_R) = \sum_{r=1}^{R} \bar{g}_r(c_r), \tag{9}$$

we can then express $\bar{X}_r$, the instantaneous throughput achieved by profile-r mobiles as:

$$\bar{X}_r = \sum_{\substack{(c_1, ..., c_R) = (0, ..., 0) | \\ c_1 + ... + c_R \leq C_{max}}}^{(C_1, ..., C_R)} \frac{\pi(c_1, ..., c_R)}{1 - p_{c_r = 0}}$$

$$\cdot \frac{\bar{N}_S}{\max\left(\bar{g}(c_1, ..., c_R), \bar{N}_S \right)} GBR_r, \tag{10}$$

with $p_{c_r=0}$ the probability that no class-r connection is active:

$$p_{c_r=0} = \sum_{\substack{(c_1,...,c_R)=(0,...,0)| \\ c_1+...+c_R \leq C_{max} \\ c_r = 0}}^{(C^1,...,C^R)} \pi(c_1,...,c_R), \quad (11)$$

and $\bar{U}_r$, the average utilization of the TDD frame by profile-r connections as:

$$\bar{U}_r = \sum_{\substack{(c_1,...,c_R)=(0,...,0)| \\ c_1+...+c_R \leq C_{max} \\ c_r \neq 0}}^{(C_1,...,C_R)} \frac{\bar{g}_r(c_r)}{\max\left(\bar{g}(c_1,...,c_R), \bar{N}_S\right)} .\pi(c_1,...,c_R). \quad (12)$$

IV. Packet Level Modeling

In this section, we show how to integrate a packet level to our connection level models.

A. Mono-Traffic Model

We consider that the packets of a connection have a mean bit length $\bar{L}_p$ and arrive at the base station at a bit rate called Arrival Bit Rate (ABR). The packet level traffic profile of a connection is thus characterized by the pair $(ABR, \bar{L}_p)$. Here, we assume that the mobiles are all statistically identical.

The transmission time of a packet depends on the number of active connections sharing the resource. To model the packet level while avoiding the analysis of a multidimensional CTMC, we separately consider the packet level behavior of a single connection for each possible number c of concurrent connections. We associate with each state $c \neq 0$ of the connection level CTMC a packet level CTMC as shown in Fig. 1. Each state p of the packet level CTMC associated with a state c corresponds to a number p of packets belonging to the same connection and waiting at the base station when exactly c connections share the resource. This CTMC is thus made of $P_{max}+1$ states as a connection can only have at most P_{max} packets at the base station. Its transitions are as follows:

- A transition out of a generic state p to state $p+1$ occurs when a new packet is received by the base station. This "arrival" transition is performed with a rate:

$$\lambda_p = \frac{ABR}{\bar{L}_p}. \quad (13)$$

- On the opposite, a transition out of a generic state p to state $p-1$ occurs when a packet has been transmitted. This "departure" transition depends on c, the number of concurrent connections sharing the resource and is performed with a rate:

$$\mu_p(c) = \frac{GBR}{\bar{L}_p} \frac{\bar{N}_S}{c\sum_{k=1}^{K} p_k g_k}. \quad (14)$$

- Finally, the transitions out of any generic state p to state 0 enables to account for the dropping of the packets belonging to the connection when it is terminated. As such, they are performed with a rate $\mu = \frac{1}{\bar{t}_{on}}$.

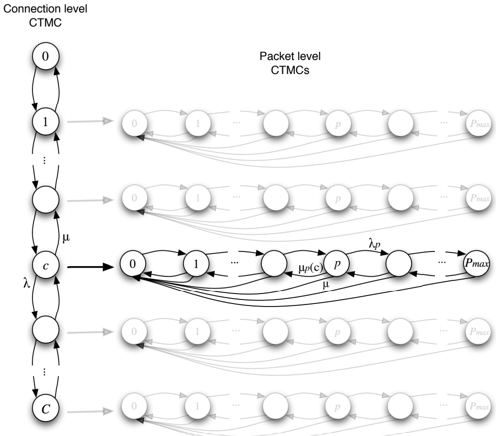


Fig. 1. Mono-traffic packet modeling.

From each of the C packet level CTMCs, we can derive the steady state probabilities $\pi_c(p)$ that an active connection has p packets waiting to be transmitted when there are c concurrent connections as:

$$\pi_c(p) = \left(\prod_{n=0}^{P_{max}-(p+1)} \frac{1}{u_n}\right) \pi_c(P_{max}), \quad (15)$$

where $(u_n)_{0 \leq n \leq P_{max}}$, is given by:

$$\begin{cases} u_0 = \dfrac{\lambda_p}{\mu_p(c)+\mu} \\ u_n = \lambda_p \left[\lambda_p + \mu + (1-u_{n-1})\mu_p(c)\right]^{-1} \end{cases} \quad (16)$$

and $\pi_c(P_{max})$ is obtained by normalization.

We deduce from these steady state probabilities the following packet level performance parameters. The average number of packets, $\bar{Q}_p$, present in the system is computed as:

$$\bar{Q}_p = \sum_{c=1}^{C} c\,\pi(c) \sum_{p=1}^{\bar{P}_{max}} p\,\pi_c(p), \quad (17)$$

and $\bar{D}_p$, the mean number of packet transmissions per unit of time is expressed as:

$$\bar{D}_p = \sum_{c=1}^{C} c\,\pi(c) \sum_{p=1}^{\bar{P}_{max}} \mu_p(c)\,\pi_c(p). \quad (18)$$

From Little's law, we can thus derive the average delay $\bar{R}_p$ between the reception of a packet by the base station and the end of its transfer to a mobile:

$$\bar{R}_p = \frac{\bar{Q}_p}{\bar{D}_p}. \quad (19)$$

Finally, we provide the reject probability $P_{rej\,p}$ of a packet arriving at the base station using the PASTA property:

$$P_{rej\,p} = \sum_{c=1}^{C} \pi(c)\pi_c(\bar{P}_{max}). \quad (20)$$

B. Multi-Traffic Extension

Now, we relax the assumption that all users have the same traffic profile. To this aim, we associate with each mobile one of the R traffic profiles, at connection level $(GBR_r, \bar{x}^r_{on}, \bar{t}^r_{off})$ but also at packet level $(ABR_r, \bar{L}^r_p)$. The packets intended for a mobile of a given profile r are thus received by the base station with an ABR_r bit rate and have an average bit length $\bar{L}^r_p$. Besides, we consider that a profile-r connection can have at most $\bar{P}^r_{max}$ packets waiting to be transmitted.

For the modeling of the packet level of this system, we follow the same approach as we did in mono-traffic. However, we now repeat this approach for each of the R traffic profiles. To model the packet level behavior of a profile-r connection, we thus associate a profile-r packet level CTMC with each state $(c_1, ..., c_R)$ of the connection level CTMC such that $c_r \neq 0$. These packet level CTMCs present similar transitions than in mono-traffic, yet their rates are different:

- A transition out of a generic state p to state $p+1$ happens with a rate:
$$\lambda^r_p = \frac{ABR_r}{\bar{L}^r_p}. \tag{21}$$
- On the opposite, a transition out of a generic state p to state $p-1$ occurs with a rate:
$$\mu^r_p(c_1, ..., c_R) = \frac{GBR_r}{\bar{L}^r_p} \frac{\bar{N}_S}{\bar{g}(c_1, ..., c_R)}, \tag{22}$$
which depends on the $(c_1, ..., c_R)$ connections sharing the resource. ($\bar{g}(c_1, ..., c_R)$ is given by relation 9).
- Finally, the transitions out of any generic state p to the state 0 are performed with a rate $\mu_r = \frac{1}{\bar{t}^r_{on}}$.

The steady state probabilities $\pi^r_{(c_1,...,c_R)}(p)$ that a profile-r connection has p packets waiting to be transmitted when there are $(c_1, ..., c_R)$ concurrent connections are derived from the packet level CTMCs as:

$$\pi^r_{(c_1,...,c_R)}(p) = \left(\prod_{n=0}^{P^r_{max}-(p+1)} \frac{1}{v_n} \right) \pi^r_{(c_1,...,c_R)}(P^r_{max}), \tag{23}$$

where $(v_n)_{0 \leq n \leq P^r_{max}}$, is defined as:

$$\begin{cases} v_0 = \dfrac{\lambda^r_p}{\mu^r_p(c_1, ..., c_R) + \mu_r} \\ v_n = \lambda^r_p \left[\lambda^r_p + \mu_r + (1 - v_{n-1}) \mu^r_p(c_1, ..., c_R) \right]^{-1} \end{cases} \tag{24}$$

and $\pi^r_{(c_1,...,c_R)}(P^r_{max})$ is obtained by normalization.

The mean number, $\bar{Q}^r_p$, of profile-r packets waiting to be fully transfered is given by:

$$\bar{Q}^r_p = \sum_{\substack{(c_1, ..., c_R) = (0, ..., 0) | \\ c_1 + ... + c_R \leq C_{max} \\ c_r \neq 0}}^{(C_1,...,C_R)} c_r \, \pi(c_1, ..., c_R) \cdot \sum_{p=1}^{\bar{P}^r_{max}} p \, \pi^r_{(c_1,...,c_R)}(p), \tag{25}$$

and the average number of transmission of profile-r packets per unit of time, $\bar{D}^{T\,r}_p$, is expressed as:

$$\bar{D}^r_p = \sum_{\substack{(c_1, ..., c_R) = (0, ..., 0) | \\ c_1 + ... + c_R \leq C_{max} \\ c_r \neq 0}}^{(C_1,...,C_R)} c_r \, \pi(c_1, ..., c_R) \cdot \sum_{p=1}^{\bar{P}^r_{max}} \mu^r_p(c_1, ..., c_R) \, \pi^r_{(c_1,...,c_R)}(p). \tag{26}$$

The average delay, $\bar{R}^r_p$, between the reception of a profile-r packet by the base station and the end of its transmission is obtained using Little's law:

$$\bar{R}^r_p = \frac{\bar{Q}^r_p}{\bar{D}^r_p}. \tag{27}$$

Lastly, we formulate the reject probability $P^r_{rej\,p}$ of an arriving packet with profile r as:

$$P^r_{rej\,p} = \sum_{\substack{(c_1, ..., c_R) = (0, ..., 0) | \\ c_1 + ... + c_R \leq C_{max} \\ c_r \neq 0}}^{(C_1,...,C_R)} \pi(c_1, ..., c_R) \pi^r_{(c_1,...,c_R)}(\bar{P}^r_{max}). \tag{28}$$

V. Validation

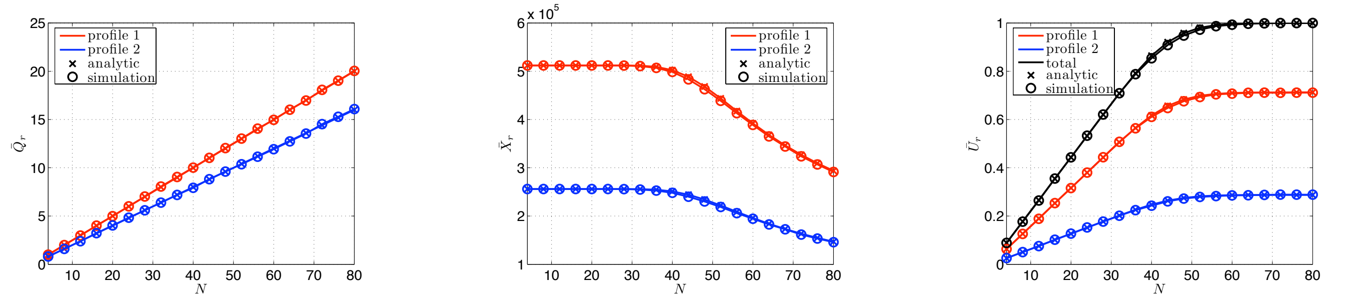
To validate our two-level modeling, we now compare the results of our models with those of simulations. To this aim, we developed a simulator that implements an ON/OFF traffic generator, a wireless channel for each user and a centralized scheduler allocating radio resources, i.e., slots, to active users at each frame. The ON and OFF durations of the mobiles are exponentially distributed and the packet arrivals follow a Poisson process. Contrary to the models which only consider averages, in simulations, the slots are allocated to each mobiles on a frame by frame basis.

We repeatedly compared results from our analytical models and simulations while considering all sorts of scenarios. We also compared our model with fixed packet length simulations. Here, we present the results of a representative scenario.

TABLE I
CHANNEL PARAMETERS

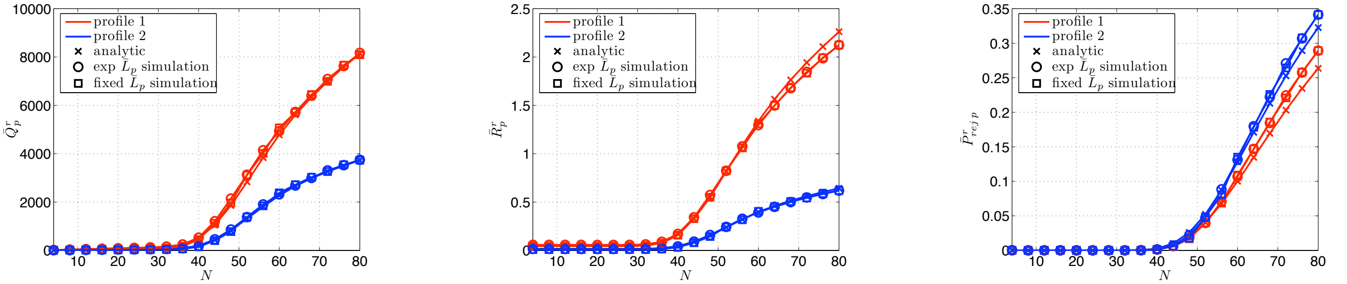
MCS	bits per slot	probability
Outage	$m_0 = 0$	$p_0 = 0.02$
QPSK-1/2	$m_1 = 48$	$p_1 = 0.12$
QPSK-3/4	$m_2 = 72$	$p_2 = 0.31$
16QAM-1/2	$m_3 = 96$	$p_3 = 0.08$
16QAM-3/4	$m_4 = 144$	$p_4 = 0.47$

We consider $\bar{N}_S = 450$ available slots per frame. This value corresponds to a system bandwidth of 10 MHz, a downlink/uplink ratio of 2/3, a PUSC subcarrier permutation and an average protocol overhead length of 2 symbols. We



(a) Mean number, $\bar{Q}_r$, of active connections. (b) Instantaneous throughput, $\bar{X}_r$, of a mobile. (c) Mean utilization, $\bar{U}_r$, of the TDD frame.

Fig. 2. Customary connection level performance parameters when the number N of mobiles in the cell increases.



(a) Mean number, $\bar{Q}_p^r$, of packets. (b) Mean latency, $\bar{R}_p^r$, of packets. (c) Reject probability $\bar{P}_{rej\,p}^r$ of packets.

Fig. 3. Customary packet level performance parameters when the number N of mobiles in the cell increases.

assume a number, N, of mobiles present in the cell ranging from 2 to 80. The mobiles are equally distributed among the two considered traffic profiles. We voluntarily accept up to $C_{max} = 50$ concurrent active connections to see the effects of congestion on the performance. The channel parameters are summarized in Table I while the traffic parameters are detailed in Table II. The results are presented in Fig. 2 and 3.

TABLE II
TRAFFIC PARAMETERS

Traffic profile, *r*	**1**	**2**
Proportion of mobiles	50%	50%
Guaranteed bit rate, GBR_r	512 Kbps	256 Kbps
Mean ON duration, $\bar{t}_{on}^r$	60 s	60 s
Mean OFF duration, $\bar{t}_{off}^r$	60 s	90 s
Arrival bit rate, ABR_r	95% of GBR_1	90% of GBR_2
Mean packet length, $\bar{L}_p^r$	1600 bits	400 bits
Limit on the packets, $\bar{P}_{max}^r$	500	250

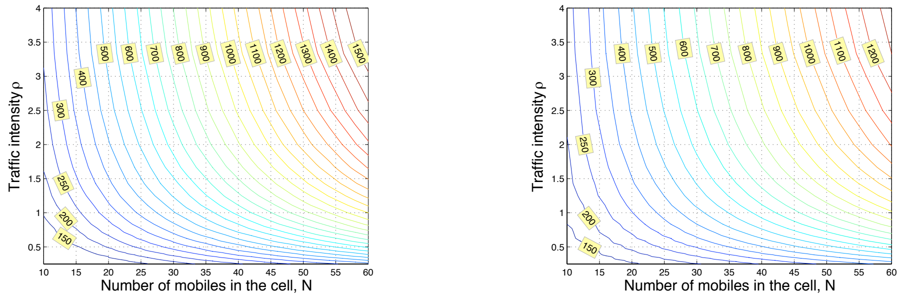
The average numbers, $\bar{Q}_r$, of active connections of both profiles obviously increase with the number of mobiles present in the cell. At first ($N < 40$), there is always enough resource to satisfy all demands: all connections get their desired throughputs ($\bar{X}_r = GBR_r$) and the average frame utilization, $\bar{U}$, linearly increases. The mean numbers, $\bar{Q}_p^r$, of packets waiting to be transmitted only increase very slowly and the mean packet latencies, $\bar{R}_p^r$ packets stay constant while reject probabilities of arriving packets, $\bar{P}_{rej\,p}^r$, are almost null. However, when there are more mobiles in the cell ($N > 40$), they begin to suffer the lack of available resource. Thus, the frames become fully occupied and the throughputs dive since more and more connections share the limited amount of resource. The congestion also has an impact on the packet level: the three parameters $\bar{Q}_p^r$, $\bar{R}_p^r$ and $\bar{P}_{rej\,p}^r$ rapidly increase.

The curves show that the results given by our analytical model match those of simulations. Indeed, the difference between them is less than 4% in most cases and less than 9% in the worst case. The only significant observed divergences appears when considering excessively overloaded scenarios.

In addition, the results of our packet level model are compared in Fig. 3 with fixed packet length simulations. The results stay very close whether we consider exponentially distributed or fixed packet length. We only observed small differences in cases of low traffic load. Finally, note that the results presented in this section are representative of the results obtained for the numerous considered scenarios. Indeed, each time, the simulations results validated our analytical models with similar accuracy.

VI. DIMENSIONING

Here, we provide a dimensioning example to demonstrate possible applications of our models. The channel parameters are provided in Table I and the traffic parameters in Table III. For the sake of simplicity, we only consider one traffic profile. Yet, results can be obtained for any other possible configuration (i.e., any mono or multi-profile traffic scenarios).



(a) Maximum number of slots, N_S^{max} to guarantee $\bar{U} \geq 0.85$.

(b) Minimum number of slots, N_S^{min} to guarantee $\bar{R}_p \leq 0.35$ s.

Fig. 4. Dimensioning of the number N_S of available downlink slots per frame.

TABLE III
TRAFFIC PARAMETERS

Parameter	Value
Limit on rtPS connections, C_{max}	50
Guaranteed bit rate, GBR	512 Kbps
Traffic intensity, ρ	0.25 to 4
Arrival bit rate, ABR	95% of GBR
Mean packet length, $\bar{L}_p$	1600 bits
Limit on the packets, $\bar{P}_{max}$	250

Our models can be advantageously used to answer dimensioning issues by drawing contour graphs. Two dimensioning criteria are proposed as examples and results are presented in Fig. 4(a) and 4(b). To draw these two graphs, we computed the considered performance parameter ($\bar{U}$ or $\bar{R}_p$) for each possible (N, ρ) pair while increasing the number $\bar{N}_S$ of available slots until the chosen criterion could not be guaranteed anymore. This straightforward method is only possible due to the instantaneous resolution of our model.

In Fig. 4(a), we find the maximum number $\bar{N}_S^{max}$ of available slots guaranteeing an average radio utilization over 85%. This kind of criterion allows operators to avoid uneconomical over-provisioning of the network resources in regard to the traffic load of their customers. To obtain the optimal value of $\bar{N}_S^{max}$ associated with a given (N, ρ) pair, we look for the point at the corresponding coordinates in the graph. This point is located between two contour lines, and the one with the lower value gives the value of $\bar{N}_S^{max}$.

The second criterion concerns the packet delay. We decided on 0.35 s as the maximum acceptable value of the mean packet delay. Now, we want to find the minimum number $\bar{N}_S^{min}$ of available slots guaranteeing this latency threshold. In Fig. 4(b), a given point is located between two contour lines. The line with the higher value gives $\bar{N}_S^{min}$.

The graphs of Fig. 4(a) and 4(b) can be jointly used to satisfy both criteria. For example, if we consider a cell with 35 mobiles and a traffic intensity $\rho = 1.5$, Fig. 4(a) gives $\bar{N}_S^{max} = 700$ slots, and Fig. 4(b) gives $\bar{N}_S^{min} = 600$ slots. So, we must consider a number of slots $\bar{N}_S \in [600; 700]$ to guarantee both a reasonable resource utilization and a satisfactory latency.

VII. CONCLUSION

In this paper, we have presented novel analytical models dedicated to rtPS, the WiMAX service class specifically designed for streaming traffic. First, we focused on the connection level only considering the bit rate constraint. Then, we extended our models to the packet level to formulate the mean packet latency. The resulting two-levels rtPS models are able to instantaneously provide closed-form expressions for all the required performance parameters even with multiple traffic profiles. Extensive simulations have validated the models' assumptions and shown their accuracy (maximum relative errors never exceeded 9%). Lastly, we have shown with an example how to use our models to perform advanced dimensioning involving crucial economical and QoS issues.

REFERENCES

[1] IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

[2] Draft IEEE std 802.16e/D9. IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems., 2005.

[3] A. Berger and Y. Kogan. Dimensioning bandwidth for elastic traffic in high-speed data networks. *IEEE Transactions on Networking*, 2000.

[4] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *IEEE Infocom*, 2003.

[5] P. Chowdhury, I. S. Misra, and S. K. Sanyal. An Integrated Call Admission Control and Uplink Packet Scheduling Mechanism for QoS Evaluation of IEEE 802.16 BWA Networks. *Canadian Journal on Multimedia and Wireless Networks*, 1(3), April 2010.

[6] C.-C. Chuanga and S.-J. Kao. Discrete-time modeling for performance analysis of real-time services in IEEE 802.16 networks. *Computer Communications Journal*, 2010.

[7] D.-J. Deng, L.-W. Chang, C.-H. Ke, Y.-M. Huang, and J. M. Chang. Delay constrained uplink scheduling policy for rtPS/ertPS service in IEEE .802.16e BWA systems. *Communication Systems*, 2008.

[8] S. Doirieux, B. Baynat, M. Maqbool, and M. Coupechoux. An Efficient Analytical Model for the Dimensioning of WiMAX Networks Supporting Multi-profile Best Effort Traffic. *Computer Communications Journal*, 2010.

[9] T. O. Engset. On the calculation of switches in an automatic telephone system. In *Tore Olaus Engset: The man behind the formula*, 1998.

[10] S. Ghazal, L. Mokdad, and J. Ben-Othman. Performance Analysis of UGS, rtPS, nrtPS Admission Control in WiMAX Networks. In *Proc of IEEE International Conference on Communications*, May 2008.

[11] A. Jensen. Truncated multidimensional distributions. *The Life and Works of A.K. Erlang*, pages 58–70, 1948.

[12] D. Niyato and E. Hossain. A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband networks. *IEEE ToC (vol. 55)*, 2006.

INTEGRATED CONTROL CHANNEL MECHANISM FOR EFFICIENT CHANNEL UTILIZATION IN TDMA FOR WLAN NETWORKS

Mohammed Amer Arafah, Nasir Hussain, Mohammed Zuheir Hourani
Department of Computer Engineering, College of Computer and Information Sciences
P.O.Box 51178, King Saud University, Riyadh 11543, KSA

ABSTRACT

In this paper we introduced a custom TDMA MAC Layer Protocol. The main idea which actuated the research was to simulate a complex MAC Protocol in a wireless environment and espy its behavior under different physical, traffic, and network configurations. The TDMA protocol itself is completely custom and not based on any one particular industry specification. Our approach is ideal for exchanging data and control information in Mobile Ad hoc Networks (MANET) and wireless mesh networks so that resource needs (in terms of slot availability) of participating nodes can be rapidly accommodated for Quality of Service (QoS) guarantees. The approach described in this paper allow an efficient resource allocation strategy such that each participating node can converge to a consistent TDMA control slot usage that is optimum for the topology and resource demand of a dynamic wireless network environment and then we further proceeded with capacity planning for such networks based on the optimum control slot usage .

Keywords: TDMA, Control Slot, WLAN, MAC Layer, OPNET Modeler.

I. INTRODUCTION

Time division multiple access (TDMA) is a channel access technique extensively adopted in shared medium networks. The users transmit in quick continuation using its own slice which is achieved by allowing several users to share the same frequency channel by dividing the signal into different time slots. Two algorithms are proposed in [1] for dynamic control slot scheduling that provide exchange of control information in an efficient and control free manner but at the cost of certain topology information provided in advance. Although its universal consent as a de-facto standard in today's market, IEEE 802.11 [2] based protocols are handicapped from bandwidth inefficiency under high contention environments [3]. Despite the new IEEE 802.11e standard experiments to take in due consideration the lack of Quality of Service (QoS) support by providing the Enhanced Distributed Channel Access (EDCA) protocol, but at times EDCA also vacillates when it comes to providing Differentiated Service support in low priority traffic under a heterogeneous environment [4]. TDMA based MAC protocols can be favorably deployed in strategy based networks where QoS guarantees and resource reservation are captious for the network services required by the user. Nonetheless, TDMA based customizations practice some weaknesses due to their demand for tight clock synchronization and consistent correspondence of control information toward achieving many of their acceded potential advantages. The problem of administering tight clock synchronization can be accommodated by distributed time synchronization techniques [5]. On the other hand, the problem of providing a reliable mechanism for exchanging control information in TDMA has not been looked at closely and is the focus of this paper. The paper is organized as follows, first we introduce the simulation environment and the network parameters then we further investigate the effect of control slot integration in different scenarios and finally we conclude.

II. TDMA PARAMETRIC STUDY

The main focus of this paper is to look at the problem of providing reliable control slot schedules for a TDMA based MAC protocol and introduce an integrated mechanism that achieve highly efficient channel utilization with low latency in the face of constant topology changes and link quality fluctuation in a MANET environment. The shaded slot in the figure below is pre-allocated for control slot scheduling. Using our approach, the pre-allocated control slot will be fairly shared among participating nodes to exchange control information so that the data portion of the frame (clear slots) can be optimally allocated based on the bandwidth need of each node.

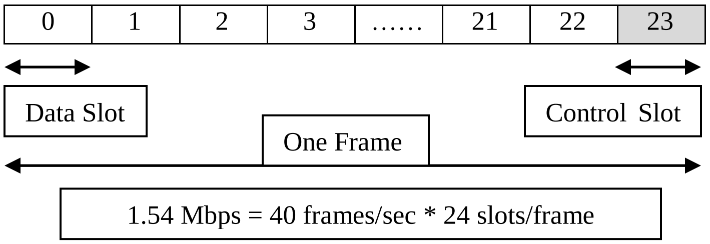


Figure 1 TDMA Channel

The simulation for this particular research was carried out in OPNET simulator having nodes in the network represented by a mobile laptop, a handheld data unit, and a VoIP cell phone. Each device has data access through the local cell tower. These nodes use the standard protocol stack (Application/TCP/IP) of a workstation in the OPNET Modeler standard model library. These client nodes will access the tower that is in the same subnet via a TDMA radio interface. There are some TDMA MAC layer attributes out of which only three are considered and are given below,

- *TDMA Channel Hold Time*

This attribute will control the length of time in seconds that a mobile unit will keep control of the timeslot after the initial packet transfer is complete. Some Applications (voice & video) have very regular packet rates and are sensitive to the TDMA system overhead of requesting a slot from the tower. For these applications a hold time of slightly longer that the period between packets will ensure a low latency data link. Bursty traffic might be able to use a zero value hold time. Ultimately, there is a trade-off between cell capacity (number of nodes that can be supported) and system access delay which will affect application response time.

- *TDMA Number of Slots to Use*

This attribute controls the number of slots that the node will request. One slot provides a data rate of 64Kbps; higher integer values will provide 128Kbps, 192Kbps, and 256Kbps.

- *TDMA Request Timeout*

This value is the length of time in seconds that a node will wait before retransmitting the slot request message. A typical retransmission will occur because there was a collision on the ALOHA [6] control slot of the original slot request.

Using two radio frequencies (One for Transmitting and the other to Receive), the cell tower provides two way IP data communication from the land based fixed network to the nodes (laptop, handheld, and cell phone). The MAC Layer interface between the nodes and the cell tower is a TDMA protocol which can support a large number of nodes with data rates from 64 Kbps to 1.54 Mbps. The landline interface to the tower is 100BaseT Ethernet.

Scenario I - Large Network Protocol with Common Control Channel

This scenario studies the network performance of a cell supporting 49 mobile clients (33 Laptops and 16 Handheld PDAs). Each type of node has a different mix of network applications including FTP, Web Browsing, Email, and Database sessions. By looking at the protocol statistics, over-utilization is not an issue. The results show that the average number of slots in use in the Tx direction is 20 out of 23. The average number of slots in use for the Rx direction is 10 out of 23.

Further analysis shows that there is a significant queuing delay for the mobiles to access the radio network. Some mobiles have a 2 second delay for each packet before it is sent to the Base station. Figure 2 shows that a significant number of client slot requests timeout. On average, 63% of all requests are being lost and causing timeouts. The loss of the control packets is being caused by radio collisions of control signals from multiple mobile units. Since the control channel is common resource to all mobiles in a cell, any mobile can access this channel at any time. This scheme is insufficient to support this number of mobile nodes with this traffic pattern.

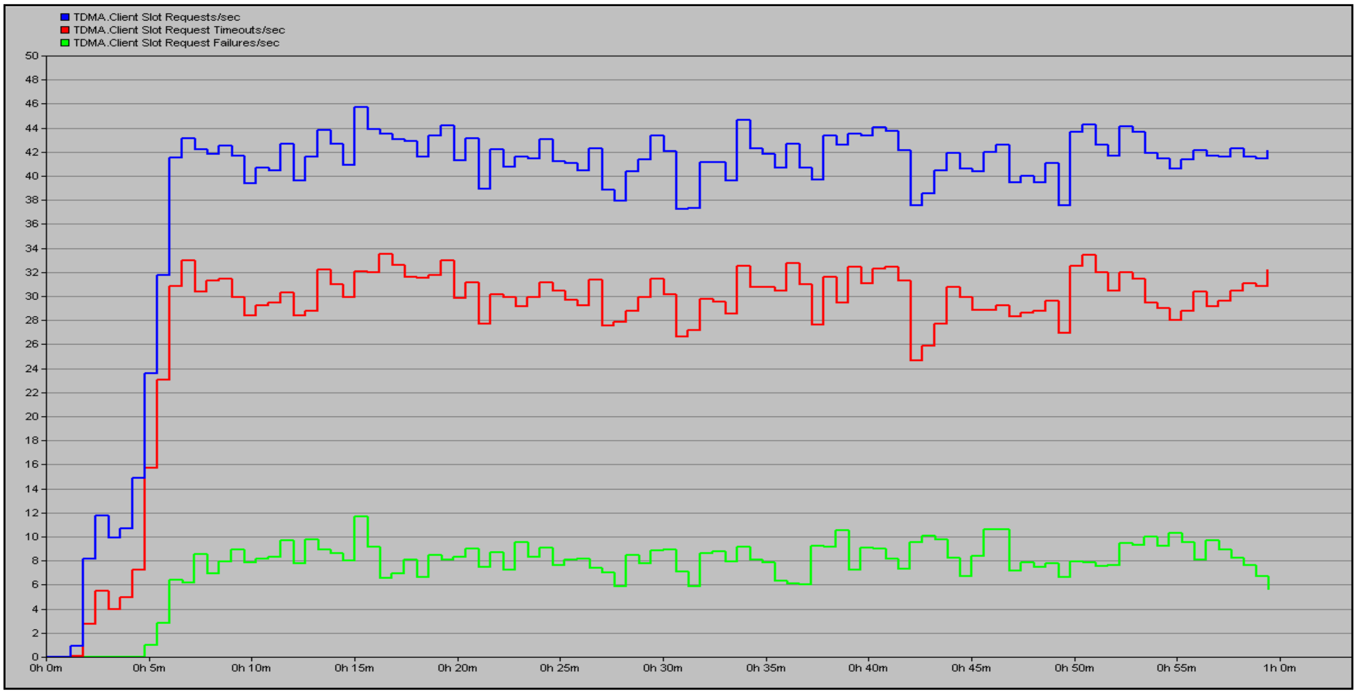


Figure 2 TDMA slot request, slot request timeout and slot request failure

Scenario II - Large Network Protocol with fixed Control Channel

This scenario attempts to fix the control channel access problem discovered in the previous scenario. Each control slot is 1600 bits long and the control messages are 82 bits long. In the previous scenario, any control message was sent at the beginning of the control slot. This method wasted a large percentage of available control channel bandwidth and caused the high number of collisions that were detected. By segmenting each control slot into sub-slots of > 82 bits, the protocol can make better use of the available bandwidth. By randomly picking a sub-slot to send the control message, and consequently reducing the number of collisions. In this scenario, the control channel is segmented into 10 sub-slots.

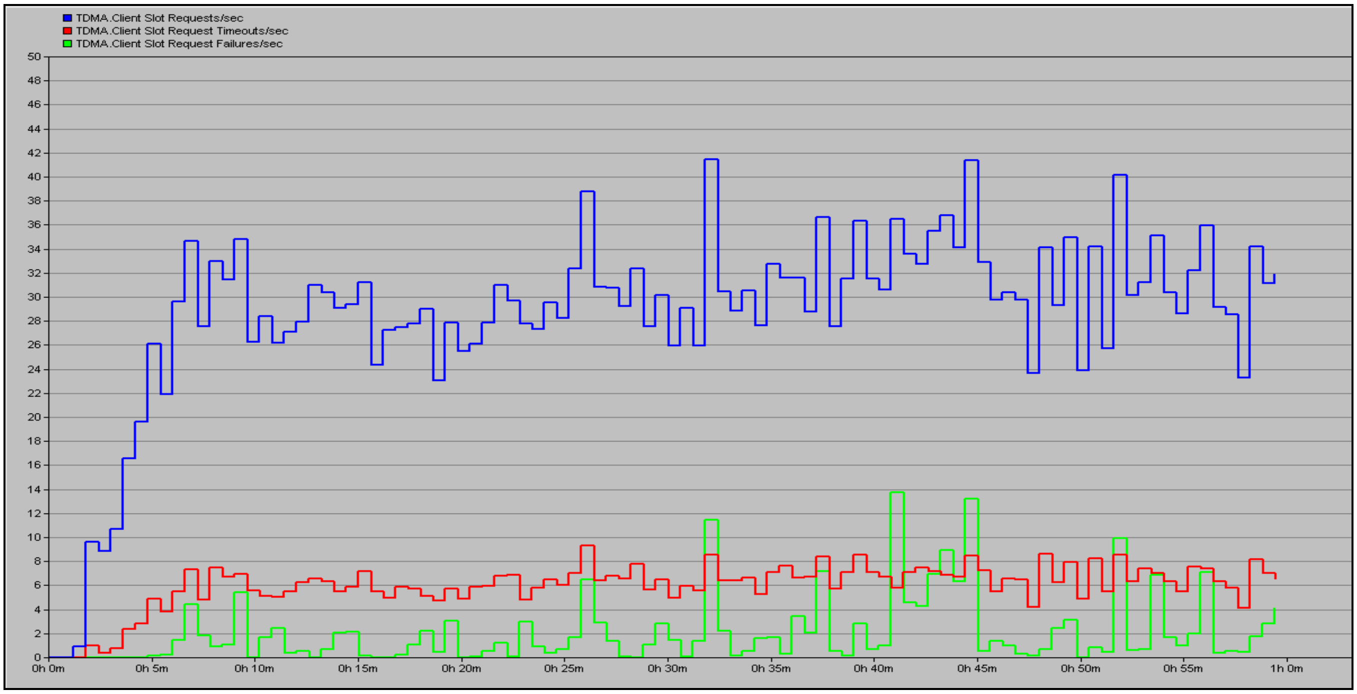


Figure 3 TDMA slot request, slot request timeout and slot request failure with fixed control channel

The results show a dramatic improvement in network performance with the new control slot configuration. Application performance has improved. The queuing delay for mobile access had been significantly reduced. The average number of slots in use for the Rx direction is 14 out of 23 as compared to 10 for the previous scenario thus capable of receiving more data. Figure 3 shows that the collisions of control messages are dropped from 63% to 18%.

$$Slot\ Capacity = (Channel\ Bandwidth\ (R))/((Slots\ (\mu))\ /Second) = 1604\ bits\ approx.$$

$$Integrated\ Control\ Slot = (Slot\ Capacity)/(Threshold\ Divisor) \cong 160\ bits/subslot$$

Comparison Based on Application Traffic

In regards to the above mentioned scenarios based on control channel integration the comparison for different traffic profiles are used to analyze the effect of control slot subdivision. By selecting traffic profiles such as the *laptop* nodes are using four types of application traffic,

- *Email (Heavy load):* Send and receive inter-arrival times are exponential and constant size of email with mean outcome 2000 and type of service is best effort.
- *File Transfer Protocol (Light load):* Inter-request time is exponential and file size is constant with mean outcome 50,000 and type of service is best effort.
- *Web browsing (Heavy load HTTP):* Page inter-arrival time is exponential with best effort as the type of service.

Now, for the *handheld* devices the traffic profile is defined as under,

- *Email (Light load):* Send and receive inter-arrival times are exponential and constant size of email with mean outcome 500 and service type is best effort.
- *Web browsing (Heavy load HTTP):* Page inter-arrival time is exponential with best effort as the type of service.

Finally, for the *cellular* devices which only consists of voice application defined as consisting a G.723.1 5.3K voice encoder and 0.02 seconds of compression delay and best effort as the type of service. Figure 4 includes the response time for different traffic profiles used during the course of simulation. Web browsing for the scenario in which there is a common control slot gives an average response time of 18 seconds which is undesirable by all means but when the control slot is subdivided into fixed sub-slots the response time goes down to 1.5 seconds which is acceptable. Similarly, for FTP traffic the average response time is reduced to 11 seconds and the response time for downloading an email on the average is reduced from 25 seconds to 3 seconds. Analyzing these results will lead to the conclusion that the response time has decreased for all kinds of traffic providing efficient channel utilization in terms of slot usage.

Also, the results show that the QoS mechanism built into the protocol (hold time) is able to provide high quality circuit behavior in a packet oriented protocol. Once a voice call is established, the unit reserves the slot for the duration of that call. This allows voice to maintain high quality even during times of congestion. Average response time for the data traffic in web browsing is 0.6 sec in the case of 90 nodes which is the optimum value and similarly, for the voice traffic the average packet ETE-Delay is 0.15 sec for the 90 nodes case. Based on these observations we can conclude that response time for voice and data traffic is best suited when we have 90 (±5) nodes. Average response times for the data (HTTP) traffic for the three cases is illustrated in figure 5 and average packet ETE-Delay for the three cases is illustrated in figure 6.

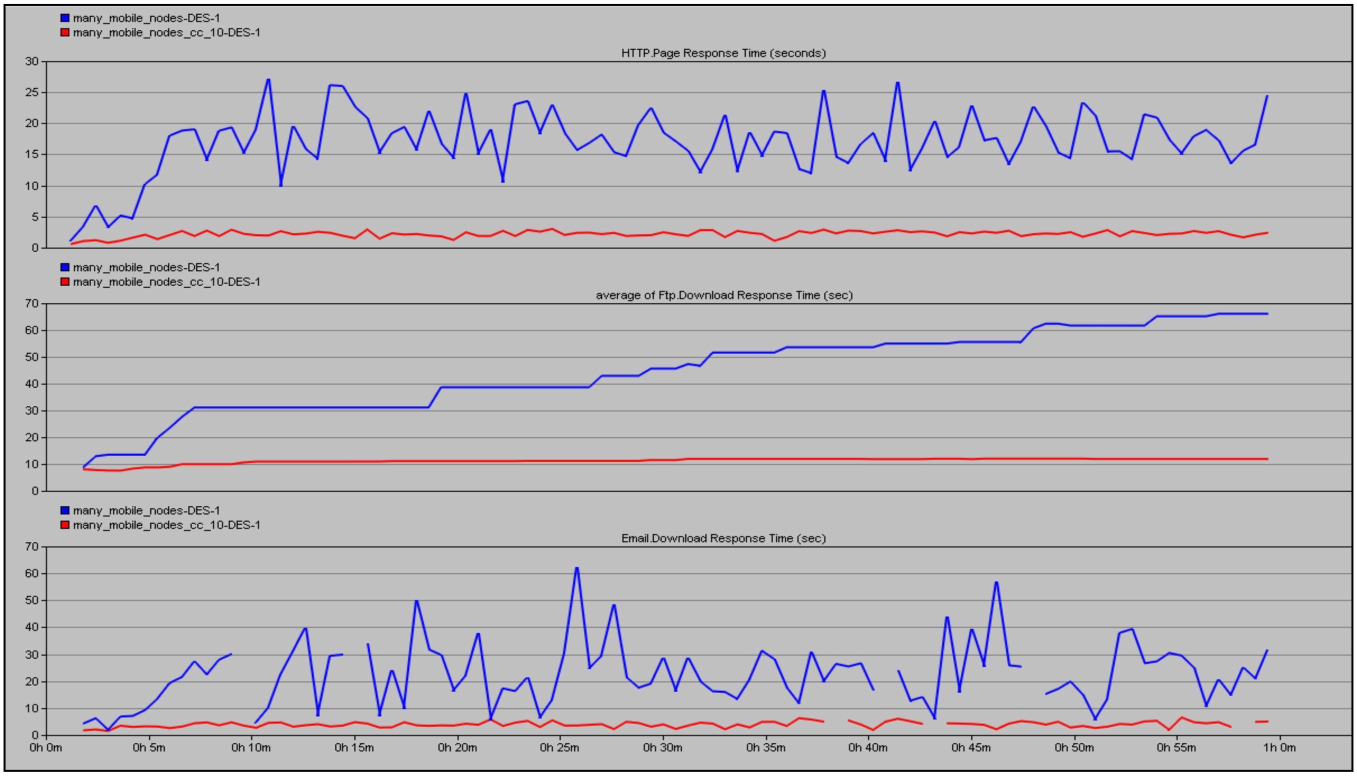


Figure 4 Response Time analysis for the traffic trunk profile

Scenario III - Network Scalability and Capacity Planning

This scenario attempts to find the maximum number of mobile nodes that a single frequency pair can support. A mix of data and voice nodes have been arranged within a cell with the control slot sub-division set at 10. The network is composed of 75, 90, and 105 users. Simulations are run to find application and network performance for each. Results show that radio utilization is highest when the number of nodes is between 90 and 105. But, when the number of nodes increases to 105, radio frequency utilization actually decreases.

III. Conclusion

Based on the above illustrated behavior of the TDMA for WLAN networks we can conclude that the channel needs to be properly administered in order to provide QoS guarantees to every node trying to connect and gain access to the channel.

Secondly, the protocol which serves the channel allocation has to be properly investigated in order to plan the capacity needs of the network and to provide efficient channel utilization and then suggest threshold value for the number of mobile nodes that can be supported with efficient bandwidth utilization for a given traffic demand.

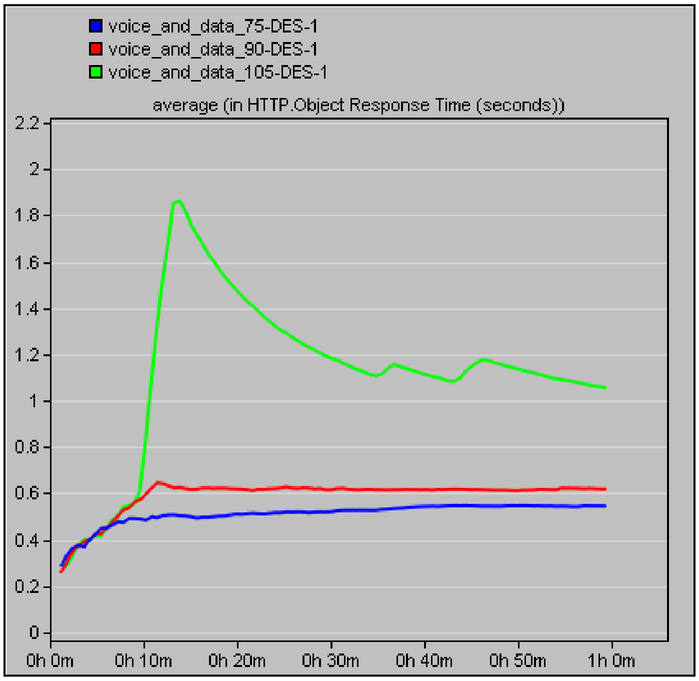


Figure 5 Comparison of Web Browsing response time

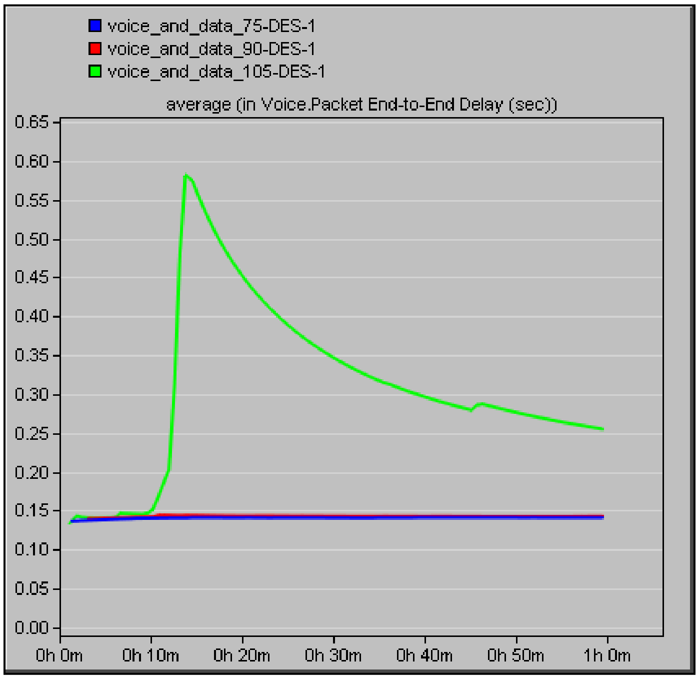


Figure 6 Comparison of Voice Traffic ETE-delay

ACKNOWLEDGMENT

Special thanks to Dr. Ridha Ouni and Dr. Hamid Abachi for their constant feedback and support. This research is supported by the Research Centre and Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh.

REFERENCES

[1]. Sung Park, Denh Sy, "*Dynamic control slot scheduling algorithms for TDMA based mobile ad hoc networks*", MILCOM 2008 IEEE. pp 1-7.

[2]. IEEE LAN MAN Standards Committee, *Wireless LAN medium access control (MAC) and physical layer (PHY) specifications*. IEEE Standard 802.11-2007.

[3]. R. Rozovsky and P. R. Kumar. SEEDEX: "*A MAC protocol for ad hoc networks*". Proceedings of the 2nd Symposium on Mobile Ad hoc Networking and Computing 2001, pp 67-75.

[4]. M. I. Abu-Tair and G. Min, "*Performance evaluation of an enhanced distributed channel access protocol under heterogeneous traffic*", Proceedings of 20th International Parallel and Distributed Processing Symposium IPDPS 2006.

[5]. J. Elson, L. Girod, and D. Estrin. "*Fine-grained Network Time Synchronization Using Reference Broadcasts*". In Proceedings of the Fifth Symposium on Operating Systems Design and Implementation (OSID 2002).

[6]. S. Keshav, *An Engineering Approach to Computer Networking*, Addison Wesley, Reading, MA, 1997, pp 150- 153.

MANET Security Schemes

Asma Ahmed[1], **A. Hanan**[2], **Shukor A. R.**[2], **Izzeldin M.** [3]
[1]Faculty of Computer Science and Information System, Universiti Technologi Malaysia, Johor, Malaysia
[2]Department of Computer Science Universiti Technologi Malaysia, Johor, Malaysia
[3]Faculty of Computer Science, Sudan University Science and Technology, Khartoum, Sudan

Abstract - *There have been various security measures proposed for protect Mobile Ad Hoc Networks (MANET). These can be categorized in two main of security measures. That is Prevention and Detection/Reaction mechanisms. Prevention mechanisms are considered as the premier defense line against attackers. On the other hand Intrusion Detection Systems (IDS) is second layer of security to defense the attacks that happen in depth. However, Clearly the problem is so broad that there is no way to devise a general solution. It is also clear that different applications will have different security requirements. This paper present prevention mechanism which are considered as the premier defense line against attackers. In Prevention mechanisms there is require for encryption techniques to provide authentication, confidentiality, integrity and non-repudiation of routing information. Various secure routing protocols proposed for MANET are investigated as well as the kinds of attacks that they can be protect.*

Keywords*:* MANET Security, prevention mechanism Secure Routing protocol,,

1 Introduction

Mobile ad-hoc networks (MANET) is a group of wireless mobile nodes, in which nodes cooperate by forwarding packets for each other to allow them to communicate beyond direct wireless transmission range. Nodes are free to move anywhere and anytime. No central administrator is required to organize the connections between nodes. Each node takes the responsibility to manage itself. MANET is being used in very sensitive applications and can be quickly and inexpensively set up as needed. An example of these applicatons are military exercises and disaster relief. However, secure and reliable communication is a necessary prerequisite for such applications. The absence of any fixed infrastructure and mobility features in MANET makes it difficult to utilize the existing techniques for network services, and poses number of various challenges to keep such a network secure. In order to overcome the vulnerabilities and achieve security goals, there is a need to find some security measures to protect the network. Clearly the problem is so broad that there is no way to devise a general solution. It is also clear that different applications will have different security requirements. The complexity and diversity of the field has led to a multitude of proposals, which focus on different parts of the problem domain. There is two main categorized of security measures proposed for MANET that is Prevention and Detection/Reaction mechanisms. Prevention mechanisms are considered as the premier defense line against attackers. Prevention is used for secure network operation from external attacks. Prevention mechanisms require encryption techniques to provide authentication, confidentiality, integrity and non-repudiation of routing information,. These can be achieved by authenticating users and nodes [1][2], and by securing routing protocols used to create routes between nodes[3] . When attacks can penetrate this line of defense, prevention mechanisms become not enough for securing the network. Intrusion Detection Systems (IDS) is second layer of security to defense the attacks that happen in depth. IDS should be able to detect the malicious activities of attackers who successfully penetrated the prevention mechanisms. Detection and response mechanisms are used to secure network against internal attacks. This can be achieved using intrusion detection systems[4][5].

The aim of this paper is to investigate the secure routing protocol that can provide security and data privacy against the external attacks.

The paper is organized as follows Section 2 provides an overview of routing protocols in MANET. Section 3 discuss different secure proactive routing protocols methods. Section 4 discuss different secure reactive routing protocols methods as well as result and analysis of the reviewed secure routing protocols methods. Section 5concludes the paper.

2 Routing protocol MANET

There are two types of routing protocols: proactive and reactive protocols. In proactive routing protocols; routing tables are created before nodes ask for the routes. Each node has one or more table containing up-to-date routing information from each node to every other node in the network. An examples of such protocols are Optimized Link-State routing protocol (OLSR)[6] and Destination Sequence Distance Vector protocol (DSDV)[8]. On the other hand in reactive routing protocols; routes are created just when nodes ask for routes. In a reactive routing protocol, control packets, namely Route Request messages(RREQ), are broadcast by the source node in order to find the optimal route to the destination node. An examples of reactive routing protocols are Ad hoc On demand Distance Vector (AODV)[7] and Dynamic Source Routing Protocol (DSR)[13].

3. Secure Proactive Routing protocols

3.1 Secure Efficient Adhoc Distance Vector (SEAD)

In [9] the authers suggested SEAD (Secure Efficient Ad hoc Distance Vector) to secure DSDV routing protocol. SEAD uses one way hash chain. It defends against modifying the sequence number or the metric value in the route updates by malicious nodes. SEAD also uses this hash function chain to authenticate metric and sequence number in the routing update. Each node selects a random seed and applies a hash function many times on this seed to generate the hash chain elements. One of these elements (authentic elements) is used for the authentication process. Many ideas were suggested for distributing the authentic element. They also suggested using asymmetric cryptography system. A trusted third party (CA) is used to sign public key for each node. Each node distributes its public key and public key credentials. This public key is used then to sign the authentic element. In[14] authers suggested using symmetric-key cryptography mechanism to secure authentic element. SEAD uses also a shared secret key between each two nodes and a Message Authentication Code (MAC) to be sure that routing updates come from authenticated neighbors. Since SEAD is robust against modifying sequence number and metric attacks, it cannot defend against tunneling and vertex cut attacks.

3.2 The OLSR Security Extension

In [10][11], schemes have been proposed for extending OLSR to make it secure against attacks. The main idea they propose is to use digital signatures for authenticating the OLSR routing messages. Such authentication may be done on a hop-by-hop basis or on an end-to-end basis. Scheme in [0] focus on the hop-by-hop approach, in which each node signs OLSR packets as they are transmitted (such packets may contain multiple OLSR messages originated by a variety of nodes). The authers in [0] and [11] discuss schemes for authenticating OLSR messages on end-to-end basis so that nodes receiving OLSR message can authenticate the node that generated the original message rather than just the node forwarding the message.

4. Secure Reactive Routing protocols

4.1 Securing AODV routing protocol

Secure Ad hoc On-Demand Distance Vector Routing Protocol (SAODV) [15][16] is an extension of the AODV routing protocol that can be used to protect the route discovery mechanism of the original AODV providing security features like integrity, authentication and non-repudiation. This extension has the format shown in figure 1.

<table>
<tr><td>Type</td><td>Length</td><td>Hash function</td><td>Max hop count</td></tr>
<tr><td colspan="4">Top hash</td></tr>
<tr><td colspan="4">Signature</td></tr>
<tr><td colspan="4">Hash</td></tr>
</table>

Figure 1: SAODV message extension

SAODV incorporates two schemes for securing AODV. The first scheme involves nodes signing the messages that they create (e.g. RREQ, RREP). This allows other nodes to verify the originator of the message. This scheme can be used for protecting the portion of the information in the RREQ and RREP messages that does not change once these messages are created. However, RREP and RREQ messages also contain a field (namely the hop count) that needs to be changed by every node. Such mutable information is ignored by the creator of the message when signing the message. The second scheme of SAODV is used for protecting such mutable information. This scheme leverages the idea of hash chains. The signing routing messages imply the various nodes need to possess a key pair that makes use of an asymmetric cipher. Therefore, a key management scheme is required and can be used for this purpose.

4.1.1 Digital signatures

Digital signatures are used to protect the integrity of the non-mutable fields in RREQ and RREP messages. The signing process is accomplished by using asymmetric cryptography. SAODV defines three types of message extensions; the first extension is called "SAODV Message Extension", it is used by other nodes to verify the authenticity of the originator node. The second extension is called "RREQ double signature extension". This extension is used to protect the non mutable fields in RREQ and RREP message. The last extension is called "RREP double signature extension". This extension is used to allow the intermediate nodes to generate RREP messages signed by the destination nodes.

4.1.2 Hash Chain

These chains are used in SAODV to authenticate mutable information such as the hop count field in routing messages RREQ and RREP. Hash chains are created by applying a hash function repeatedly to a seed number. This scheme provides protection against nodes manipulating (more precisely decreasing) the hop count when forwarding AODV routing messages assuming a strong hash function.

4.2 ARAN Protocol

ARAN [17] stand of Authenticated Routing for Ad Hoc Networks. ARAN is a security scheme, which can apply to any on-demand routing protocol. ARAN is similar to SAODV in many points; both of them are based on digital signature and also both of them uses control messages. Routing operations of ARAN are performed using three data structures: Route Discovery Packet (RDP), Reply message (REP) and error message (ERR). These messages have the same functionality of RREQ, RREP and RERR messages in SAODV. Each of these messages is secured by digital signatures. These messages use the forward path and the reverse path during the routing discovery process. The messages use certificate revocation for detecting expired public keys.

4.3 Security Aware Ad Hoc Routing

Security Aware Ad Hoc Routing (SAR) [18] is selecting route paths using trusted nodes in the routing discovery process is better than selecting the shortest path using unchecked nodes. SAR uses AODV protocol in a trusted hierarchy structure. Nodes in higher level are more trusted than nodes in lower levels. SAR adds a field to each RREQ message; this field represents the security level needed for this route. Intermediate nodes ignore this RREQ if they cannot achieve the security level required by the requester node. SAR also adds a field to each RREP message; this field represents the maximum security level that can be supported by the discovered route. SAR uses a key shared by trusted nodes to encrypt SAR messages..

4.4 Securing DSR protocol

In [19] the authers proposed Secure Routing Protocol (SRP) to secure Dynamic Source Routing Protocol (DSR). SRP secures routing messages by adding SRP headers to these messages. Each header contains a type field that represents the message type, a sequence number that is used to ignore old route messages, a query identifier that is used to verify the freshness of the route, and a MAC (Message Authentication Code) that is used to verify the message validation. SRP uses a secret key shared between the node requester and the final destination. This key is used to sign the non-mutable fields of the packet. Originator nodes generate MAC value by applying a hash function on the non-mutable fields. Destination node verifies the route request validation by applying the same hash function on the non-mutable fields and comparing the result with the MAC value coming with the request. Intermediate nodes just add their address to the route request addresses list and rebroadcast the request. SRP does not need to protect the mutable fields (hop counts) in the route messages; because even the hop count was altered maliciously; it will be detected since SRP is a source routing protocol.

4.5 ARIADNE and ENDAIRA protocols

Secure On-Demand Routing Protocol for Ad hoc Network, ARIADNE [20], is also proposed to secure DSR. Similar to SRP, it requires pre-deployment of authentication keys between the source and destination. Ariadne provide three key sharing approaches corresponding to three authentication methods: pair wise shared secret keys, TESLA keys; Shared secrets between communicating nodes combined with broadcast authentication; and digital signature. Pair wise shared secret keys authenticate DSR routing messages by using secret key between each pair of nodes. This requires n(n-1)/2 keys for a network consisting of n nodes. Pair wise shared secret keys avoid need for synchronization. TESLA requires time synchronization which is difficult to achieve in MANET environments. Each node should have a hash chain; the authentic element of each hash chain should be distributed to all network nodes. Also digital signature requires pre-deployed asymmetric cryptography for the authentication process.

In [21] another routing protocol called ENDAIRA which is the reverse word of ARIADNE is signing the route replays instead of signing the route requests as in ARIADNE. ENDAIRA is more suitable than ARIADNE for MANET environments that have limited resources.

Secure reactive routing protocol present in Table 1 shows that :

- The main difference between ARAN and SAODV is that SAODV uses the route that has the least number of hops, while ARAN uses the first route discovered without comparing the hop counts value. Another difference is that Intermediate nodes in SAODV can respond to RREQ if they keep a valid route to destination. While in ARAN, intermediate nodes cannot respond to RDP.
- SAR is not suitable for some MANET environments because of the overhead caused by the authenticity checking processes.
- SRP cannot prevent malicious nodes from sending wrong route error messages which affect the performance of the protocol.
- ENDAIRA is more suitable than ARIADNE for MANET environments that have limited resources.

Table 1 : Secure Reactive Routing protocols

Protocol	Analysis
SAODV	**a.** able to handle many attacks leveraging modification, fabrication, or impersonation **b.** on ensuring that nodes do not impersonate other nodes and that nodes forwarding routing

	messages do not alter them while those messages are in transit. **c.** cannot protect against that does not increment the hop count when it forwards a routing message(wormhole attack)
ARAN	**a.** uses the first route discovered without comparing the hop counts value **b.** intermediate nodes cannot respond to RDP.
SAR	a. overhead caused by the authenticity checking processes.
SRP	**a.** does not need to protect the mutable fields (hop counts) in the route messages **b.** cannot prevent malicious nodes from sending wrong route error messages which affect the performance of the protocol.
ENDAIRA and ARIADNE	a. ENDAIRA is more suitable than ARIADNE for MANET environments that have limited resources. b. ENDAIRA requires signing the route replays coming just from intended nodes. This requires less resources power than signing the route requests broadcasted to all network nodes.

5 Discussion and Summary

Security is the most important concern for the basic functionality of the network. Any network should be provided with security services to the users. All these services integrate each other to give complete protection. There is no single mechanism that can provide all the security services. Attack prevention measures, such as authentication and protocol encryption can be used as the first line of defense for reducing the possibilities of attacks in MANET. However, these techniques have a limitation to the effects of prevention techniques in general, and they are designed for a set of known attacks. Intrusion detection system comes as second layer of defense for strengthening security in MANET. One of the MANET vulnerabilities comes from the weaknesses of routing protocols. There are many attacks, such as black hole, selfish, rushing, wormhole, and DoS attacks can be generated by malicious node to cripple MANET operation. Unfortunately, most proposed routing protocols at present day do not specify schemes to protect against such attacks.

This paper consolidated various works related to prevention mechanisms that can achive by secure routing protocols. The previous sections described and discussed all major proposed solutions to secure routing protocol against external attack.

Overall, a significant amount of work has been done on prevent MANET from malicious node that do modification, fabrication, or impersonation. Clearly the problem is so broad that there is no way to devise a general solution. Secure routing cannot protect the network from the malacious node that autherized as apart of the network.

References

[1] Zhou, L., Schneider, F. and Van Renesse, R. (2003). COCA: a secure distributed online certiffication authority. Foundations of Intrusion Tolerant Systems, 2003.

[2] Zhou, Z. and Huang, D. (2008). Computing cryptographic pairing in sensors.

[3] Kim, J. and Tsudik, G. (2009). SRDP: Secure route discovery for dynamic source routing in MANETs.

[4] Subhadrabandhu, D., Sarkar, S. and Anjum, F. (2004). Efficacy of misuse detection in ad hoc networks. Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004.

[5] Vigna, G., Gwalani, S., Srinivasan, K., Belding-Royer, E. M. and Kemmerer, R. A. (2004). An intrusion detection tool for AODV-based ad hoc wireless networks.

[6] Clausen, T. and eds, J. (2003). Optimized Link State Routing Protocol (OLSR). IETF RFC 3626. Retrievable at http://www.ietf.org/rfc/rfc3626.txt.

[7] C. E. Perkins, E. M. B. Royer, and S. R. Das, Ad hoc On-Demand Distance Vector (AODV) routing, RFC 3561, July 2003.

[8] Perkins, C. E. and Bhagwat, P. (1994). Highly dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers. In SIGCOMM '94: Proceedings of the conference on Communications architectures, protocols and applications.

[9] Hu, Y.-C., Johnson, D. and Perrig, A. (2002). SEAD: secure efficient distance vector routing for mobile wireless ad hoc networks. Mobile Computing Systems and Applications, 2002.

[10] A. Halfslund, A. Tonnesen, et al., “Secure Extension to the OLSR Protocol,” OLSR Interop and Workshop, 2004.

[11] C. Adjih, T. Clausen, et al., “Securing the OLSR Protocol,” Proceedings of Med-Hoc-Net, June 2003

[12] D. Raffo, T. Clausen, et al., “An Advanced Signature System for OLSR,” SASN’04, October 2004.

[13] 24. D. B. Johnson et al., “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR),” IETF Draft, draft-ietf-manet-dsr-10.txt, July 2004.

[14] Hu, Y.-C., Johnson, D. and Perrig, A. (2002). SEAD: secure efficient distance vector routingfor mobile wireless ad hoc networks. Mobile Computing Systems and Applications, 2002.

[15] M. Zapata and N. Asokan. "Securing ad hoc routing protocols". In Proceedings of the ACM Workshop on Wireless Security (WiSe 2002), Atlanta, GA, September 2002.

[16] Manel Guerrero Zapata, "Secure Ad hoc On Demand Distance Vector (SAODV) Routing", Technical University of Catalonia (UPC), Mobile Ad HocNetworking Working Group, Internet Draft, 15 September 2005.

[17] Sanzgiri, K., Dahill, B., Levine, B., Shields, C. and Belding-Royer, E. (2002). A secure routing protocol for ad hoc networks. Network Protocols, 2002. Proceedings. 10th IEEE International Conference on, 78{87. ISSN 1092-1648.

[18] Yi, S., Naldurg, P. and Kravets, R. (2001). Security-aware ad hoc routing for wireless networks. In Proceedings of the 2001 ACM International Symposium on Mobile Ad Hoc Networking and Computing: MobiHoc 2001. 6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002)

[19] Papadimitratos, P. and Haas, Z. J. (2002). Secure Routing for Mobile Ad hoc Networks. In in SCS Communication Networks and Distributed Systems.

[20] Hu, Y.-C., Perrig, A. and Johnson, D. B. (2005). Ariadne: A secure on-demand routing protocol for ad hoc networks.

[21] Buttyan, L. and Vajda, I. (2004). Towards provable security for ad hoc routing protocols. In SASN'04 - Proceedings of the 2004 ACM Workshop on Security of Ad Hoc and Sensor Networks.

[22] Y. A. Huang and W. Lee, "Attack analysis and detection for ad hoc routing protocols," in The 7th International Symposium on Recent Advances in Intrusion Detection (RAID'04), pp. 125-145, French Riviera, Sept. 2004.

[23] S. Basagni, K. Herrin, et al. "Secure pebblenets". In Proceedings of the 2001 ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2001), Long Beach, CA, October 2001.

[24] J. Binkley and W. Trost. "Authenticated ad hoc routing at the link layer for mobile systems". Wireless Networks, 7(2): 139–145, 2001.

[25] K. Sanzgiri, D. LaFlamme, B. Dahill, B. N. Levine, C. Shields, and E. M. B. Royer, "Authenticated routing for ad hoc networks," IEEE Journal on Selected Areas in Communications, vol. 23, no. 3, pp. 598-610,Mar. 2005.

[26] Y. C. Hu and A. Perrig, "A survey of secure wireless ad hoc routing," IEEE Security & Privacy Magazine, vol. 2, no. 3, pp. 28-39, May/June 2004.

[27] H. Deng, W. Li, and D. P. Agrawal, "Routing security in ad hoc networks, Oct. 2002.

Adaptive coded modulation and power control over Cognitive Radio Network for QoS provisioning

Kyungsup Kim
Department of Computer Engineering
Chungnam National University
Goong-dong, Yuseong, Daejeon, Korea
Email: kskim@cse.cnu.ac.kr

Abstract— *To facilitate the efficient support of quality of service(QoS) in Cognitive Radio network, it is essential to model a wireless channel and traffic in terms of QoS metrics such as data rate, delay and error rate. We propose and develop a wireless channel model termed the effective capacity in Cognitive Radio network. We model the effective capacity as a function of SNR and BER in fading channel. We propose adaptive power control and adaptive rate methods in adaptive coded modulation scheme to maximize the effect capacity. We illustrate the advantage of our approach with a set of numerical simulation experiments.*

1. Introduction

Cognitive radio technology is the key technology that enables an next generation to use spectrum in a dynamic manner. We have witnessed a huge increase in the demand for radio spectrum in wireless communication. Since Spectrum is inherently a limited natural resource, the government agencies in each country regulated the radio resource access. The traditional approach to spectrum management is very inflexible due to exclusive license for most of frequency band and tight limits to the permitted user. Recent measurements for spectrum use indicate that many portions of the spectrum are either unused or lighted used [1]. A new communication paradigm is necessary to solve the limited available spectrum and the inefficiency usage. A new dynamic spectrum access method is proposed, which is called as cognitive radios.

A cognitive radio should support the capability to select the best available channel. Specially, we are interested in the cognitive capability to capture or sense the information from its radio environment. This capability cannot simply be realized by monitoring the power of frequency band and traffic rate. More sophisticated techniques are required in order to capture the temporal and spatial variations in the radio environment [1].

Considering the spectrum sharing channels, where different users share the same spectrum, a constraint on the received power at a third party user's receiver may be a more relevant constraint than a maximum on the transmit power [2]. A QoS driven power and adaptation scheme maximizing the system throughput in terms of the effective capacity was introduced in [3]. We investigate the maximum throughput of the secondary user's channel in Cognitive radio network with delay QoS constraint by obtaining the effective capacity of the channel. We assume that the successful operation of the primary user requires a minimum rate to be supported by its channel for a certain percentage of time, and determine a lower bound on the effective capacity of the secondary user's link in fading channel, when the transmission power of the secondary user adheres with the above mentioned interference limiting.

To facilitate the efficient support of quality of service(QoS) in Cognitive Radio network, it is essential to model a wireless channel and traffic in terms of QoS metrics such as data rate, delay and channel capacity. We propose and develop a wireless channel model termed the effective capacity for Cognitive Radio network. Integrating information theory with the effective capacity, we investigate the QoS driven adaptive power control, coding and modulation over cognitive radio network. The problem is how to maximize the throughput subject to a given QoS constraint.

The remainder of this paper is organized as follows. The system modeling of cognitive radio Network, the definition of effective capacity and the optimizing problem to maximize the effective capacity will be explained in Section 2. In Section 3, we propose the optimal power control and rate adaptation in adaptive MQAM modulation to maximize the effective capacity. In section 4, we discuss the optimal power control and rate adaptation in non-binary coded modulation schemes. Finally, the paper conclude with section 5.

2. Optimization problem in CRN

2.1 System modeling of cognitive Radio Network

A cognitive radio shall sense the environment (cognitive capability), analyze and learn sensed information (self-organized capability) and adapt to the environment (reconfigurable capabilities(CF)). CF must have reconfigurable capabilities such as Frequency agility, Dynamic frequency selection, Adaptive modulation/coding (AMC) and Transmit power control (TPC) and Dynamic system/network access. We consider an adaptive modulation and coding and variable power control scheme for high-speed data transmission over fading channels.

We assume that n secondary transmitters are trying to send their data to secondary receivers in the presence of m primary transmitter. We consider a spectrum-sharing system in which a secondary user is allowed to use the spectrum occupied by a primary user. CR is required to determine the spectrum band allocation that meets the QoS requirements for different users with various applications.

We consider a discrete-time block fading channel with perfect channel side information(CSI) at the receiver and transmitter of the secondary user. The received signal at the secondary receiver $y_s[n]$ depends on the transmitted signal $x_s[n]$ according to

$$y_s[n] = \sqrt{h_s[n]}x_s[n] + z_s[n] \tag{1}$$

where $h_s[n]$ denotes the channel gain between the transmitter and the receiver of the secondary user and z_s represents the additive white Gaussian noise (AWGN). We treat the interference from the primary transmits with with constant power P_p and define the channel gain between the secondary's transmitter and the primary's receiver by $h_{sp}[n]$ and the gain between the transmitter and the receiver of the primary user by $h_p[n]$. We assume that h_s, h_{sp} and h_p are independent from each other and from the noise. We assume that perfect knowledge of h_p is available at the receiver and transmitter of the secondary user. The information about h_p can be carried out by a band manager.

2.2 Effective Capacity for link-layer channel

We consider the link-layer channel model termed the effective capacity defined in [4] [5] [6]. The effective capacity has been introduced by Wu and Negi [4] as a link layer channel model for supporting QoS requirements. The effective capacity is the dual of the effective bandwidth [5]. Effective capacity-based QoS measure model translated into connection level QoS measure such as data rate, delay and delay-violation probability was introduce in [6]. The effective capacity is defined as the maximum constant arrival rate that a given service process can support in order to guarantee a QoS requirement specified by θ.

Let $R(t)$ be the instantaneous channel capacity at time t. Define $\mathfrak{S}(t) = \int_0^t R(t)dt$, which is the service process provided by the channel. The service process $\mathfrak{S}(t)$ depends on the instantaneous channel capacity and is independent of the arrival $\mathfrak{A}(t)$. We note that the channel service $\mathfrak{S}(t)$ is different from the actual service received by the source. We assume that $\Lambda(-\theta) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{-\theta\mathfrak{S}(t)}]$ exists for all $u \geq 0$. Then the effective capacity function of the service process $R(t)$ is defined as

$$E_C(\theta) = \frac{-\Lambda(-\theta)}{\theta} = -\lim_{t\to\infty} \frac{1}{t\theta} \log\left(\mathbb{E}\{\exp(-\theta\mathfrak{S}(\gamma))\}\right).$$

for $\theta \geq 0$ and $\mathbb{E}(X)$ is defined by the expectation of random variable X. When the $R(t)$ is an uncorrelated process, the effective capacity $E_c(\theta)$ reduces to

$$E_C(\theta) = -\frac{1}{\theta}\log(\mathbb{E}\{\exp(-\theta R)\}). \tag{2}$$

The effective capacity can be considered as the maximal throughput under the QoS exponent θ. We can formulate an optimization problem to maximize the effective capacity for a given θ. For a dynamic queueing system with stationary ergodic arrival and service process, the queue length process $Q(t)$ converges such as $\theta = \lim_{x\infty} \frac{\log(Pr\{Q(\infty)>x\}}{x} = \theta$. A smaller θ corresponds to a slower decay rate(looser QoS), while a larger θ leads to a faster decay rate (stringent QoS). So θ is called the QoS exponent [4].

2.3 The problem Optimizing the effective capacity

We formulate the designs of CR networks with diverse QoS guarantees into optimization problems. To provide explicit QoS guarantees such as a data rate, delay bound, and loss probability, it is necessary to analyze CR network as a QoS provisioning system for the fading channels. We consider the rate R termed as capacity. Assume an additive white gaussian noise (AWGN) fading channel with and ergodic channel gain $g(i)$. If the channel fade level is known at the transmitter, then Shannon capacity $R(\gamma) = B\log(1+\gamma)$ is achieved by adapting the transmit power, date rate and coding scheme [7]. Usually, the instantaneous received SNR is $\gamma(t) = \frac{\bar{S}}{N_0 B}$, where $\bar{S}$ denote the transmit power power, B denote the received signal bandwidth and $N_0/2$ is the density of noisy signal. Assuming Gaussian signal inputs, γ is denoted by the signal per interference-plus-Noise rate (SINR) $\gamma = \frac{Sg_i}{\sigma}$ where σ is interference-plus-noise power.

For give SNR γ and the power control law μ, the adaptive code modulation can be achieve capacity. Thus, the instantaneous service rate R can be expressed as

$$R(\gamma) = B\log_2(1+\mu\gamma). \tag{3}$$

The important problem is to maximize the rate R of the secondary user subject to individual peak transmission power constraints of each SU, the interference power constraints to the PUs and the error probability constraints (Bit error rate(BER) or Symbol error rate (SER)). Set $E_C(\theta) = \frac{-1}{\theta}\log(\mathbb{E}\exp(-\theta R(\gamma)))$. The design problem of CR networks is formulated by the optimization problem to maximize the effective capacity $E_C(\theta)$ with constraints as follows:

$$\begin{aligned} \max_{P} \quad & E_C(\theta) \\ \text{subject to} \quad & S_s \leq S_{max} \\ & \mathbb{E}(S_s) \leq \bar{S} \\ & S_p \leq S_{th} \\ & P_b \leq \bar{P}_b \end{aligned} \tag{4}$$

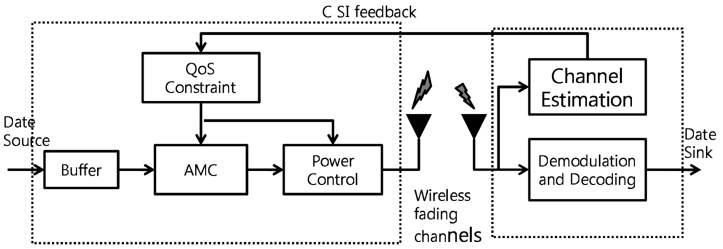


Fig. 1: System model Adaptive coded modulation

for given thresholds S_{max}, $\bar{S}$, $\bar{P}_b$ and QoS exponent θ, where S_s is the transmission power of SU, S_p is the transmission power of PU , and P_b is bit error rate. We study the constraint on the service outage level of the primary user considering that the information about h_{sp} is available to the secondary user. The transmission power of the secondary is limited such that the primary receiver is provided with a minimum-rate R_{min}. Applying the power adaptation, the instantaneous transmit power becomes $S = \mu\bar{S}$. Then the third constraint satisfy the mean power constraint: $\mathbb{E}(\mu) \leq 1$.

3. Adaptive Power controls and Modulation

We first assume that there is no restriction on the constellation size of adaptive MQAM, which implies that the rate of the service process can be adapted continuously. We study the scenario where the transmitter employs adaptive MQAM modulation.

The spectral efficiency of MQAM scheme equals its data rate per unit bandwidth (R/B). M denotes the number of points in each signal constellation. For rectangular signal constellation for M, the exact symbol error rate for the M-QAM under Gaussian channel is given as [8]

$$P_b(\gamma, M) = 1 - \left(1 - 2(1 - \frac{1}{\sqrt{M}})Q\left(\sqrt{\frac{3}{M-1}\gamma}\right)\right) \quad (5)$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-t^2/2)dt$. The Q function is not easily inverted to obtain the optimal rate for a given target BER. In [9], a tighter bound of the BER for $M \geq 4$ and $0 \leq \gamma \leq 30dB$ can be written in the following form:

$$P_b(\gamma, M) \leq 0.2\exp\left(\frac{-1.5\gamma}{M-1}\right).$$

Note that these expressions are only bounds, so they donąŕt match the error probability expressions. We use these bounds since they are easy to invert, so we can obtain M as a function of the target P_b and the power adaptation policy.

We now consider adapting the transmit power $S(\gamma)$ relative to γ, subject to the average power constraint $\bar{S}$ and and instantaneous BER constraint $P_b(\gamma) \geq P_b$. Define a power-control policy μ. The approximate probability of BER P_b for each value of γ can be written by

$$P_b(\gamma, M) \approx 0.2\exp\left[\frac{-1.5\gamma}{M-1}\mu\right] \quad (6)$$

as a function of (γ, M). We adjust M and μ to maintain the target P_b. M is chosen such that P_c is maintained

$$M(\gamma) = \max\{M : P_b(\gamma, M) \leq P_b\}$$

For each given received SNR γ , the corresponding constellation size $M(\gamma)$ is determined by

$$M(\gamma) = 1 + K\mu\gamma \quad (7)$$

where K is defined as $K = -\frac{1.5}{\log(5P_b)}$. Continuous rate MQAM is originally proposed to investigate the insight relationship between the Shannon capacity and the achievable spectral efficiency of MQAM modulation [9].

The variable rate and power control techniques can be applied for other M-ary modulations. The transmit power and constellation size are adapted to maintain a given fixed BER for each symbol while maximizing average data rate. The approximate probability of bit error for the modulation can be written in the following form:

$$P_b(\gamma, M) = c_1\exp\left[\frac{-c_2\gamma\mu}{2^{c_3 R} - c_4}\right] \quad (8)$$

where $R = \log_2 M$ and c_i for $i = 1, \cdots, 5$ are positive fixed constants. The probability of bit error for most M-ary modulation can be approximated in this form with appropriate curve fitting. We compute the coefficients by fitting tool. Set $K = \frac{-c_2}{2^{c_3 R} - c_4}$. The result of computation result $\begin{bmatrix} c_1 & K \end{bmatrix}$ is $\begin{bmatrix} 0.1720 & 2.2635 \end{bmatrix}$. In Fig. 2, we can see that the BER from the computed approximation are close to exact BER and simulation BER.

We can invert the BER in (8) to express the rate R as a function of SNR γ, power adaptation policy μ and the error probability P_b as follows

$$R(\gamma, \mu, P_b) = \frac{1}{c_3}\log_2\left(c_4 - \frac{c_2\gamma\mu}{\log(P_b/c_1)}\right). \quad (9)$$

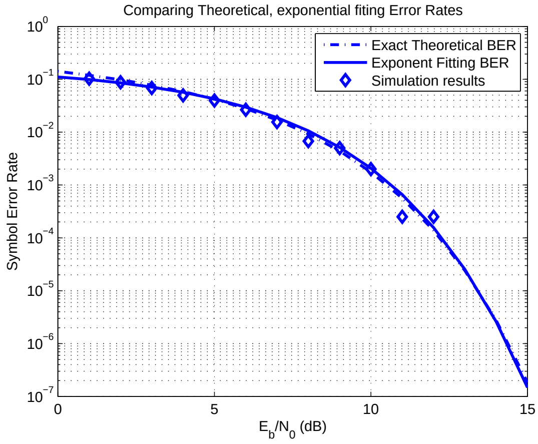


Fig. 2: Exponent fitting BER, Exact BER and Simulation BER for MQAM

We design the resource allocation policy in the concept of the optimization problem of effective capacity. We define the effective capacity by

$$E_C = -\frac{1}{\theta}\log \mathbb{E}\{\exp(-\theta R)\} \tag{10}$$

where R is defined by (9) and θ is the QoS exponent. The resource allocation policy can be expressed as a function of the instantaneous network CSI γ and the QoS exponent θ. Let us define $\nu = (\theta, \gamma)$ as network state information. For a adaptive power control policy ν, the optimize resource allocation policy is to maximize the effective capacity. Our power adaptation policy, denoted by $\mu(\theta, \gamma[i])$ is a function of not only the instantaneous SNR $\gamma(i)$, but also the QoS exponent θ. Applying the power adaptation, the instantaneous transmit power becomes $P(i) = \mu(\theta, \gamma(i))\bar{P}$ where the mean transmit power is upper-bounded by $\bar{P}$. We assume that given network information $(\theta, \gamma[i])$ and the corresponding power-control law $\mu((\theta, \gamma[i]))$, the adaptive modulation and coding scheme can achieve the effective capacity $E_C(\theta)$ for given θ as in (2). By solving the optimization problem with constraints for a given QoS exponent θ, we can derive well-known power controls as follows:

1) The optimal policy, which maximizes the effective capacity, is determined by

$$\mu_{opt}(\theta,\gamma) = \begin{cases} \dfrac{1}{\gamma_0^{\frac{1}{\beta+1}}\gamma^{\frac{\beta}{\beta+1}}} - \frac{1}{\gamma}, & \text{if } \gamma \geq \gamma_0; \\ 0, & \text{otherwise.} \end{cases}$$

2) The water-filling formula, which is well-known the optimal power allocation strategy without delay constrain, is determined by

$$\lim_{\theta\to 0} \mu_{opt}(\theta,\gamma) = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma}, & \text{if } \gamma \geq \gamma_0; \\ 0, & \text{otherwise.} \end{cases}$$

3) As the QoS exponent $\theta \to \infty$, the optimal power control converges to the policy of the total channel inversion determined by

$$\lim_{\theta\to 0} \mu_{opt}(\theta,\gamma) = \frac{\sigma}{\gamma}$$

where $\sigma = \frac{(m-1)\bar{\gamma}}{m}$ for $m \geq 1$.

We plot the instantaneous power adaptation policy in Fig. 3. We can verify that as QoS exponent θ, the corresponding optimal power-adaptation policy swings between the water-filling and the total channel inversion schemes as the results in [3].

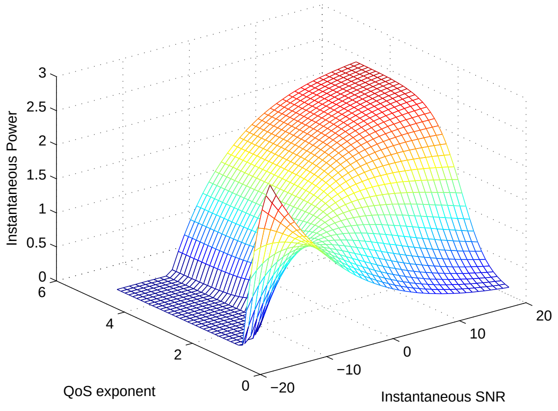


Fig. 3: The optimal power-adaptation policy

4. Adaptive coded modulation

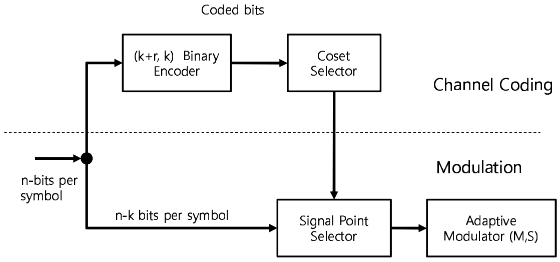


Fig. 4: Adaptive coded Modulation

We consider channel coding and decoding to yield performances close to the Shannon capacity. There are two major coding classes, namely, linear block codes and convolutional codes [8]. Efficient error control on time-varying channels can be performed, independent of modulation, by implementing an adaptive control system in which the optimum code is selected according to the actual channel conditions. We consider on adaptive trellis-coded modulation of varying complexity for the channel coding as in [7] and [10]. In order to implement the adaptive coding scheme, it is necessary again to use a return channel state information. The channel state estimator determines the current channel state, on the basis of the number of erroneous blocks. Once the channel state has been estimated, a decision is made by the reconfiguration block whether to change the code, and the corresponding messages are sent to the encoder and locally to the decoder. The adaptive error protection is obtained by changing the code rates.

The structure of adaptive trellis-coded modulation is shown in Fig. 4. In the channel coded modulation, a binary encoder E, block or convolutional, operates on k uncoded data bits to produced $k+r$ coded bits. The coset selector uses these coded bits to choose one of the 2^{k+r} cosets from a partition of the signal constellation. When the encoder E is a convolutional encoder, the coded modulation is referred as a trellis code; for D a block encoder, it is called a lattice code [11]. The channel is assumed to be slowly fading so that $\gamma(t)$ is relatively constant over many symbol periods. During a given symbol period $T(\gamma)$ are functions of the channel SNR γ the size of each coset is limited to $2^{n(\gamma)-k}$, where $n(\gamma)$ and $T(\gamma)$ are functions of the channel SNR γ.

A channel code is designed for the source code to minimize end-to-end distortion over the given channel with some distortions. We call it source-optimization channel coding. We employ a simple but very effective implementation of source-optimization channel coding using rate-compatible punctured convolutional (RCPC) codes introduced by Hagenauer [12]. In order to explain RCPC, a convolutional code with the mother code rate $R = 1/N$ and memory L is punctured periodically with period P. We get a compatible family of codes with code rate $R = \frac{P}{P+l}$ for $l = 1, \cdots, (N-1)P$. Since there are k bits per code word, $E_s = kE_b$ and since each code word conveys k bits of information, the energy per information bit is $E_b = \frac{E_s}{k}$. The bit error rates for a convolutional code can be upper bounded by the union bound

$$P_b \leq \frac{1}{P} \sum_{d=d_{free}}^{\infty} c_d P_d \tag{11}$$

where P_d is the probability that the wrong path at distance d is selected and c_d is a small information error weight on all paths with $d \geq d_{free}$. The error coefficients c_d are tabulated by Hagenauer [12]. For an AWGN channel, P_d is given by

$$P_d = \frac{1}{2} Q(\sqrt{2d\frac{E_s}{N_o}})$$

If the BER approximated is adjusted for the coding gain, the approximated probability of BER P_b is expressed by

$$P_b(\gamma) = c_1 \exp(-c_2 \frac{\gamma G_c}{2^{c_3 R} - c_4}) \tag{12}$$

for a particular SNR γ, where M is the size of transmit signal constellation and G_c denote denoted the coding gain of the coset code. As in the uncoded case, we can adjust the number of constellation points M and signal power policy relative to instantaneous SNR to maintain a fixed BER. Rearranging (12) yields the maximum constellation size M as a function of γ and P_b:

$$M(\gamma, P_b) = c_4 - \frac{c_2 G_c \mu \gamma}{\log(P_b/c_1)}. \tag{13}$$

The selected point in the selected coset is one of $M(\gamma) = 2^{n(\gamma)+r}$ points in the transmit signal constellation. Since r redundant bits are used for the channel coding, $\log_2 M(\gamma) - r$ bits are sent for a received SNR of γ. The average rate of the adaptive scheme is given by

$$R = \int_{\gamma_0}^{\infty} (\log_2 M(\gamma) - r) p(\gamma) d\gamma \tag{14}$$

where $\gamma_0 \geq 0$ is a cutoff fade depth below which transmission is suspended. We define the effective capacity by $E_C = -\frac{1}{\theta} \log \mathbb{E}\{\exp(-\theta R)\}$. The resource allocation of adaptive coded modulation is achieved by the adaptive rate, adaptive error correct and power control. The resource allocation problem is formulated by maximizing Effective capacity as the adaptive modulation.

5. Conclusion

We proposed and analyzed the adaptive modulation coding(AMC) problem with QoS constraint in CR networks by applying the concept of effective capacity. We developed the optimal power and rate adaptation method for adaptive M-QAM modulation and adaptive coding using rate-compatible punctured convolutional code. We formulated the effective capacity for adaptive M-QAM modulation and adaptive coding. We derived the optimal power policy and rate adaptation to maximize the the effective capacity. The simulation and numerical computation results verified that the optimization to maximize capacity can be applied reasonably to AMC problems.

.

References

[1] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[2] M. Gastpar, "On capacity under received-signal constraints," in *In Proc 2004 Allerton Conference*, 2004.

[3] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *Wireless Communications, IEEE Transactions on*, vol. 6, no. 12, pp. 4349 – 4360, 2007.

[4] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. vol. 2, no. no. 4, pp. 630–643, 2003.

[5] C.-S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, pp. 913–931, 1994.

[6] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *Mob. Netw. Appl.*, vol. 11, no. 1, pp. 91–99, 2006.

[7] A. Goldsmith and S.-G. Chua, “Adaptive coded modulation for fading channels,” *Communications, IEEE Transactions on*, vol. 46, no. 5, pp. 595 –602, 1998.

[8] J. G. Proakis, *Digital Communications*, 4th ed. McGraw Hill Higher Education, 2000.

[9] A. J. Goldsmith, S. ghee Chua, and A. Member, “Variable-rate variable-power mqam for fading channels,” *IEEE Trans. Commun*, vol. 45, pp. 1218–1230, 1997.

[10] D. Goeckel, “Adaptive coding for time-varying channels using outdated fading estimates,” *Communications, IEEE Transactions on*, vol. 47, no. 6, pp. 844 –855, 1999.

[11] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[12] J. Hagenauer, “Rate-compatible punctured convolutional codes (RCPC codes) and their applications,” *IEEE Transactions on Communications*, vol. 36, pp. 389–400, 1988.

Data EPC Acquisition System Middleware Device Manager: DEPCAS MDM

I. Abad, C. Cerrada, J. A. Cerrada, R. Heradio[1]
Software Engineering and Computer Systems Department. UNED: Distance Learning University of Spain.
Madrid. Spain

Abstract - *One of the main purposes of RFID middleware is to manage and hide the broad range of device readers existing in acquisition RFID networks. There are two basic solutions to override. First is to define a specification that should be accomplished for every RFID reader inside a network. The second solution is to define an abstraction layer that can translate proprietary protocols and specific characteristics from a vendor to a general solution. In this paper we present the Middleware Device Manager in our proposal of RFID middleware called DEPCAS: Data EPC Acquisition System.*

Keywords: Radio Frequency Identification, RFIT: Radio Frequency Information System, DEPCAS: Data EPC Acquisition System, SCADA: Supervisory Control and Data Acquisition

1 Introduction

Radio frequency identification (RFID) is one of today's most rapidly proliferating technologies for enterprises. Based on the promise of lower operating costs combined with more accurate product and asset information, organizations are introducing RFID technologies in different productive and organization areas: tracking, supply chain, optimize stock levels, enforce tighter security, people surveillance and industry regulations [1] are examples of these uses.

As enterprises explore these potential benefits [2], many are realizing that scaling from a pilot program to a production scale deployment is not a trivial exercise. Without careful planning, it is possible to destabilize business operations as huge volumes of RFID messages generated at the edge of the enterprise flood toward the data center. Over time, this problem could get magnified as RFID tagging at the item level becomes mandated competence pressures.

Most existing application and network infrastructures were not designed with RFID in mind. At the edge, if RFID infrastructure is deployed on general-purpose servers that require installation and ongoing support, manageability problems alone might render RFID cost-prohibitive. In addition, without a unified deployment model of centralized control and distributed enforcement, disconnected areas of RFID infrastructure will grow, creating a familiar set of application and data integration challenges. On the network side, RFID implementations will generate large amounts of new data that will fan in from the edge of an enterprise toward its data center core, placing further demands upon the existing network.

To optimize the use of available network bandwidth and simplify the deployment of RFID technology, solutions are needed to collect and filter raw RFID tag data close to its source and subsequently to correlate, aggregate, and transform this data into meaningful business-level events. These events in turn should be acted upon locally, where appropriate, or securely routed and delivered to back-end inventory management, sales, and financial applications. It is this "intelligent information" that will enable RFID-ready applications to deliver on their promise of increased operational productivity.

The fundamental requirements of this type of middleware are related to the management of infrastructure device readers, the processing information received from auto identification, the business needs to incorporate auto identification and sharing of information [3]. To address these requirements, software architectures that implement this type of middleware proposed a set of layers that specialize in solving different requirements. These layers are based primarily on the architecture proposed by EPCglobal organization[4], including the resolution of a protocol for reading between the middleware and the tag reading devices, called LLRP (Low Level Reader Protocol) and higher RP (Reader Protocol) according to the standards [5] [6], an event process from the reading of RFID tags [7], an assignment of semantic information to the object through an RFID identification

[1] This research has been carried out under contract with the Spanish CICYT through the DPI-2008-05444 project. It also belongs to the activities carried out in the fram of the Robocity2030-II excellence network of the CAM (ref. S2009/DPI-1559)

service [8], and a service to receive or transmit information to business applications from the RFID middleware [9].

Furthermore, the implementations include what is called generically device manager that allows the configuration and the management of every device incorporated to the RFID network. In this article we focus on this software component that is usually included in every RFID middleware and we present the alternative in which we are working for the development of our proposed architecture called DEPCAS (Data EPC Acquisition System). DEPCAS define software architecture for solving the RFID middleware based on system architecture for monitoring and control (SCADA) process. DEPCAS propose the scenario concept to generalize the process that can be performed from the acquisition of auto identification information. The processing of a scenario produces results that contain elements relevant to their semantic direct use in business applications. The graphical interface is called in DEPCAS GUV (Graphical User Viewer) and is designed to standardize access to the information managed in the middleware regardless of the scenario being driving.

Henceforth this paper is structured as follows: Section 2 presents middleware device manager in existing proposals for middleware, which functions are included and how they are integrated with other elements of the middleware, in Section 3 we presents the middleware device manager (MDM) defined in DEPCAS, in Section 4 we presents the MARC (Minimun Abstract Reader Commands) commands to manage different readers vendor inside a unique RFID network, , and finally the section 5 describes implementation prototype of RRTL (RFID Readers Topology Language) using MARC sentences developed in DEPCAS to solve MDM.

2 The RFID device management

There are several existing RFID middleware implementations (Table 1): OSS RFID middleware as Fosstrak, Rifidi or Aspire, RFID middleware platforms as Sun Java System RFID software or IBM WebSphere RFID system and RFID middleware proprietary as Detego RFID Suite Solution or OAT RFID Suite. Every of these solutions include a solution to infrastructure management with two main positions. One solution is to force the use of a specific protocol inside RFID reader's network while the other extreme is to accept any kind of reader and the middleware will provide bundle mechanism to uniform the devices.

Table 1. RFID Middleware and Device Management Components

	Name	Device Management
AutoIDLabs	Fosstrak	LLRP Adaptor + HAL Adaptor
Primari/Univ. Arkansas	Rifidi	Sensor Abstraction Layer
Aspire	Aspire	Hardware Abstraction Layer
Sun Micro Systems.	Sun Java RFID System v3.0	RFID Execution Agent
IBM	IBM WebSphere RFID	WebSphere RFID Device Infrastructure
RF-IT Solutions	Detego	You-R Device Management
OAT Systems	OAT Foundation Suite	OAT Device Manager

The study of these different RFID Device Management components allows obtaining a set of characteristics that are implemented to solve it in the RFID middleware. Between these features we include:

- Device Agnostic.
- Configuration and parameterization of RFID devices.
- Topological RFID devices.
- Start-up Devices.
- Statistical and Report RFID devices data.
- Monitoring Devices in live operation
- Diagnosis of device in live operation
- LLRP implementation
- OSGi Technology

These features and functions define the operation of the middleware device manager in RFID middleware [10][11]. For example, Device Agnostic Feature is provided when middleware infrastructure acts as an integration point for RFID readers, printers, and other infrastructure devices, such as any sensors and actuators. RFID reader and printer manufacturers typically provide complex configuration options, proprietary interface languages, and communication protocols. If RFID Device Infrastructure eliminates the need for RFID integrators to be aware of these complexities and provides a device agnostic interface with which supported devices can be configured hiding the complexity. Other features are related to scope and technological implementation. Some of the middleware device manager support the EPC Global Network specification to connect with RFID Reader. The Low Level Reader Protocol standardizes a set of instructions to configure, send and receive reader information. In the technological side, several middleware device manager have used different software solutions. Between them, the OSGi Architecture is one of the most common specification framework used in our days. The OSGi specification enable components to hide their implementation from others components while communicating through services.

Next table shows the cross results between existing RFID middleware implementation and the device manager set of characteristics presented in this section

Table 2. Middleware RFID/MDM Characteristics[2]

	FOSSTRACK	RIFIDI	ASPIRE	SUN	IBM	DETEGO	OAT F.S.
Device Agnostic	Y	N	Y	Y	Y	Y	Y
Device Configuration	Y	Y	Y	N	N	Y	Y
Topological Configuration	N	N	Y	Y	N	Y	Y
Needs Programming Starting-up Device	Y	Y	Y	N	N	N	N
Statistical Data from Device	Y	N	Y	N	N	Y	Y
Monitoring Data from Device	N	N	N	Y	Y	Y	Y
Diagnosis capabilities	Y	N	Y	N	Y	N	Y
LLRP Implement.	Y	Y	Y	N	N	N	N
OSGi Technologies	N	Y	Y	N	N	N	N

.

3 Middleware Device Manager in DEPCAS

DEPCAS (EPC Data Acquisition System) is a proposed architecture for RFID middleware systems dedicated to data acquisition in real and heterogeneous environments. The scheme proposed is based on the supervisory and control systems known as SCADA (Supervisory Control and Data Acquisition). In this case, the remote terminal units in SCADA systems are replaced by the antenna-reader RFID systems to receive auto identification information. The communication networks are used to connect reader (communication via serial, Ethernet...) to the system Central acquisition software. The basic structure of DEPCAS is divided into four sub-systems: the MDM (middleware device manager) for the management of infrastructure, MLM (middleware logic manager) for auto-identification process, the GUV (graphical user viewer) or man-machine interface, and finally the EPCIS (EPC Information System) as software component to communicate with other systems.

The main objectives of MDM are:

- Manage the topology RFID reader's network.
- Agnostic management of different RFID vendors.
- Homogeneous process of different RFID tag reads.
- Support LLRP and ALE specification process.
- Provide configurable RFID network
- Hide operational environment
- Minimum tag management operations.
- Monitoring and controlling RFID readers.

To support these objectives the MDM definition includes the concept of Minimum Abstract Reader Commands (MARC) and the RFID Reader Topology Language (RRTL). The MARC allow to hide the particularities of every RFID reader vendor through a set of operations that are mapped to specific vendor API. The RRTP includes the description of RFID topologies including control and management of readers.

The MDM core architecture prototype have been developed using TCP and HTTP for transporting reader protocol messages, while the message content can be either XML or Text. In addition, it support synchronous and asynchronous messaging (through the reader's protocol Notification Channels mechanisms).

4 Minimum Abstract Reader Commands (MARC)

MARC is the minimum set of instructions that an RFID reader must supply to support the operational instructions inside a network. The set of MARC is closed and give an independent layer of RFID management. The implementation of MARC is related to RFID reader API and gives a dependent hardware layer. The commands included in MDM are described in Table 3:

Table 3. MARC Commands

MARC	*Operations*
setAntennaSequence	Set an order sequence to read from multiple antennas installed in one reader
setAutoFalseOutput	Order to autonomous function mode to ignore tag reads
setAutoFalsePause	Order to autonomous function mode to pause working
setAutoMode	Start/Stop autonomous reader functions
setAutoStartTrigger	Activate the trigger sending functions in reader autonomous function
setAutoStartTimer	Activate periodic timer reader

[2] Y: Yes/ N: No

	function in reader autonomuos function
setAutoStopTimer	Reset periodic timer reader function in reader autonomous function
setAutoStopTrigger	Reset the trigger sending functions in reader autonomous function
setAutoTrueOutput	Order to autonomous function mode to process tag reads
setAutoTruePause	Order to autonomous function mode to start reading process
setConnectTCP	Establish a TCP connection
setNotifyAddress	Init a listening TCP port to instruct the reader to send notification messages
setNotifyFormat	Configure the format message to be used by the reader
setNotifyTime	Establish the period that the reader notifier should use
setPassword	Identify the password reader access
setRun	Start to execute a configuration reader mode
setTagType	Establish the expected tag type property
setTime	Synchronize reader time
setUsername	Identify the user reader access

The MDM prototype includes MARC implementation for RFID reader ALIEN8780, FEIG2000, SKYWRAEM2 and we are working in MARC implementation for ThingMagic Mercury 4, Intermec IF61 and Impinj Speedway R420.

5 RFID Reader Topology Language (RRTL)

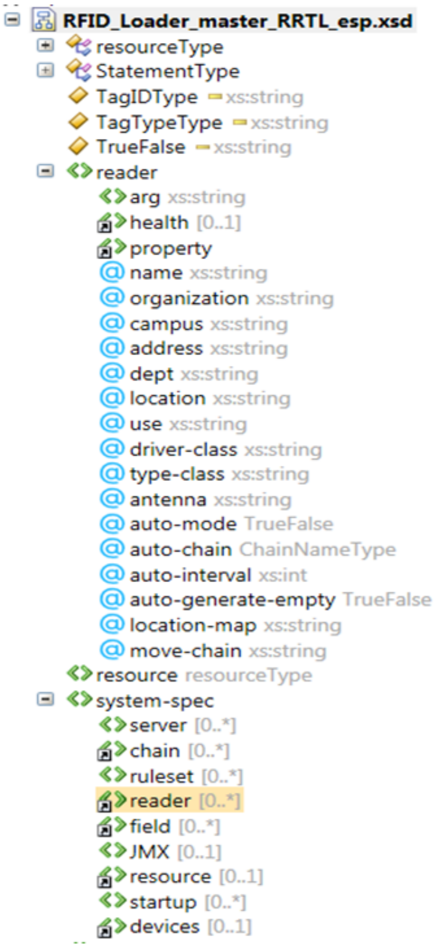
RFID Reader Topology Language (RRTL) is a language to define RFID network topologies that includes control and management of RFID devices. The language manages the process and filter of tag reads to uniform them to DEPCAS system.

The RRTL command could be classified by their application:

1. Resource definition.
2. Time management.
3. Conditional configuration
4. Tag management
5. Device management
6. Control and flow configuration
7. Database management
8. Web service control
9. File management
10. Tag position and code management
11. Process synchronization
12. User commands

There are a set of miscellaneous sentences that are related with different kind of operation like start/stop operation, locks, etc.

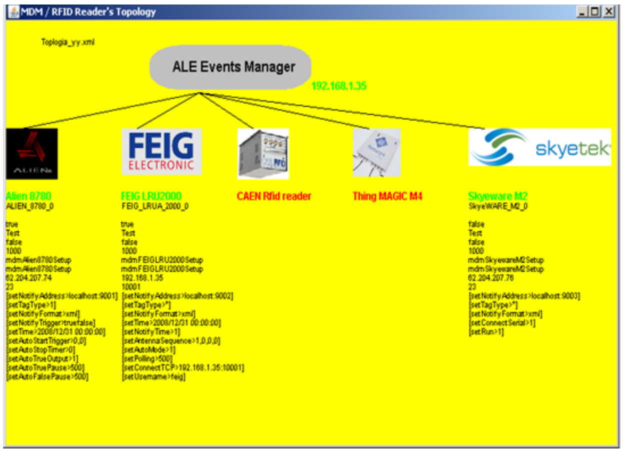
The RRTL definition is supported by an XSD squeme and the results with XML files. An example of some RRTL commands is shown in figure 1:



```
RFID_Loader_master_RRTL_esp.xsd
  resourceType
  StatementType
  TagIDType xs:string
  TagTypeType xs:string
  TrueFalse xs:string
  reader
    arg xs:string
    health [0..1]
    property
    name xs:string
    organization xs:string
    campus xs:string
    address xs:string
    dept xs:string
    location xs:string
    use xs:string
    driver-class xs:string
    type-class xs:string
    antenna xs:string
    auto-mode TrueFalse
    auto-chain ChainNameType
    auto-interval xs:int
    auto-generate-empty TrueFalse
    location-map xs:string
    move-chain xs:string
  resource resourceType
  system-spec
    server [0..*]
    chain [0..*]
    ruleset [0..*]
    reader [0..*]
    field [0..*]
    JMX [0..1]
    resource [0..1]
    startup [0..*]
    devices [0..1]
```

Figure 1. RRTL Reader Command Example

One simple topology with RRTL commands is shown in fig. 2:



```
ALIEN_8780_0

true
Test
false
1000
mdmAlien8780Setup
mdmAlien8780Setup
62.204.207.74
23
[setNotifyAddress>localhost:9001]
[setTagType>1]
[setNotifyFormat>xml]
[setNotifyTrigger>truefalse]
[setTime>2008/12/31 00:00:00]
[setAutoStartTrigger>0,0]
[setAutoStopTimer>0]
[setAutoTrueOutput>1]
[setAutoTruePause>500]
[setAutoFalsePause>500]
```

```
FEIG_LRUA_2000_0

true
Test
false
1000
mdmFEIGLRU2000Setup
mdmFEIGLRU2000Setup
192.168.1.35
10001
[setNotifyAddress>localhost:9002]
[setTagType>*]
[setNotifyFormat>xml]
[setTime>2008/12/31 00:00:00]
[setNotifyTime>1]
[setAntennaSequence>1,0,0,0]
[setAutoMode>1]
[setPolling>500]
[setConnectTCP>192.168.1.35:10001]
[setUsername>feig]
```

```
SkyeWARE_M2_0

false
Test
false
1000
mdmSkyewareM2Setup
mdmSkyewareM2Setup
62.204.207.76
23
[setNotifyAddress>localhost:9003]
[setTagType>*]
[setNotifyFormat>xml]
[setConnectSerial>1]
[setRun>1]
```

Figure 2. Topology example

The RRTL commands for XML data for FEIG configuration in this example is:

```
<?xml version="1.0" encoding="UTF-8" ?>
<system-spec xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="RFID_Loader_master_RRTL_esp.xsd">

<resource>
    <timer-job chain="Test" name="ALIEN_8780_0">
   <time-span milliseconds="1000" /
      </timer-job>
    <timer-job chain="Test" name="FEIG_LRUA_2000_0">
    <time-span milliseconds="1000" />
    </timer-job>
    <timer-job chain="Test" name="SkyeWARE_M2_0">
    <time-span milliseconds="1000" />
    </timer-job>
    <file name="DataFile"
       filename="c:/Uned/MDM/DataFile.mdm"
         append="true"newline-style="Windows" />
</resource>

<reader driver-class="mdmFEIGLRU2000Setup"
         type-class="mdmFEIGLRU2000Setup"
         name="FEIG_LRUA_2000_0"
         organization="UNED"
         campus="Madrid / Univ.EDE"
       address="c/ Juan, 28040-MADRID"
         dept="ISSI"  location="Planta: 1,  Desp: 112"
         use="Test"  auto-mode="true"  auto-interval="1000"
         auto-chain="Test"   >
         <arg>192.168.1.35</arg><arg>10001</arg><arg>feig</arg>
         <arg>password</arg>
          <property>
                <key="setNotifyAddress"/
                <value="localhost:9002"/>
        </property>
         <property key="setTagType"  value="*" trim="false" />
         <property key="setNotifyFormat" value="xml" />
         <property key="setTime" value="2008/12/31" />
         <property key="setNotifyTime"    value="1"/>
         <property key="setAntennaSequence" value="1,0,0,0"/>
         <property key="setAutoMode"     value="1"/>
         <property key="setPolling"  value="500" />
         <property>  <key="setConnectTCP" value="192.168.1.35:1001"/>
         <property key="setUsername"  value="feig"/>
         <property key="setPassword" value="password"  />
         <property key="setRun"    value="1"/>
</reader>

<chain name="Test">
         <poll  generate-empty="false" reader="ALIEN_8780_0" />
         <poll generate-empty="false" reader="FEIG_LRUA_2000" />
         <poll generate-empty="false" reader="SkyeWARE_M2_0" />
         <purifier period="1000" reads="2" generateempty="false"/>
         <echo message="Reading tag ..........id=${tagId}" />
         <transfer resource="RMI to DBMonitor.java"
             forward-resource="localhost:3232"/>
</chain>
</system-spec>
```

6 Conclusions

Regarding the development of MDM for RFID, there are two main focuses. One is to build an agnostic RFID device acquisition system in the same time it accomplish the existing standard proposals. The other one is how to manage topological RFID network questions like RFID levels reader point (antennas, edge readers, reader, PLCs with RFID readers, etc.) or redundant reader points. This two main subject are covered in or solution of DEPCAS MDM, one with the use of MARC to hide device management or standardization protocols and topological questions with RRTL to express network configuration.

About future works there are different issues to be resolved. First, we want to extend the MARC functionality to new devices as Intermec IF61, Impinj Speedway R420 or Siemens RF6660A. Second there are other topological concepts to work in with RRTL as tag tables, conditional operations master/submaster and peer to peer network reader organizations, etc.

7 References

[1] R. Russell, “Manufacturing Execution Systems: Moving toi the next level”, Pharmaceutical Technology, January 2004, pp 38-50.

[2] H. Ramamurthy, B.S. Prabhu, Rajit Gadh. “ReWINS: A distributed multi-RF Sensor Control network for industrial automation”, IEEE Wireless Telecommunications Symposium (WTS 2005), Pormona, California, 2005.

[3] Xiaoyong Su, Chi-Cheng Chu, B. S. Prabhu, Rajit Gadh, “On the creation of Automatic Identification and Data Capture infrastructure via RFID and other technologies” The Internet of Things: from RFID to the Next-Generation Pervasive Networked Systems, Lu Yan, Yan Zhang, Laurence T. Yang, Huansheng Ning (eds.), Auerbach Publications, Taylor & Francis Group, 24 pp. , 2007.

[4] Brewer, A., Sloan, N. and Landers, T.L., “Intelligent tracking in manufacturing”. Journal of Intelligent Manufacturing. v10 i3. 245-250. 1999

[5] C. Floerkemeier, C. Roduner, and M. Lampe, “RFID Application Development with the Accada Middleware Platform”. IEEE Systems Journal. 2007

[6] Q.Z. Shenz, X. Li, S. Zealdally, “Enabling next-generation RFID applications: Solutions and challenges”. Computer, Col 41, Issue 9. 2008.

[7] M C. Bornhövd, T. Lin, S. Haller, and J. Schaper, “Integrating Automatic Data Acquisition with Business Processes - Experiences with SAP’s Auto-ID Infrastructure”. In Proceedings of the 30st international conference on very large data bases (VLDB). Toronto, 1182–1188. 2004.

[8] Curtin, J. Kauffman, R.J., Riggins F.J., “Making the “MOST” out of RFID Technology: a research agenda for the study of the adoption usage and impact of RFID”, Springer Science+Business Media. Abril 2007.

[9] N. Kefalakis, N. Leontiadis, J. Soldatos, D. Donsez, “Middleware Building Blocks for Architecting RFID Systems”, Mobile Lightweight Wireless Systems, Vol 13, pp 325-336. August 2009

[10] Wells,A.: “RFID Detail”, IMPO Magazine, Advantage Businness Media. June 2009

[11] “EPC Information Services (EPCIS)”, EPCGlobal, version 1.0.1, 2007

[12] J. Han, H. Gonzalez, X. Li, D. Klabjan. “Warehousing and Mining Massive RFID Data Sets”, ADMA 2006: 1-18. 2006

Heuristic Resource Allocation with Satisfaction for WiMAX Services

Yung-Liang Lai
945402011@cc.ncu.edu.tw

Jehn-Ruey Jiang
jrjiang@csie.ncu.edu.tw

Department of Computer Science and Information Engineering
National Central University

No.300, Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan

Abstract—**The WiMAX technology offers a versatile mobile services platform, which enables operators to provide differential Quality of Service (QoS) for fulfilling subscribers' various requirements. WiMAX operators need to allocate enough resources to make all subscribers satisfied to avoid "user churn" that subscribers will unsubscribe services due to dissatisfaction with allocated resources. This resources allocation in WiMAX can be modeled as a constrained nonlinear integer problem (Constrained NLIP), which is hard to solve for a large subscriber scale. This paper proposes a heuristic optimization algorithm to allocate available resources for maximizing the operator revenue under the subscribers' satisfaction constrains for differential services. The efficiency of the proposed algorithm is also evaluated.**

Keywords-WiMAX, resource allocation, Quality of service, Constrained Nonliner Integer Problem, Genetic algorithm

I. Introduction

WiMAX (Worldwide Interoperability for Microwave Access) [1] is a wireless broadband access technology that provides performance similar to IEEE 802.11 Wi-Fi networks with the coverage and QoS (quality of service) of cellular networks. It is initiated by WiMAX forum [2] and IEEE 802.16d [3] and 802.16e [4] are two important standards in the WiMAX specifications. The former is to provide last-mile connectivity up to 50km for fixed stations. The latter is to enable convergence of mobile and fixed broadband networks through a common wireless-access technology; it provides connectivity up to 15 km for mobile stations. WiMAX subscribers have broadband network access anytime and anywhere, leading to a rich set of diverse services. Some of the services, such as Voice over IP (VoIP) telephony service and Multimedia Broadcast Multicast Service (MBMS), demand differential QoS requirements. WiMAX operators need to allocate resources to different services to satisfy their requirements. With QoS support [5], the operators can provide better experience for all subscribers in different service levels.

Since resources (e.g., spectrum) are scare in the WiMAX network, they are usually costly. The operators have to invest a very large capital for acquiring the resources in order to provide enough allocation for subscribers, which may debase the revenue. However, if the subscribers are not satisfied with the allocation of resources, they will unsubscribe the services and migrate to another operator, leading to the *user churn* problem [6], which will also reduce the revenue. It is challenging to allocate available resources for maximizing the operator revenue under the subscribers' satisfaction constrains.

The IEEE 802.16e standard [4] defines the physical (PHY) layer and medium access control (MAC) layer for broadband wireless access systems. The OFDMA has been selected by the WiMAX forum [2] as the standard model in the PHY layer. The OFDMA is defined as a multiplexing technique for subdividing the wireless spectrum into a set of frequency resources, which are called subchannels. IEEE 802.16e provides a diversity classes for providing QoS services, which are unsolicited grant service (UGS) class, real-time-polling-services (rtPS) class, non-real-time-polling services (nrtPS) class, and best effort (BE) class. The UGS class supports the reliable wireless access services, such as T1/E1 bandwidth in internet service provider. The rtPS class supports voice (or video) transmissions in real-time. The nrtPS class supports non-real time transmissions, such as file transfer. The BE service supports non-delay sensitive transmissions. Some schemes are proposed to manage the resources for the QoS services of the WiMAX. The authors in [7] proposed a pricing model for providing UGS and rtPS. The paper [8] proposed a variable pricing model for nrtPS. The authors in [9] proposed a mechanism to assign costs for services based on different bandwidth requirements. However, the problem of WiMAX wireless resources allocation with subscribers' satisfaction constraints is not fully studied.

In this paper, we study how to allocate the WiMAX resources in order to maximize the revenue for differential QoS levels under the subscribers' satisfaction constrains. The resource allocation is modeled as a constrained nonlinear integer problem (Constrained NLIP) [10]. Due to the hardness of solving the Constrained NLIP problem, we develop a heuristic genetic optimization algorithm to find the solution. We also evaluate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. In Section II, some related work is described. A resource allocation with satisfaction problem is formulated in Section III for QoS resource allocation in the WiMAX network. The proposed heuristic genetic algorithm is introduced and evaluated in Section IV and Section V, respectively. And finally, some concluding remarks are drawn in Section VI.

II. REALTED WORK

In this section, we introduce the related work about the QoS model of WiMAX, user satisfaction model, and resources allocation schemes.

A. *QoS Model of WiMAX*

The WiMAX network is an IP based mobile network, which provides a well-defined framework to support the QoS based services, such as the Voice over IP (VoIP) telephony service, Internet Access service, and Multimedia Broadcast Multicast Service (MBMS), etc. Based on the 3GPP Rel. 5 specification [11], the network consists of multiple mobile bearers (e.g., mobile phones, notebooks) to provide end-to-end quality of services, where the WiMAX base station (BS) allocates the spectrum resources for providing the wireless services in mobile bearers. The BS can provide diverse classes for differential QoS services, which are unsolicited grant service (UGS) class, real-time-polling-services (rtPS) class, non-real-time-polling services (nrtPS) class, and best effort (BE) class.

Based on the QoS framework, the operator can design and sell products combining different QoS services. As shown in Fig. 1, the subscribers are equipped with different wireless devices to access the WiMAX network, so the base station needs to provide differential services for the subscribers.

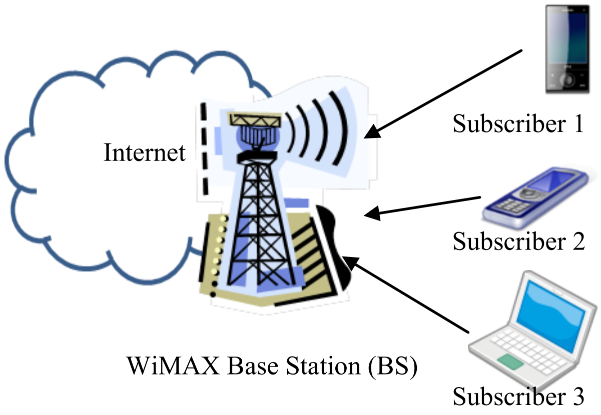


Figure 1. Illustration of WiMAX Services

B. *User Satisfcation Model*

The satisfaction of subscriber is important for an operator, since unsatisfied subscribers will migrate to another operator, leading to the "user churn" problem reported in [6]. Subscribers will feel unfulfilled when some satisfaction requirement is not met. For example, the acceptable one-way (speaker's mouth to listener's ear) delay of voice communication for VoIP applications recommended by ITU [12] is at most 150 ms. The subscriber uses VoIP service will feel unsatisfied if transmission delay is more than 150 ms. A Sigmoid distribution function is proposed in [6] to approximate the subscriber's satisfaction of resource allocations. The function is useful for modeling natural processes or system learning curves shown in Fig. 2, since it can represent a history dependent progression to approach a limit over time. The Sigmoid function depends on a random variable *x* to represent the occupation of resources for the subscriber, which is the utility function formulated as:

$$\Psi(x) = \frac{1}{1+e^{-\alpha(x-\beta)}} \tag{1}$$

In Eq. (1), α and β are steepness and middle of the curve. In the subscribers' point of view, the α is the sensitivity of QoS and the β is the tolerable limit of failure of QoS.

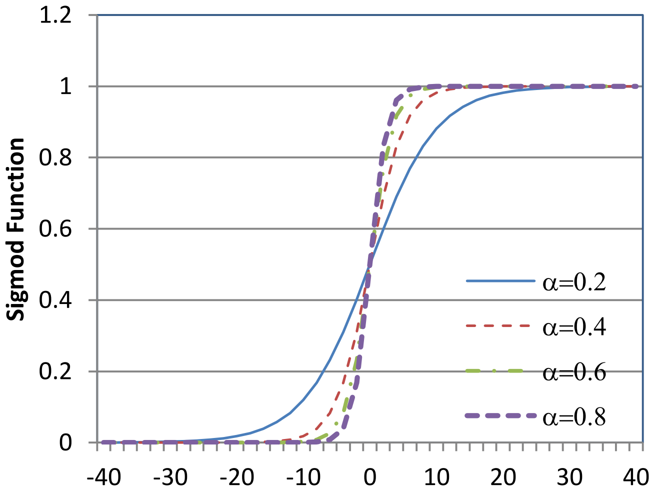


Figure 2. Curves of Sigmod function with β=0 and different α values

C. *Resource Allocation Schemes*

The resource allocation is one of the most important research topics in the WiMAX technology. In [7], the authors proposed a pricing model for unsolicited grant services (UGS) and real time polling services (rtPS). They proposed an evaluation function which defines the price per unit time for controlling subscribers' transmission rates. The paper [8] proposed a variable pricing model for non-real-time-polling service (nrtPS). The paper [9] proposed a cost based admission mechanism by assigning costs for services based on different bandwidth requirements. However, the wireless resources allocation with subscribers' satisfaction constraints problem is not fully studied.

III. Problem Formulation

We formulate the problem of resource allocation with satisfaction in this section. The main goal of the problem is to maximize the operator's revenue under the subscriber's satisfaction constraints.

We assume an operator sells S services to subscribers. The operator allocates SR_s resources for the service s, where $s \in$ services set Ω. The subscribers need to pay the dynamic price $P_{dynamic}$ for their dynamic usage of resources and the static price P_{static} for their subscribed services. The operator has the cost C_{res} for the resources. The subscribers demand DR_s units of resources for the service s; however, the resources consumed should be not more than SR_s. The revenue (R) for the operator can be defined as:

$$\mathrm{R} = \sum_{s\in\Omega} \mathrm{MIN}\{\mathrm{DR}_s, \mathrm{SR}_s\} \cdot \mathrm{P}_{\mathrm{dyamic}} + \sum_{s\in\Omega} \mathrm{SR}_s \cdot \mathrm{P}_{\mathrm{static}} - \sum_{s\in\Omega} \mathrm{SR}_s \cdot \mathrm{C}_{\mathrm{res}} \qquad (2)$$

The first constraint is that at least one resource should be allocated for each service:

$$\forall s\in\Omega,\ \mathrm{SR}_s \geq 1 \qquad (3)$$

We also have to enforce another constraint such that the total number of allocated resources is equal to the resource budget (B) to ensure that every resource is utilized. We thus have

$$\textstyle\sum_{s\in\Omega} \mathrm{SR}_s = \mathrm{B} \qquad (4)$$

Moreover, we have to ensure the service's satisfaction is not less than the satisfaction level (γ_s); i.e.,

$$\forall s\in\Omega,\ \Psi(\mathrm{R}_s) \geq \gamma_s \qquad (5)$$

Based on above definitions, the problem can be expressed by the problem of finding a resource allocation set $\{SR_s|s\in\Omega\}$ for maximizing the operator's revenue, which is defined as following:

$$\text{Maximize R} \qquad (6)$$

subject to resource constraints (3) and (4) and satisfaction constraint (5).

IV. Proposed Heuristic Allocation Algorithm

Our algorithm is based on the genetic algorithm (GA) approach, which can be used to heuristically find the solutions of combinatorial optimization problems. The GA approach is to mimic natural selection in the biology, where individuals with higher fitness can survive to next generation [13].

A population (a set of individuals) is randomly generated in the initial step. Then, the population will be evolved in the generation-loop for MAXIMUM_GENERATIONS times. In each generation, the steps to produce good individuals in the next generation consist of: Selection, Crossover, Mutation, and Production. When the generation-loop is terminated, the solution is made by decoding the best individuals in the Decode step.

Based on the above steps, we design a heuristic resource allocation algorithm. The algorithm is presented as follows.

Heuristic Resource Allocation Algorithm

```
Init_population();
DO
    Selection();
    Crossover();
    Mutation();
    Production();
UNTIL (Terminate_Condition)   /*Condition is
  repeated times reaches MAXIMUM_GENERATIONS */
Solution = Decode();
Return Solution
```

A. Chromosome Representation

We use the chromosome (CH) to represent one feasible solution to allocate B resources to S services. Each chromosome contains (B-S) genes. Each gene is a binary string, which contains $\lceil \log_2 S \rceil$ bits, where $\lceil \cdot \rceil$ stands for the ceiling function. Each gene represents the identity of a service, whose value range is in [0, S-1]. In such design, we can allocate B resources and the allocation solution does not violate the constraint (4). In addition, the algorithm will add one resource to each service in the Decode step due to the constraint (3). According to the above-mentioned design, each feasible solution does not violate constraints (3) and (4).

Below, we use an example to illustrate the design of chromosome (CH) and the Decode step. We assume an operator has to allocate B (B=7) resources for 3 services. The following chromosome is feasible for the resource allocation problem.

CH= [1, 2, 2, 2]

In the Decode step, the above CH represents that the operator needs to allocate 2 units of resources for service 1 and allocate 4 units of resources for service 2.

B. Initialization Population

The population is randomly assigned. It is notable that setting of initial population is fairly simple to our algorithm; according to studies in [13], the population can have good resilience for randomly setting of initial population.

C. Fitness Function

The fitness function is used to evaluate the performance of individuals of the objective. In our design, the performance is the revenue, which implies only the individuals with higher revenue can survive. It is notable that we use the penalty term in (7) to enforce the constraint (5). The λ is the plenty weight, which is of a very large value. Thus, we have the fitness function as follows:

$$F = R - \lambda \cdot PLT \tag{7}$$

The PLT is the plenty function deified in Eq. (8), where the γ_s is the minimum number of required satisfaction levels.

$$PLT = \sum_{s \in \Omega} MAX\{\gamma_s - \psi(SR_s), 0\} \tag{8}$$

V. PERFORMANCE EVALUATIONS

In this section, we evaluate the performance of our proposed heuristic resource allocation algorithm. Our evaluation objective is to evaluate the operator's revenue and subscribers' satisfactions. We implement our algorithm on the Matlab platform [14]. The problem set for testing our algorithm contains the testing cases of 18, 19, …, and 24 resources. The number of services is 16. The setting parameters used in the simulation are listed in the Table I.

TABLE I. SIMULATION SETTING PARAMETERS

Parameter	Setting Value
Crossover Rate	0.7
Mutation Rate	0.0175
Generations	1000
Genre Representation Scheme	Bit string
Plenty weight (λ)	1000
Satisfaction level	0.4
Cost, Price parameters	P_{static} = 3 units, $P_{dynamic}$ =2 units, C_{res} = 10 units

The simulation results show the operator's revenue increases with the number of resources, as shown in Fig. 3. Moreover, the users' satisfactions are larger than satisfaction level (0.4), which implies that the solution in our algorithm does not violate the subscribers' satisfaction constraints, as shown in Fig. 4.

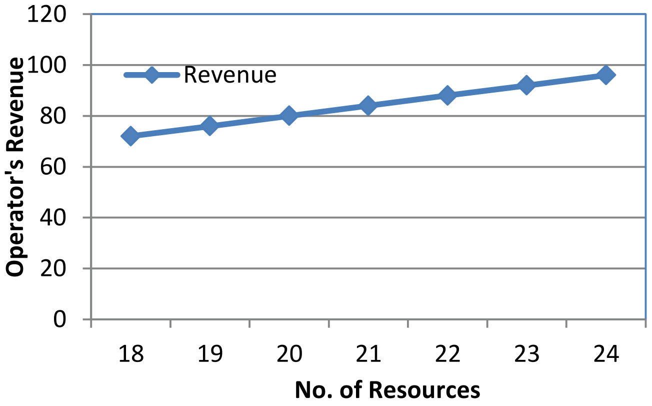


Figure 3. Simulation results of the proposed algorithm in terms of the operator's revenue

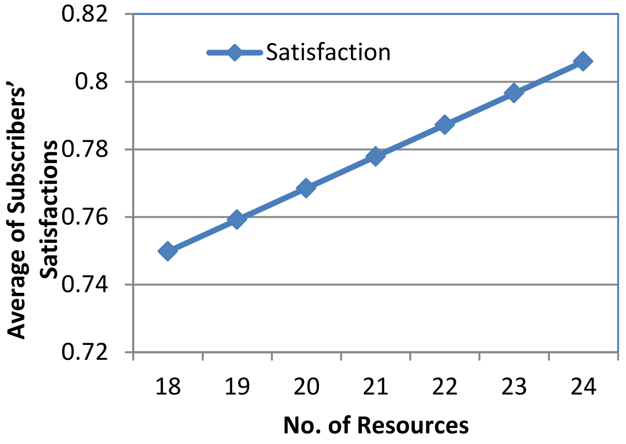


Figure 4. Simulation results of the proposed algorithm in terms of the average of subscribers' satisfactions

VI. Conclusions

WiMAX is a promising technology for wireless broadband Internet access. To successfully deploy a commercial WiMAX system, both the operator's revenue and subscribers' satisfaction constraints must be taken into account. In this paper, we have formulated a problem of resource allocation with satisfaction for properly allocating resources in WiMAX networks. We have proposed a heuristic genetic algorithm for solving the problem. Simulation results demonstrate that the proposed algorithm is effective. Our future work will focus on using different user satisfaction models and different QoS models to formulate the resource allocation problem in WiMAX networks.

References

[1] S. Ahmadi, "An overview of next-generation mobile WiMAX technology, " *IEEE Communications Magazine*, vol. 47, no.6, pp.84-98, June 2009.

[2] WiMAX Forum, 2011, URL: http://www.wimaxforum.org/

[3] IEEE 802.16-2004, IEEE Standard for local and metropolitan area networks, Air Interface for Fixed Broadband Wireless Access Systems, Oct 2004.

[4] IEEE 802.16e, IEEE Standard for local and metropolitan area networks, Air Interface for Fixed Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed, 2005.

[5] B. Li, Y. Qin, C. P. Low, and C. L. Gwee, "A survey on mobile wimax [wireless broadband access], " *IEEE Communications Magazine*, vol. 45, pp. 70-75, 2007.

[6] H. Lin, M. Chatterjee, S. K. Das, and K. Basu, "ARC: an integrated admission and rate control framework for CDMA data networks based on non-cooperative games, " *ACM Mobicom 2003*, pp. 326-338.

[7] A. Belghith, L. Nuaymi, and P. Maille, "Pricing of Real-Time Applications in WiMAX Systems, " *IEEE 68th Vehicular Technology Conference*, Calgary, Canada, 21-24 September 2008.

[8] A. Belghith, L. Nuaymi, and P. Maille, " Pricing of differentiated-QoS services WiMAX networks," *IEEE Global Communications Conference*, New Orlean USA, 30 Novembre - 4 December 2008.

[9] B.-J. Chang Y.-L. Chen, and C.-M. Chou, "Adaptive Hierarchical Polling and Cost-based Call Admission Control in IEEE 802.16 WiMAX," *Wireless Communications and Networking Conference*, 2007, , pp. 1954 - 1958, Hong Kong, 11-15 March 2007

[10] D. Li, X. Sun, *Nonlinear Integer Programming*, Springer, 2010.

[11] 3GPP TR 23.836, "Quality of Service (QoS) and policy aspects of 3GPP – Wireless Local Area Network (WLAN) interworking", December 2005.

[12] One-Way Transmission Time, ITU-T Recommendation G.114, Feb. 1996.

[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[14] Matlab, 2011, URL: http://www.matlab.com

SESSION

NOVEL APPLICATIONS + ROUTING + LANS + PERFORMANCE ISSUES + AD-HOC AND SENSOR NETWORKS + WI-FI

Chair(s)

Prof. Hamid R. Arabnia

Using content-based addresses as a programming construct for scenario specific routing and multicasting

R.P.Bosman[1], **J.J.Lukkien**[1], **and P.H.F.M.Verhoeven**[1]
[1]Department of Mathematics and Computer Science,
Eindhoven University of Technology, Eindhoven, The Netherlands

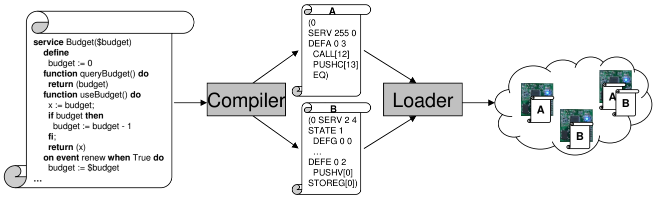
Abstract— *In this paper we extend the notion of content-based addressing with a programmable forwarding scheme which yields a powerful programming construct for expressing scenario specific routing and multicasting. We demonstrate the expressive power of this construct through examples of over-the-air code dissemination and composition of heterogeneous services and communication between them. Furthermore, we show an efficient implementation and indicate how the underlying forwarding mechanism can be replaced to further enhance the expressivity and efficiency*[1]. *The realization of the content-based addresses does not require periodic address announcements and construction of routing tables and can therefore be applied in highly dynamic networks at low cost.*

Keywords: Sensor Networks, content-based addressing, routing, language

1. Introduction

Wireless sensor networks consist of small battery powered nodes with sensing, actuating, processing and communication capabilities. In recent years sensor networks have become a popular topic of research. Sensor nodes are extremely resource constrained and require a significant amount of programmer expertise to manage their energy usage efficiently. To cope with the complexities of programming these nodes, service oriented programming models are becoming popular [1], [2], [3]. In these models, each sensor node provides one or more heterogeneous services which are composed to form an application. This allows programmers to focus on problems in the small while other services in the system address wider issues.

In *macroprogramming*, the sensor network is regarded as a platform on which applications are built as opposed to programming individual nodes. An application is implicitly or explicitly written as a “program” for the entire network. Proposed macroprogramming models include special-purpose ones (e.g. TinyDB [5] that regards the network as a database that can respond to queries) and general purpose ones (e.g. MagnetOS [6], which regards the network as an extended Java virtual machine). A typical approach is to have an interpreter running on the nodes in the network; a compilation and mapping procedure generates the virtual machine code (VMC) for each node. In such a case there are two languages involved: the global macroprogramming language (usually domain-specific) and the virtual machine code language (either fairly general, or sensor node specific). Alternatively, the compilation and mapping procedure can generate native code from the macroprogram. Fig. 1 shows macroprogramming in the Open Service Architecture for Sensors (OSAS), our platform for programming sensor networks.



```
service Budget($budget)
  define
    budget := 0
  function queryBudget() do
    return (budget)
  function useBudget() do
    x := budget;
    if budget then
      budget := budget - 1
    fi;
    return (x)
  on event renew when True do
    budget := $budget
...
```

```
A
(0
SERV 255 0
DEFA 0 3
  CALL[12]
  PUSHC[13]
  EQ)
```

```
B
(0 SERV 2 4
STATE 1
  DEFG 0 0
  ...
DEFE 0 2
  PUSHV[0]
STOREG[0])
```

Fig. 1: *Macroprogramming in OSAS. A (collection of) service(s) is compiled into one or more components, e.g. virtual machine code. These components are then heterogeneously distributed onto nodes by a loader using a deployment algorithm.*

In our work we have developed a Virtual Machine (VM) platform that interprets a bytecode especially designed for sensor nodes and a macroprogramming language that is compiled to this bytecode. Both languages provide the concept of *services*. Such services are loosely coupled components which implement some well defined functionality and are accessible from the network. Each VM in the sensor network must provide a standard set of system services. These system services provide functionality such as service installation, composition and introspection, sensor hardware abstraction, storage and communication (i.e. routing and flooding). From the point of view of an application programmer, such system services are indistinguishable from other (application level) services.

New services can be defined through our macro programming language WaSCoL (the Wasp Service Composition Language). A typical program in this language consists of a series of imports of library services (such as e.g. a service discovery protocol) and the definition of several scenario

[1]The research described in this paper was funded partially from the IST project WASP, IST.034963.

specific services. Services are composed (i.e., interconnected) by *subscriptions* contained in the WasCoL program as well. Deployment directives indicate where and how services should be installed. A distinguishing feature of our design is that newly defined services live on equal footing with the system services and can, when required, provide alternate implementations of system level functionality. For example, a service can reimplement part of the network stack, providing alternate ways of message routing or of code distribution.

A challenging aspect in wireless sensor network is the concept of addressing. Typical applications do not involve communication with targets based on a particular address but determine their destinations or sources based on a property or data value. For example, data may be needed from nodes that are warmer than a certain temperature or nodes that are located in a certain area. Such properties can again be used to route messages or optimize filtering as has been done in specialized routing protocols such as geographic routing[15], directed diffusion[14] and Content-based routing[9]. In this paper we study how this data and node dependent addressing can be specified as part of the macroprogram. With others [8] we call such addresses, *content-based addresses* (CBA) though our CBAs are more general. We study these CBAs from the perspective of address specification and examine how the program may extend it with directives to admit efficient implementations. We regard CBAs primarily as a means to specify a set of destinations from the perspective of a sending node. This set is determined partly by properties ("nodes that are too hot") and partly by forwarding behavior ("nodes that are nearby"). We study as an application the dissemination of code. We design a more efficient means of doing this dissemination using CBAs than the standard dissemination installed in our system.

This paper is further structured as follows. In the next section, related work is discussed. Section 3 presents the addressing scheme of the WaSCoL language and the underlying code dissimination procedure. Using the mechanisms of WaSCoL we construct more efficient code dissemination for our application domain. Section 4 evaluates our results. Finally, section 5 concludes the paper and discusses future work.

2. State of the art

In [8], Carzaniga et al. present a model for content based addressing. They envisioned a model of networked nodes with receiver predicates where each packet is to be received by a node whenever the predicate applied to the packet holds. In [9], [10] they elaborate on how such packets can be routed in a network by utilizing a broadcast routing protocol with a content-based routing protocol on top which prunes the broadcast distribution paths.

The authors of [11] implement a CBA mechanism for mobile networks with multiple sinks. They handle forwarding by maintaining a distance and content table (a table of the sinks interests). For the actual forwarding a probabilistic, counter-based approach is used. A similar approach is utilized for broadcasting periodic beacons to update these tables.

A major application of CBAs within the OSAS system is code distribution. Many algorithms have been proposed for code distribution in wireless and ad hoc networks. In [4] the problem of a broadcast storm is discussed and analyzed. The authors consider redundant broadcasts, contention and collision and propose five mechanisms to limit rebroadcasting (probabilistics, counter-based, distance-based, location-based and clustering).

The Trickle [12] and Deluge [13] algorithms provide rapid and scalable code propagation with low maintenance overhead. A major shortcoming of these protocols which makes them unsuitable for service oriented sensor networks is the assumption of a single uniform code image to be distributed within an entire sensor network. With heterogeneous services one can no longer assume that the neighbourhood of a node can provide the services that it requires.

In contrast to CBAs in the literature, where each node has its own predicate expressing its specific interest in data, we specify a set of predicates on all nodes, each of which identifies a (possibly dynamically changing) group of destinations. Whenever a message is transmitted, this message explicitly states which address (i.e. group) it targets. Since code distribution is a major application in our approach, the predicates themselves tend to focus more on attributes of the nodes rather than the values in a packet, although this is supported as well.

3. Content-based addresses

In a content-based addressing scheme, the destination of a message is specified by means of properties or data values instead of by a unique identifier. This decreases the coupling between sender and receiver and can be beneficial in situations where the endpoints are not known a priori, are costly to discover, frequently fail and join or are irrelevant, which are all very relevant properties for sensor networks. If nothing is further known about a content-based address, sending messages to a CBA reduces to a flooding operation, as only endpoints can evaluate the CBA, in principle. In order to improve on this the handling of the message must be included in the addressing, either by installing special purpose handling or by giving directives to the flooding using application knowledge. For example, in the case of an interest in exceeding a temperature threshold, handlers may be installed that filter temperature messages in a tree-like fashion. In this paper we focus on adding application knowledge though our language and system are powerful enough to also install special handlers, e.g., to do specialized routing for a particular class of addresses. By leveraging application knowledge, CBAs can be utilized to transform flooding into an efficient routing or multicasting mechanism with very low overhead.

In essence, a content based address (CBA) is a guard which determines if a message is to be processed by its receiver. Any message broadcasted (Fig. 2-a) over a single hop is processed by all those neighbours where the CBA evaluates to *True* (Fig. 2-b). In order to reach more than just a single-hop neighbourhood the message containing the address needs an underlying forwarding protocol such as e.g. a flooding mechanism (Fig. 2-c) to propagate it to its intended destinations (Fig. 2-d). Although a flood is often a basic primitive and otherwise easily implemented, it is overkill in most situations and suffers from poor scalability. The main challenge for efficient message propagation is therefore to define a proper forwarder *Fwd* (Fig. 2-e). The optimal forwarder involves the minimum number of nodes that do not satisfy the CBA such that it can still deliver the message to all intended recipients (those satisfying the CBA). This can be generalized to other metrics as well such as e.g. minimizing the total amount of energy spent. The metrics to optimize depend on the scenario realized.

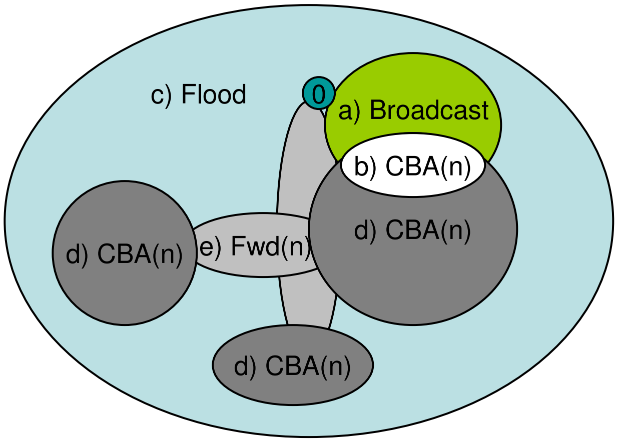


Fig. 2: *Propagation of content based addresses. In order to deliver a message to each node n, such that CBA(n) holds (sets b and d), a forwarding mechanism is required (sets a, c, e) which involves as few nodes as possible while still delivering to all destinations. We seek constructs for expressing forwarders.*

Though such algorithms have been proposed in the literature [10], there is never a single algorithm which is optimal in every imaginable scenario. In this paper we present content based addresses as a programming construct to specify such forwarders or more precisely to tailor an underlying forwarder towards a specific scenario.

In the following part we describe the content based addresses themselves and illustrate how they can be utilized to specialize a simple flooding protocol to improve its forwarding efficiency.

In the OSAS system a content based address consists of three parts:

$$[forwarding\ scope|hopcount|address\ predicate] \quad (1)$$

The *address predicate* is what is referred to as a content based address in the literature where it is usually a predicate over the payload of a message. Within the OSAS system the foremost application of CBAs is to specify service deployment and interconnection. The use of content based addresses in communication eliminates the need for an explicit service discovery, matching and selection operation as commonly found in service oriented systems since these occur during address resolution instead. The merger of these operations is possible since the macroprogramming nature of the OSAS system guarantees that all services are known at compilation time and can be mapped onto unique and compact IDs without the need for large service descriptions which ensures that services can be referenced in addresses at very low cost. As a consequence of this, OSAS CBAs tend to focus on predicates over nodes rather than message payloads, although both are supported.

The state of a node comprises the installed services, their state and operations as well as node properties (key, value) pairs which can be set (remotely) on the nodes. Address predicates are evaluated on the same virtual machine as the services and as such inherit the full expressivity of expressions in the WaSCoL programming language ranging from simple comparisons to invocation of service specific functions and all the code and computation they encompass. Any address that is either undefined, calls an undefined function or fails in any other way will evaluate to *False*.

In practice, a CBA often refers to specific capabilities of a sensor node. For example, $[Network| * |HasSysCall(Temp)]$ specifies all nodes in the entire sensor network which have a temperature sensor (or, more precisely, have the correspondent system call).

The *forwarding scope* and *hopcount* are modifiers which specify forwarding limitations on the message and thereby prune the set of nodes reached by the underlying forwarder.

Like the *address predicate*, the *forwarding scope* is a predicate *considered and evaluated in a particular context*, where context includes the current message location, time, values of variables, node state and message content. The scope predicate limits the nodes participating in forwarding the message to only those nodes which satisfy the predicate. For example, the scope $Forward(True)$ indicates there is no scope restriction: all nodes in the network may forward the messages. We abbreviate this scope as *Network* as shown in the CBA example above. Other examples of scoping are discussed below.

The *hopcount* defines an upper limit on how far the message is allowed to travel. E.g. the CBA $[Network|2|True]$ limits the packet to the two-hop neighbourhood of the sender. An unlimited amount of hops is specified by either "$*$" or "0".

In dense or large sensor networks using a full network flood to propagate messages towards their destination will waste a lot of energy and creates the risk of causing flood storms [4] which increase collisions and unneeded packet transmissions. The following examples demonstrate how the *forwarding scope* can be utilized to improve this situation.

$$[Forward(NodeType() == "StaticNode")| * |HasSysCall(Temp)] \tag{2}$$

$$[Forward(NodeType() == "NeckNode")| * |HasSysCall(AccelX)] \tag{3}$$

$$[Forward(self.rssi < 230)| * |Temp() < 30] \tag{4}$$

Addresses 2 and 3 are taken from a scenario for a testbed setup we deployed in collaboration with the life science university of Wageningen, the Netherlands. In this scenario cows were spread out around a barn where each cow had two nodes with accelerometers on a front and rear leg respectively and one node attached to their neck which had its type property set to "NeckNode". Furthermore, static sensor nodes were deployed throughout the barn providing an infrastructure to reach the sink node which was attached to a pc.

Messages targeted at address 2 are transmitted to all nodes with a temperature sensor, but the forwarding is only performed by nodes tagged as being static infrastructure, which are much easier to recollect and/or replace their battery. Similarly, address 3 was used to transmit messages to all nodes on the cows legs, but the flood is only forwarded by the neck nodes (the leg nodes lose connectivity when the cows lie down on them). This is a lightweight form of assigning the neck nodes as a clusterhead of a mobile body area sensor network.

Besides the initial (over-the-air) uploading of the address definitions and the address evaluation itself there is no periodic overhead for refreshing or rebuilding any paths or tables for forwarding. This also holds when much more dynamic addresses are used such as e.g. address 4 where both the set of forwarding nodes as well as the set of destination nodes are dynamic. In this particular case the message is targeted at nodes below a critical temperature threshold and the received signal strength indicator (RSSI) is utilized to trim the amount of forwarders.

Address 3 demonstrated a simple way to define clusterheads, sometimes the need arises to logically group a subset of the network such as e.g. all nodes within a single body area network. The cluster scope $[Cluster(P)|{*}|Q]$ is a shorthand notation for $[Forward(P)| * |P \;\&\&\; Q]$ just like how $Network$ is a shorthand notation for $Forward(True)$. This scope is utilized to confine all communication to only those nodes which are part of the cluster. For example,

$$[Cluster(PatientID() == \ll PatientID() \gg)| * |True] \tag{5}$$

addresses a cluster formed by all (connected) nodes with the same patientID (the sub-expression $\ll PatientID() \gg$ is evaluated sender side and passed as parameter to the address). Hence, this address refers to all (connected) nodes that have the same *PatientID* as the sender.

As an illustration of the expressive power of these content based addresses, imagine a hospital with static nodes deployed throughout. Each of the nodes is configured with a location property which globally indicates their deployment site (e.g. "Wing1"). Due to some viral infection in wing one, medical staff wants these body hubs to start monitoring the temperature of all patients in this wing. In terms of WaSCoL this means that body hubs must subscribe to temperature services of the body sensors.

The OSAS runtime provides a third-party subscription which directs a receiver of a subscription request to subscribe to a service. This mechanism can perform many-to-many subscriptions from one content based address to another. A simple flooding implementation would perform a two-stage flooding of the entire network. First, a subscription request would be flooded to all potential subscribers (body hubs in wing one). Second, each subscriber would flood his subscription to all providers, which results in a network flood per patient!

The program fragment below shows how a programmer can deploy such a subscription in a scenario-specific fashion.

```
for [Forward(NodeType()=="StaticNode"                                1
      && GetProperty("location","metadata")=="Wing1")
     |*|NodeType()=="BodyHub"]                                       3
install TemperatureSubscription
on [Cluster(PatientID()==<<PatientID()>>)                           5
    |*|HasService(TemperatureService)]
```

Operationally, the meaning of this fragment is: “static nodes in Wing1 forward subscription requests to body hubs; these bodyhubs subscribe to temperature services in their own cluster and these subscriptions are only forwarded within their own cluster”. This replaces a full network flood by a directed regional flood. Subsequently, all body hubs broadcast their subscriptions, but these broadcasts are constrained to the body area network of the individual patients (line 5).

3.1 Realization of CBAs

At the syntactical level, CBAs look powerful but difficult to implement. In practice they can be realized in a concise and efficient manner which means it is a powerful programming primitive. In this section we take a closer look at how CBAs are handled by the OSAS system.

In our programming model each node has a virtual machine for the execution of services, CBAs reuse this virtual machine for evaluation of the predicates. Each occurrence of a CBA in a WaSCoL program is split into two separate parts by the compiler: an address definition and an address reference.

The address definitions consist of the virtual machine code of both the forwarding predicates and the address predicates. They are loaded onto sensor nodes prior to loading any services and/or subscriptions. Each address is assigned a unique ID (one byte) by which it can be referenced. To upload an address the message handler *DEFA* (define address) is introduced which stores an array of bytecode in a lookup table under specified address ID. The following message shows the format of the *DEFA* handler.

DEFA	id,n	$b_0, \ldots, b_n$

where $b_0, \ldots, b_n$ is the bytecode for the address.

In order to check validity of an address the handler *CBA* is introduced. This handler needs to be inserted in transmitted messages along with the reference of the address to test and zero or more parameters to the address (data values such as e.g. $\ll PatientID() \gg$ in the previously mentioned hospital example). The following message shows how the *CBA* handler is integrated into the message format

HDR	FLD	src,seq	CBA	id,n	$p_1 \ldots p_n$	⟨PL⟩	CRC

The OSAS header *HDR* contains, amongst others, the hopcount (maximum number of hops the message is allowed to travel). The payload ⟨*PL*⟩ consists of a series of handlers to be executed (active messages [7]). The *FLOOD* handler (*FLD*) forwards the message throughout the network if the hopcount allows it. The *CBA* handler verifies if address *id* holds and if so processes the embedded payload ⟨*PL*⟩.

The pseudo code of CBA is given by the following fragment

```
if executebytecode(address_table[addrId], message) then
  ExecuteHandler(message, indexOf(<PL>), indexOf(CRC))
fi
```

Whenever a forward predicate evaluates to *False*, the hopcount field in the OSAS header is cleared indicating that the message should no longer be forwarded.[2] In order for this to have the intended behaviour, any forwarder should satisfy the following criteria. 1) It should evaluate its embedded payload before forwarding the message. 2) It forwards the packet only if the hopcount indicates that at least one more hop is allowed *after* evaluation of the embedded payload.

3.2 Building routing protocols

In this section we present an example of how CBAs can be utilized to simply implement more complex routing mechanisms. The focus is on the ease of expressing behaviour and not so much the optimality of the solution itself. The goal is to construct a location dependent routing mechanism, an example taken from the aforementioned testbed setup in Wageningen, the Netherlands. In this particular case, cows equipped with sensors could be either grazing outside in the field or be inside in the barn (most time is spent inside). Outside no static infrastructure was provided aside from one static node at the edge of the field, so multi-hop routing through mobile nodes needs to be employed. Inside, mobile nodes can transmit their data to static infrastructure over a single hop. Because cows are mobile the topology changes frequently when outside, this requires the routes outside to be computed more often. Periodically, routes need to be constructed towards a sink (which is part of the static infrastructure). The wanted behaviour is to periodically construct a spanning tree towards a sink (which is part of the static infrastructure) such that inside the barn mobile nodes will only occur as leaves of the spanning tree while they can also be intermediate nodes when outside.

The first step is to express a service for constructing a spanning tree.

```
service SpanningTree()
  define
    parent := 0    # NodeID of parent, 0 == no parent

  function SendToRoot(payload) do
    # API for sending a payload to the sink.
    Send(parent, Header() ++ Message(Forward) ++ payload)

  action SetParent(sender) do
    parent := sender;
    sender := NodeID()

  action Forward() do
    if parent != 0 then # intermediate node, forward
      Send(parent, self.message)
    else # root node, i.e. sink
        # process the next handler in the payload
      InvokeHandler(self.message, self.handler_start+1,
                    self.handler_end)
    fi
```

This is a generic service for constructing a spanning tree, along with a function which allows delivery of content towards the root of the network. The SetParent message modifies the received message in-place (*sender := NodeID()*) to insert its own ID into the message payload. By flooding a SetParent message, a (non-minimal) spanning tree is created with the initial sender as root. This is because the flooding mechanism prevents double execution of received messages (no cycles) and because it executes its payload before forwarding (one of the requirements of forwarders) such that the sender variable (i.e. the new parent node) is updated as the message traverses the network.

The next step to express the required behaviour is to define the route construction for both inside and outside the barn. For localization we periodically broadcast beacons from the static infrastructure. If these beacons are missed several times in a row, a mobile node is outside, otherwise it is inside. How often a beacon can be missed is a parameter which depends on scenario specific factors such as length of a period, walking speed of the animals, density of the beacons and amount of interference.[3] The static infrastructure itself is always considered to be inside. For brevity we do

[2] This behaviour is contained within the virtual machine code of the address that is generated by the compiler. The CBA handler itself only evaluates the code and verifies the result to determine if the embedded payload needs to be processed.

[3] During most measurements in our testbed setup, we experienced about 30 percent package loss. Requiring a beacon to be lost two to three times in a row ensured that less than 2 precent of the locations were reported incorrect and in all but a few cases this false localization lasted only one period.

not show this service, but its complexity is similar to the tree construction. This localization service exposes a function *getLocation* which is used in the services below.

```
service InsideRoutes()
  on event reconstruct when True do # flood message:              2
    # construct spanning tree w/ mobile nodes as leaves
    SendMessage([Forward(NodeType()=="StaticNode")|*],             4
                SetParent, NodeID())
                                                                   6
service OutsideRoutes()
  on event reconstruct when True do # flood message:              8
    SendMessage(
      # address to prevent it from being propagated inside:       10
      [Cluster(getLocation()=="outside")
      |*|NodeType()=="MobileNode"]                                 12
      # payload, construct a spanning tree:
      SetParent, NodeID())                                         14

# Activate inside route reconstruction, every 15 minutes          16
subscription ReconstructInside
to InsideRoutes()                                                  18
with (period=15m, deadline=1m, send="Normal", exec="Normal")
                                                                   20
# Activate outside route reconstruction every 5 minutes
subscription ReconstructOutside                                    22
to OutsideRoutes()
with (period=5m, deadline=1m, send="Normal", exec="Normal")       24

# Deploy InsideRoutes on sink                                      26
# (max 2 hops because we know it's close to the loader)
for [Network|2|NodeType()=="Sink"]                                 28
  install InsideRoutes
                                                                   30
# Deploy OutsideRoutes on static node next to field
# (this was node 25, reachable through static infrasture)         32
for [Forward(NodeType()=="StaticNode")|*|NodeID()==25]
  install OutsideRoutes                                            34

# Deploy subscriptions, in this case it needs exactly             36
# one subscriber, so we pick the loader. This just
# activates the InsideRoutes and OutsideRoutes services.          38
for [Network|2|NodeType()=="LoaderNode"]
  install ReconstructOutside on                                    40
    [Forward(NodeType()=="StaticNode")|*|NodeID()==25]
  install ReconstructInside on                                     42
    [Network|2|NodeType()=="Sink"]
```

The address used inside the InsideRoutes services ensures that mobile nodes only execute SetParent messages and do not forward them. This results in the mobile nodes only occuring as leaves. In a similar way, OutsideRoutes uses an address to ensure that the outside route reconstruction messages are neither forwarded nor executed inside the barn.

The period of the two subscriptions can easily be adjusted, even at runtime, to optimize the refreshing period though that is outside the scope of this paper.

The program above shows how a generic tree construction service can be specialized for scenario specific behaviour. In this particular case, only two helper services are introduced and through careful selection of content based addresses, exactly the required behaviour is established. Furthermore, from line 27 onwards, content based addresses are used to specify service and subscription deployment each with their own forwarder such that code distribution differs per service.

Though we do not show it here, more complex forwarders can be defined in a similar manner as the spanning tree and routing mechanism shown here. As long as the forwarder first executes its payload and subsequently forwards the message if and only if there are hops remaining after evaluation of the payload then it will be compatible with the content based addresses.

4. Evaluation

As mentioned before, the use of clustering and forwarding predicates has been tested in a testbed setup in cooperation with Wageningen University & Research centre (WUR), in the Netherlands. Several examples in this paper have been derived from this testbed. At this location we have deployed static nodes throughout a barn as permanent infrastructure as well as motion sensing nodes onto the legs of cows walking within the barn. Addresses like (2) and (3) have been used with success to restrict forwarding to only static nodes. Aside from the initial definition of addresses, there is no overhead for computation of forwarding tables (unless a custom forwarder is utilized which introduces this overhead, but the default flood does not). Due to this, with properly chosen forwarding predicates, our CBAs scale quite well. In a lab setup we have deployed up to 127 nodes. In the barn we have successfully deployed up to 80 nodes in a highly dynamic environment (mobile cows blocking each others signals) for the purpose of performing step detection and activity monitoring. These numbers were limited by availability of nodes and animals though traffic density starts to play a role as well and as a consequence the amount of messages per node per time unit needs to be adjusted downwards.

Table 1: Size (in bytes) of various content based addresses used throughout this paper.

content based address	size
$[Network\vert 2\vert True]$	0
$[Network\vert 2\vert NodeType() == "Sink"]$	3
$[Network\vert * \vert HasSysCall(Temp)]$	3
$[Forward(NodeType() == "StaticNode")\vert * \vert True]$	9
$[Forward(NodeType() == "StaticNode")$ $\vert * \vert HasSysCall(AccelX)$	11
$[Forward(NodeType() == "StaticNode")$ $\vert * \vert NodeID() == 25]$	12
$[Forward(self.rssi < 230)\vert * \vert Temp() < 30]$	14
$[Forward(NodeType() == "StaticNode"\&\&$ $GetProperty("location", "metadata") == "Wing1"$ $\vert * \vert NodeType() == "BodyHub"]$	20
$[Cluster(PatientID() == \ll PatientID() \gg)$ $\vert * \vert HasService(TemperatureService)]$	12
$[Cluster(getLocation() == "outside")$ $\vert * \vert NodeType() == "MobileNode"]$	15
	99

As discussed in section 3.1 our CBAs can be implemented very efficiently. In particular, the combination of a virtual machine based programming model and communication based on active messages allows the CBAs to be added by the addition of only two message handlers. The handler to evaluate addresses is a total of 202 bytes, the handler to install new addresses is integrated into the configuration handler which totals 1474 bytes, about 10 to 20 of those are specific to addresses. The virtual machine itself is currently 4206 bytes (compiled for a 16-bit MSP 430 microcontroller). Since our macro-programming approach assigns a unique ID to each content based address, transmitting a message towards a specific address incurs an overhead of three bytes plus one or two bytes for each parameter passed to the address (usually zero or one). Table 1 shows the size of the bytecode for each of the addresses mentioned in this paper (excluding some very similar ones).

In our current implementation all CBAs are uploaded to nodes prior to uploading of services and subscriptions. So far, none of the programs which we have deployed in practice have required a large number of extremely complex CBAs which kept the memory overhead for storing them around 100 bytes and often much smaller. It is not unimaginable that in the future both the number and the complexity (and therefore also the size) of these addresses themselves increase such that it becomes unfeasible or unwanted to statically store them all on every node.

Memory space can be saved by including the definition of an address in a message instead of a reference. Such an approach is only viable for rather small addresses (i.e. around 5 bytes) as otherwise the overhead on the payload becomes too large. This way, only the sender needs to store the address. This is viable in case of scenarios with many, small addresses.

Finally, addresses can be stored more compact. An *address predicate* which always yields the same value for a given node only needs to be evaluated once and the result can be stored instead. Our compiler needs to be modified to recognize such addresses. Furthermore, if an address is known (by the programmer) to be $False$ on a specific set of nodes then the address doesn't need to be stored there as any unknown address is understood to be $False$. Conditional definition of CBAs can be achieved by utilizing non-conditional CBAs as a guard for the installation of the conditional ones.

5. Conclusion and future work

In this paper we have extended content-based addressing (CBA) with predicates over nodes and forwarding restrictions in the form of forwarding predicates and an upper limit on travelling range. Through these extensions, CBAs become a powerful and expressive programming construct for implementing simple, compact and lightweight routing and multicasting mechanisms.

Content based addressing schemes lend themselves particularly well to service oriented sensor networks by merging service discovery and service utilization into one operation. We have shown how to utilize CBAs for code distribution and deployment of heterogeneous services in sensor networks. The advantage of this approach lies in the extent of control which an application programmer has over the code distribution mechanism. The nature of the OSAS programming language allows programmers to consisely specify a forwarding policy on the granularity of individual services and subscriptions.

Finally, we have demonstrated how the underlying forwarding/broadcasting mechanism can be refined through definition of OSAS services which makes it easy to incorporate application level knowledge into their decision process.

Current ongoing work focuses on enhancing our routing protocols for mobile networks. Future work includes generalization of the hopcount range restriction to a per hop message modifier such that a content based address is defined by a *forwarding predicate*, *message modifier* and *address predicate* instead.

References

[1] D. I. Tapia , R. S. Alonso , et al., Introducing a Distributed Architecture for Heterogeneous Wireless Sensor Networks, in Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, 2009.

[2] M. Kushwaha, I. Amundson, X. Koutsoukos, et al., OASiS: A Programming Framework for Service-Oriented Sensor Networks, COMSWARE, 2007.

[3] R.P.Bosman, J.J.Lukkien, P.H.F.M.Verhoeven. An integral approach to programming sensor networks, CCNC, 2009.

[4] S.-Y.Ni, Y.-C.Tseng, Y.-S.Chen, J.-P.Sheu, The broadcast storm problem in a mobile ad hoc network, MobiCom, 1999.

[5] S.R. Madden, M.J. Franklin, J.M.Hellerstein, W. Hing, TinyDB: an acquisitional query processing system for sensor networks, ACM Transactions on Database Systems, Vol.30, Issue 1, March 2005.

[6] H. Liu, T. Roeder, K. Walsh, et al., Design and Implementation of a Single System Image Operating System for Ad Hoc Networks, Mobisys, 2005.

[7] T. von Eicken, D. Culler, S. Goldstein, K. Schauser, Active messages: a mechanism for integrated communication and computation, Proc. of the 19th annual international symposium on Computer architecture, 256-266, May 1992

[8] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf. Content-based addressing and routing: A general model and its application. Technical Report CU-CS-902-00, Department of Computer Science, University of Colorado, Jan. 2000.

[9] A. Carzaniga, M. J. Rutherford, A. L. Wolf, A routing Scheme for Content-Based Networking, Infocom 2004.

[10] A. Carzaniga and A.L. Wolf, Forwarding in a Content-Based Network, ACM SIGCOMM 2003. p. 163-174. Karlsruhe, Germany. August, 2003.

[11] G. Cugola, M. Migliavacca, A Context and Content-Based Routing Protocol for Mobile Sensor Networks, EWSN 2009.

[12] P. Levis, N. Patel, D. Culler, S. Shenker, Trickle: a self-regulating algorithm for code propagation and maintenance in wireless sensor networks, NSDI 2004.

[13] J. W. Hui, D. Culler, The dynamic behavior of a data dissemination protocol for network programming at scale, SenSys 2004.

[14] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heideman, F. Silva, Directed diffusion for wireless sensor networking, Trans. on Netw. 11(1) (Feb 2003) 2-16

[15] A. Rao, S. Ratnasamy, C. Papadimitriou, S. Shenker, I. Stoica, Geographic routing without location information, In Proc. of ACM MOBICOM.

[16] F. Cao, J. P. Singh, Efficient event routing in content-based publish-subscribe service networks, Infocom 2004.

Location Determination Systems for WLANs*

Stanley L. Cebula III, Aftab Ahmad, Luay A. Wahsheh, Jonathan M. Graham, Aurelia T. Williams, Cheryl V. Hinds and Sandra J. DeLoatch
{s.l.cebula@spartans.nsu.edu}, {aahmad, law, jmgraham, atwilliams, chinds, sjdeloatch}@nsu.edu
Information Assurance Research, Education, and Development Institute (IA-REDI), College of Science, Engineering and Technology (CSET), Norfolk State University (NSU)
700 Park Avenue, Norfolk State University, Norfolk VA 23504

Keywords: NLOS, Geolocation, WLANs, Information Assurance

Abstract

In this paper we discuss three location determination mechanisms for WLANs based on geolocation models (Global Positioning System, Angle of Arrival, Time-based Models and a Signal Strength based model). We also present a new mechanism that determines the location of a WLAN station within a triangle or quadrangle with probabilistically the strongest vertices. The Signal Strength based model is identified to be the most appropriate location determination system in a local wireless network, because of its proven accuracy. Such an algorithm can be used to locate all users in an infrastructure type of WLAN, including an attacker. Thus, it provides privacy in a wireless local environment that is equivalent to the inherent privacy in wired local environments due to the tools available to detect the location of an attacker on a cabled network.

1. INTRODUCTION

Even though the IEEE 802.11-2007 specifies a robust medium access control (MAC) sublayer, as show in [1], it has not earned the same level of trust from networking community as, for example, IPsec. Another weakness of WiFi is a lack of signal privacy on the physical layer. With wired LANs, one can physically trace each packet's source and destination machine using commonly available tools, and by running cable through designated areas, thus providing strong privacy. However, there is no way to physically trace a packet on a WiFi network to see the source or destination machine. Attackers can remain anonymous when attacking WLANs. There is no way to control or view the WiFi signal according to the 802.11-2007 protocol. Directional antennas restrict the service availability but won't provide signal containment. Signal spilling can occur when a WiFi signal is transmitted further than intended. This makes it possible for attacks to occur outside of the physical building where the WiFi network is located. A tool that would allow system administrators to view each machine connected to a WiFi network on a map adds privacy to the network. Furthermore, such a tool can be used to create a map of the signal strength of the WiFi network. This will allow system administrators to view where the WiFi network is physically located. In order to plot the signal strength of a WiFi network, a channel model is needed to predict the signal strength at any given distance. [2] focused on the development of a custom channel model to use in signal strength mapping software. This paper will focus on a comparison between five geolocation models in order to ascertain the most accurate location determination system for use in our environment.

The experiments for the channel modeling were performed in the Information Assurance Research, Education, and Development Institute (IA-REDI) located on the sixth floor of the Marie V. McDemmond Center for Applied Research (MCAR) at Norfolk State University. This area is a computer lab (approximately twenty feet by sixty-five feet) next to an office, three conference rooms, and one long hallway. Along with many computers, printers, and other machines, the system under-consideration will be outfitted with a sensor grid to aid in location determination system. Figure 1 represents the layout of the assumed environment. The circle represents the access point (AP). The squares are computers, and the rectangle is a printer. The diamonds are sensors that are all connected to the access point.

The remainder of this paper is organized as follows: obstacles that affect signal strength are discussed in Section 2. Existing geolocation models are outlined in Section 3. We conclude the paper in Section 4. Lastly, we discuss our future work in Section 5.

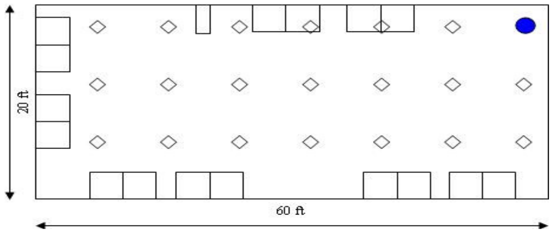


Figure 1. Environment

2. OBSTACLES

In any type of WiFi signal transmission, the outputted signal from the access point will differ from the signal that is received at the client. There are many factors that affect the signal while it is in transit including: attenuation, free space loss, fading, reflection, diffraction, scattering, refraction, and noise. Attenuation occurs when the strength of a signal falls off with distance [3]. Basically, the further the signal travels, the weaker the signal will get. This can be represented logarithmically [1, 2, 3]. Free space loss is a form of attenuation that means the signal disperses with distance [3]. In other words, the further the signal travels, the more the signal spreads out in other directions. The spread of the signal makes the signal weaker. When variation of the signal power occurs due to changes in the transmission medium or path, fading occurs [3]. Basically, any interruption in the transmission medium (atmospheric changes) or path (objects) can affect the strength of the signal. Reflection exists when the signal bounces off large objects causing the signal to change. These changes can increase or decrease the signal strength. This usually happens when the signal reflects off walls, floors, or ceilings. Diffraction is produced when the signal runs into a large object. The secondary waves resulting from the obstructing surface are present throughout the space and behind the large object negatively affecting the strength of the transmitted signal [3]. This can occur when the signal runs into a wall partition or cubicle. Scattering exists when the transmitted signal passes through many small objects that cause the signal to go in many different directions. Scattered waves are produced by rough surfaces, small objects, or by other irregularities in the channel [3]. Refraction is defined as a change in direction of a transmitted signal resulting from changes in velocity [3]. This usually occurs when only part of the line of sight transmitted signal reaches the destination. Noise can be characterized as various distortions imposed by the transmission medium or additional unwanted signals [3]. Noise is usually caused by interference or reception of unwanted signals from other electronic devices. Due to the large number of obstacles that affect the strength of a transmitted WiFi signal, the channel models used to represent the environments must be very specific to each environment. Furthermore, geolocation models must take these obstacles into account in order to be consistently accurate in their calculations.

3. GEOLOCATION MODELS

The geolocation models that are discussed include Global Positioning System (GPS), Angle of Arrival (AOA), Time-based models, Ahmad's Algorithm, and a Signal Strength-based model (SS). We will briefly describe each model and determine if it can be used in our environment. Later, we will pick the most accurate geolocation model for our specific environment.

3.1. GPS

One of the most accurate geolocation systems in use is GPS. As stated in [4, 5], GPS consists of a constellation of twenty-four satellites (synchronized), equally spaced in six orbital planes 20,200 kilometers above the earth. Figure 2 represents the GPS satellite constellation. GPS receivers are used to calculate their exact position (longitude, latitude, and altitude) based on measured signals from at least four (must be able to have line-of-sight to at least four satellites) of these twenty-four satellites. In order to calculate the GPS receiver's location, the GPS receiver compares the time the messages are sent and the satellites' locations. In terms of accuracy, GPS can be exact to around ten meters [4, 6, 7].

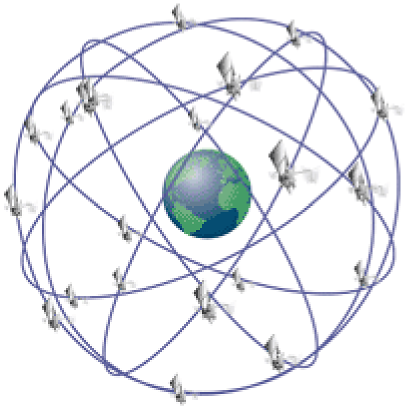


Figure 2. GPS Satellite Constellation [8]

The first shortcoming of implementing GPS concerns calculation time. If the GPS receiver starts without any knowledge of the GPS constellation's state, it may take as long as several minutes for locations to be calculated [4]. Also, as mentioned in [4, 5, 6, 7], in order for GPS to operate properly, the GPS receiver needs to have line-of-sight to at least four GPS satellites. If the GPS receiver cannot connect to at least four satellites, this system will not work at all. Therefore, GPS will not work indoors [4, 6, 7, 9].

Based on the shortcomings of GPS previously listed, GPS is not a system that would work for our location determination system in our environment. First, GPS can take several minutes to calculate locations. Our location determination system will not be able to wait several minutes while locations are being calculated. Our location determination system must be able to calculate locations in a few seconds. Second, GPS will only work with line-of-sight to at least four satellites. Our system will be deployed indoors, so no line-of-sight is possible.

GPS is not a system we will use for our location determination system.

3.2. AOA

AOA geolocation systems use antenna arrays and the angle of the array from the client to multiple base stations to calculate specific locations [5, 6, 10, 11]. Figure 3 represents how the measurements of antenna arrays can be calculated into location (based on [5]). Node C represents a client connected to the sensor network. Nodes S represent sensors, and all nodes have x-y coordinates. The distance between node C and nodes S are represented by (a) and (b). The angle between the antenna arrays and sensor nodes are represented by (c) and (d).

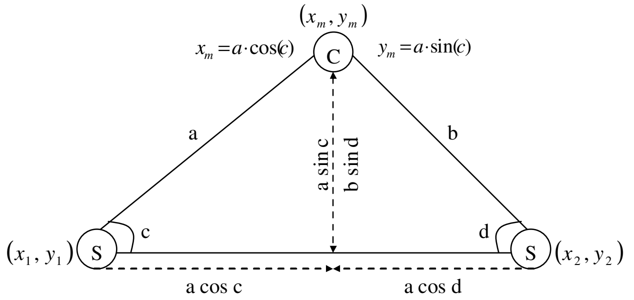


Figure 3. AOA Measurements and Calculations [5]

According to [5, 9, 10, 11, 12], AOA geolocation systems have accuracy issues indoors due to multipath interference. In order for the measurements to be extremely accurate, line-of-sight is required from the client to the sensors. If there is no line-of-sight, measurements will not be accurate.

Due to the fact AOA measurements are not accurate all of the time, AOA is not a system that would work for our location determination system. Our system is located in an indoor environment that is deployed in a multipath channel. The multipath interference will cause AOA measurements to not be accurate all of the time. While it is possible AOA will work some of the time, we need a location determination system that will be accurate on a consistent basis. AOA is not a system we will use for our location determination system.

3.3. Time-based Models

There are two types of time-based geolocation systems; Time of Arrival (TOA) and Time Difference of Arrival (TDOA). As stated in [9], TOA geolocation systems measure distance based on an element of propagation delay between a transmitter and a receiver since in free space or air, radio signals travel at the constant speed of light. In order for the calculations to be exact, the internal clocks of the sensors and client need to be synchronized. There are many equations used to calculate the distance estimates as discussed in [5, 6, 9, 10, 11]. TDOA takes the formulas used in TOA and adds more estimation in order to account for the lack of synchronization between the client and sensor nodes. TDOA still requires the sensors' clocks to be synchronized. In more detail, the TDOA of two signals traveling between the client and two sensors is estimated, which determines the location of the client on a hyperbola, with foci at the two reference nodes (a third sensor is used for localization) [10, 11, 13].

One of the shortcomings for TOA is the requirement for the sensors and client to have synchronized clocks [5, 6, 9, 10, 11, 13]. If the client or sensors are not synchronized, the TOA output will be inaccurate. Even though TDOA does not require synchronization between the client and sensors, it still has accuracy issues due to the estimation of clock delay between the client and sensors [5, 6, 10, 11, 13]. If the estimate for TOA between the client and sensors is not accurate, the location computed in the TDOA calculation will not be accurate. Finally, as with AOA geolocation systems, TOA and TDOA will perform poorly if there is no line-of-sight [10, 11, 12]. Multipath interference will significantly reduce the accuracy of TOA and TDOA geolocation systems.

As reported previously, TOA requires synchronization between sensors and client. While it is acceptable to assume the sensors will be synchronized in our sensor grid, it is not acceptable to assume clients will be synchronized with the sensors. In order to eliminate the need of synchronization, TDOA makes estimates about the difference in specific TOA values. However, if these estimates are not accurate, the results of the TDOA calculation will not be accurate. We would rather employ a geolocation system that does not depend on estimates. Lastly, TOA and TDOA will not be consistently accurate in our environment, because our environment has a multipath channel. Due to synchronization requirements, calculations based on estimates, and inadequate resistance to multipath interference, we will not implement TOA or TDOA geolocation systems for our location determination system.

3.4. Ahmad's Algorithm of Closest Vertices[1]

This algorithm compares the signal strength values for the wireless station received at all of the sensors in the grid. The objective is to determine the quadrant (or triangle) where the client is located. In more detail, a list is compiled of the signal strength of the client at each sensor. Next, the list is sorted from the strongest signal to

[1] Due to Professor Aftab Ahmad of Computer Science Department, Norfolk State University. Also, a co-author of this paper.

the weakest. Finally, the algorithm selects which quadrant the client is in based on four rules:

i. if the four strongest signal strength nodes form one quadrant, the client is located in that quadrant,
ii. if the three strongest signal strength nodes are from one quadrant, the client is located in that quadrant,
iii. if the two strongest signal strength nodes are from one quadrant and they form a vertical line, two neighbors (left and right) of one of the nodes are compared where the strongest signal strength results in the quadrant where the client is located,
iv. if the two strongest signal strength nodes are from one quadrant and they form a horizontal line, two neighbors (above and below) of one of the nodes are compared where the strongest signal strength results in the quadrant where the client is located.

Figure 4 represents a sensor grid with quadrants to further explain Ahmad's Algorithm. The sensors are denoted with labeled squares, and the circle is the client. The dashed lines form a quadrant. The following example refers to Figure 4. As rule one stipulates, if the four strongest signal strengths are for nodes F, G, J, and K, the client is located in quadrant FGJK. As rule two stipulates, if the three strongest signal strengths are for nodes F, G, and K, the client is located in quadrant FGJK. As rule three stipulates, if the two strongest nodes are G and K, the signal strength values for nodes F and H or J and L are compared. If nodes F or J have a stronger signal strength value for the client than H or L, the client is located in quadrant FGJK. Otherwise, the client is located in quadrant GHKL. As rule four stipulates, if the two strongest nodes are G and F, the signal strength values for nodes C and K or B and J are compared. If nodes C or B have a stronger signal strength value for the client than J or K, the client is located in quadrant BCFG. Otherwise, the client is located in quadrant FGJK.

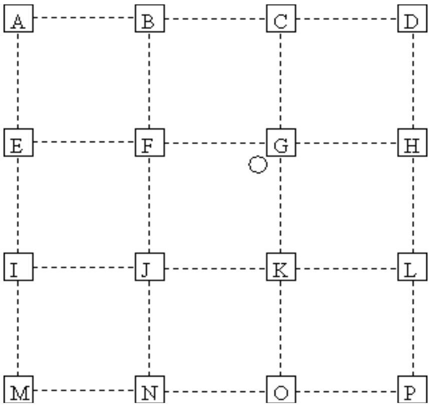


Figure 4. Sensor Grid with Quadrants

Ahmad's Algorithm is simple and convenient for networks with sensor grids, but the efficiency can depend on the size of the quadrants. For example, if the quadrants are very small, mobile clients will be changing quadrants constantly. If the client moves randomly while attacking a network with this design and the quadrants are small, the quadrant where the client is located will change too quickly to provide a reliable location. Also, if the quadrants are very large, it will take more time to search a quadrant for a specific client. Furthermore, if a client were to walk around node G in Figure 4, quadrants BCFG, CDGH, FGJK, and GHKL would need to be searched. There have been no implementations of Ahmad's Algorithm yet to test the ideal size of quadrants.

Even though Ahmad's Algorithm is straightforward and accommodating for networks with sensor grids, there are still many questions about it that needs answering. For example, the ideal quadrant size needs to be determined along with the performance and computational requirements. After these characteristics are known, there will be enough information to determine whether this algorithm is appropriate for use for a location determination system in a WLAN. At this point in time, we will not use Ahmad's Algorithm for our location determination system, because we do not know how it would perform (accuracy and speed) against other geolocation systems.

3.5. SS

SS based geolocation systems do not rely on line-of-sight. SS systems are well suited for indoor use, because they account for multipath interference [6, 10, 11, 13]. The combination of measured signal strength and a path loss model will produce a value that represents the distance between the client and sensor [6, 10, 11, 12, 13]. In more detail, a channel model represents the path loss a signal will experience in a given transmission medium. The signal strength of a client compared to a sensor can be entered into a channel model to produce a distance. If this calculation is completed between one client and three sensors, the client's location can be determined. Figure 5 represents the SS location determination process. The client connected to the network is represented by (C). There are three sensors (Sa), (Sb), and (Sc) that have signal strength values for the client. These signal strength values are plugged into the channel model to get the distance the client is from each sensor (Da), (Db), (Dc). Next, circles are drawn to represent the points where the client could possibly be located. Finally, the intersection of all three circles represents the client's physical location.

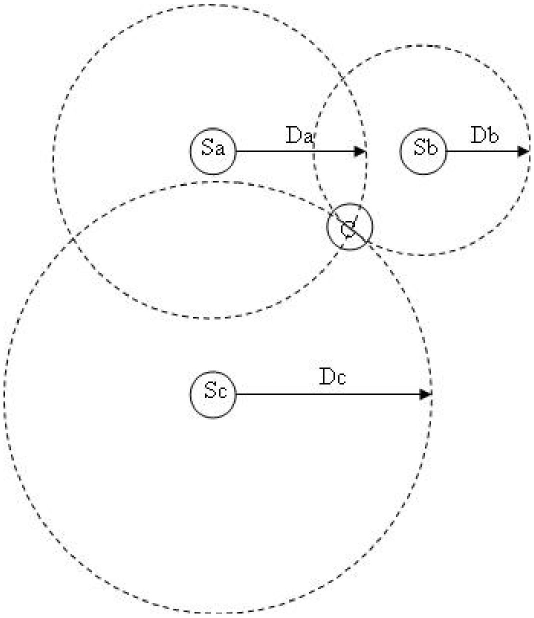


Figure 5. SS Location Determination Process [5]

While SS geolocation systems eliminate the need for line-of-sight, estimation still occurs due to multipath interference of the radio signal. As stated in [10, 11, 13], a channel model is needed to predict the path loss in a given medium. This channel model estimates the relationship between distance and signal strength. Therefore, the accuracy of SS geolocation systems depends on the exactness of the channel model used in the SS calculation process. If the channel model is not accurate, the results of the SS calculation process will not be accurate.

In our environment, we measured the signal strength in various locations throughout a working week 810 times in order to calculate an accurate channel model [2]. Our channel model is accurate, because it is based on actual measurements. If we use our custom channel model in the SS calculation process, the results would be extremely accurate. Also, this would require no additional equipment other than the sensor grid we would already have in place. Until the writing of this paper, the SS geolocation system in combination with our custom channel model is identified to be the best location determination system for a WLAN.

4. CONCLUSION

In this paper, we have compared salient features of several location determination systems to be employed in a Wireless Local Area Network. The location determination system will plot the physical location of clients for an entire WiFi network based on signal strength measurements by a sensor grid and a channel model. The research presented in this paper identified four existing geolocation models and one new geolocation model to consider for implementation in a WLAN environment. In our conclusion, we should not use GPS, because our network is inside an office environment. Also, we cannot afford to wait several minutes for results. AOA and Time-based models are not resistant to the multipath channel where our location determination system will be deployed. Furthermore, the accuracy of TOA is based on synchronization between the client and network grid (we do not always have this), and the accuracy of TDOA is based on estimates that guess the time propagation delay (we do not want to depend on estimates). While Ahmad's Algorithm is uncomplicated and suitable for networks with a sensor grid, further work is underway to verify its performance versus cost tradeoff. The SS model will work well in our environment, because we have a channel model based on measurements specific to our environment. This leads us to state that the combination of the SS based model and our custom channel model is appropriate to use for a location determination system in our environment.

5. FUTURE WORK

In order to compare and contrast the performance of our system, a sensor grid network will need to be deployed. After the network is deployed, we will be able to test Ahmad's Algorithm against the SS based model to determine which geolocation system performs better and is more accurate. Furthermore, once a location determination system is selected, it will need to be coded in the signal strength monitoring system program. The combination of the signal strength monitoring system and location determination system will greatly increase the security of WiFi.

References

[1] Cebula, S. L., Ahmad, A., Wahsheh, L. A., Graham, J. M., DeLoatch, S. J., and Williams, A. T., "How Secure is WiFi MAC Layer in Comparison with IPsec for Classified Environments?". *In Proceedings of the 14th Communications and Networking Simulation Symposium*, April 2011.

[2] Cebula, S. L., Ahmad, A., Graham, J. M., Hinds, C., Wahsheh, L. A., Williams, A. T., and DeLoatch, S. J., "Empirical Channel Model for 2.4GHz IEEE 802.11 WLAN." *In Proceedings of the 2011 International Conference on Wireless Networks*, July 2011.

[3] Tummala, D. "Indoor Propagation Modeling at 2.4GHz for IEEE 802.11 Networks". M.S. Thesis, University of North Texas, 2005.

[4] Djuknic, G. M. and Richton, R. E. "Geolocation and assisted GPS," *Computer*, vol.34, no.2, pp.123-125, Feb 2001.

[5] Sayed, A. H., Tarighat, A., and Khajehnouri, N. "Network-based wireless location: challenges faced in developing techniques for accurate wireless location information," *Signal Processing Magazine, IEEE* , vol.22, no.4, pp. 24- 40, July 2005.

[6] Gustafsson, F. and Gunnarsson, F. "Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements," *Signal Processing Magazine, IEEE* , vol.22, no.4, pp. 41- 53, July 2005.

[7] Xiujun Li, Gang Sun, and Xu Wang. "Mobile Positioning System Based on the Wireless Sensor Network in Buildings," *In Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing, 2009,* pp. 96-100, Sept. 2009.

[8] http://www.declarepeace.org.uk/captain/murder_inc/site/911-7.html

[9] Pahlavan, K., Xinrong Li, and Makela, J. P. "Indoor geolocation science and technology," *Communications Magazine, IEEE* , vol.40, no.2, pp.112-118, Feb 2002.

[10] Gezici, S. "A Survey on Wireless Position Estimation," *Wireless Personal Communications*, vol. 44, no.3, pp.263-282, Feb 2008.

[11] Gezici, S., Zhi Tian, Giannakis, G. B., Kobayashi, H., Molisch, A. F., Poor, H. V., and Sahinoglu, Z. "Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks," *Signal Processing Magazine, IEEE* , vol.22, no.4, pp. 70- 84, July 2005.

[12] Tsung-Nan Lin and Po-Chiang Lin. "Performance comparison of indoor positioning techniques based on location fingerprinting in wireless networks," *2005 International Conference on Wireless Networks, Communications and Mobile Computing,* pp. 1569-1574, June 2005.

[13] Yihong Qi. "Wireless Geolocation in a Non-Line-of-Sight Environment". Ph.D. dissertation, Princeton University, Nov. 2003.

[*] *Acknowledgement:* "This material is based upon work supported by the Department of Energy National Nuclear Security Administration under Award number DE-FG52-09NA29516/A000." *Disclaimer:* "This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof."

Study of Wireless Network and How to Enhance its Performance

Waqqas ur Rehman Butt[1], **Sohail Abbas**[2]
[1]Centre of Excellence in Water Resources Engineering, University of Engineering & Technology, Lahore, Punjab, Pakistan Email: wkbutt@hotmail.com
[2]MIS, SNGPL, Lahore, Punjab, Pakistan

Abstract - *Wireless networking is the most efficient and convenient technology and rapidly increasing from last few years and applying on almost all fields such as home, offices and public places. Currently, this technology is becoming more popular because of its easy installation without cables. This communication has revolutionary impact on data networking. Using security measurements wireless network become reliable and without cord mobility of user is possible. Data transferred in the form of Radio waves in air medium. Wireless network based on network standard defined by IEEE, infrastructure and base station. Wireless network standards and different types are briefly described in this paper. Wireless network is also referred as WI-FI and WLAN (Wireless Local area Network). This paper is showing that how wireless network work, design or merge with existing network, Usages, requirements and techniques for improving the performance of wireless environment. The most optimum techniques and methods are also analyzed.*

Keywords: Wireless Network, WLAN, Wi-Fi, Wireless Infrastructure, Wireless Network security.

1 Introduction

Any type of network in which network devices and computers are connected without wires is called wireless network. Wireless network used to connected these devices or computer without any physical medium (Wire). Telecommunication and Large scale network used this network avoid to spend a lot of money on wires; they prefer to use wireless network for the communication and connection. Initially it was introduced by IEEE. It is also certified by After getting the certification from IEEE 802.11 Technology, it gives more functions (i.e. security and privacy of computers) than previous. Wireless network have no Wired so it use air as medium for data communication. In this network, connection is created with electromagnetic waves (Radio Waves). In other words Wireless network is based on radio waves which are generated from sender and receiver both ends for wireless communication. In other words we can say that data transferring in the form of radio signals. These radio waves works on the physical layer of the OSI model (Open System Interface). Radio waves used to transmit the data. This data can be in the form of picture, music or voice. Radio waves cannot be seen or detect by human which have great impact on society. In current era it is using in different ways like mobile phone, cordless phones etc.

Wireless networking is going more popular due to its easy configuration and setup. Spending less efforts you can create and manage wireless network. Some time it is also referred as WIFI or WLAN (Wireless LAN). It can be implemented on public areas very easily. For example, currently mostly international air ports of the globe giving the free Wifi facility to use internet for valuable customers. People can use this facility on their mobile as well as laptop computers without any configuration. Wireless network is very popular in young generation because they can connect with their buddies from their cell phones by using Wifi, mostly small business places like Coffee shops, Malls giving the free of cost Wifi internet. New form of this technology gives more functions i.e. low cost technology and flexible with high standards. Figure 1 shows that how the workless network works.

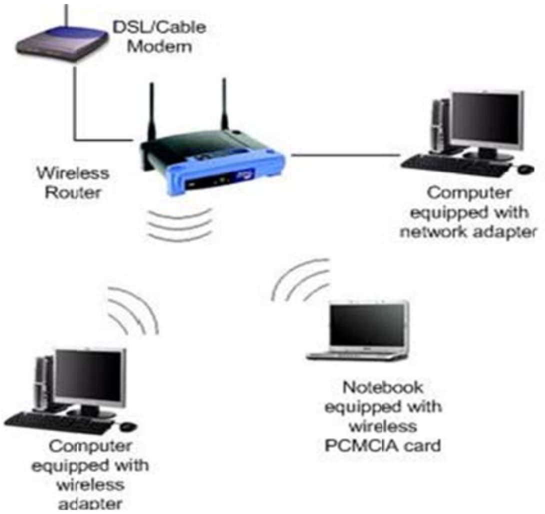

DSL/Cable Modem
Wireless Router
Computer equipped with network adapter
Computer equipped with wireless adapter
Notebook equipped with wireless PCMCIA card

Figure 1: Wireless Network

2 Types of Wireless Networks

Every wireless network type uses the spread spectrum, which gives facility to the end user for mobility within coverage areas of wireless network. There are three types of wireless network WAN, WLAN and PAN. WAN and WLAN are the most popular.

2.1 WAN Wide Area Networks

WAN used in wide area. Internet is an example of WAN. WAN consist of different LANs. These LANs are connected with WAN with the help of routers. WAN is responsible to maintain the addresses of LAN and WAN. Wide Area

Networks support both voice and data services. Mobile phone network use this type. Initially it support only voice but in the enhancement it gives the data service as well. User can use both these services (Data and Voice), where the mobile signals is available. In large area network where is WAN is best solution for your business availably of connection (Signals) is important rather speed have less importance. GSM (Global System for Mobile Communication) and GPRS (General packet radio service) are using in mobile companies for data and voice communication. Both GSM and GPRS are compatible with each other. Different mobile companies offered both. In Pakistan Ufone, Warid and Mobilink are the most popular companies which are providing these GSM and GPRS for their users. EDGE technology is next step of GSM and GPRS. Mobile companies of the Developed countries are also offering this technology for users. (2G, 3G, 3GS and 4G) currently Apple is producing 4G capable mobile phones.

2.2 WLAN (Wireless Local Area Network)

WLAN is the extension on Local area network (LAN) but without wires. Access point AP is used to create connectivity between devices and computers. This technology also offered the secured networking. Network becomes secured with the help of WEP. It is using from last few years in market, especially if you want to develop the wireless network in some limited area of public places i.e. in campus, building, airports, mega shopping malls and hospitals. In these days this technology is becoming more popular to use in widely. WLAN uses the 802.11n which gives the high speed data transferring rate. WLAN gives the mobility for users and easy installation which gives the more reliability of network. One of the big advantages of this technology is Installation cost, which is less as compare to other wire technology. Spread spectrum come from Military effort to become more secure service. There are two kinds of spread spectrum which are frequency and direct sequence. Frequency spectrum is gives single channel and it also maintain this channel, chip terminology is given by direct frequency. Direct spectrum is more useful and also gives the infrared. Wireless adapter is key hardware of wireless network for connection between computers and devices. Currently there are three types of WLANS is offering which are peer to peer, Bridge and WDS (Wireless Distribution System).

2.3 WMAN (Wireless Metropolitan Area Network)

Speedy wireless network within area of town or a metropolitan is called wireless metropolitan area (WMAN). Maximum supported range is fifty kilometer, within this range wireless connectivity is possible and have fast communication. This is useful for a wide scale area comparatively than Local area network (LAN). Maximum speed is 155 Mbs/s within its slandered area which is fifty kilometer. Currently it is using in university campuses and colleges. High network speed is very necessary for better communication for WMAN. IEEE is certified its name WMAN. Air is affected on WMAN because it is also used scheme of single carrier. This is the first step to create a large scale to provide wifi network to users. It is also reliable and has ability for IP Voice and avoids network congestion as well.

3 Wireless Networks Usage

As we already describe that wireless network give the communication between computers and devices without any physical medium (Cable). In the current era telecom is purely relate with wireless networking and its use is becoming more popular and have change the world. Most of the developed as well as developing courtiers are using Wireless network in people and business (small and large scale). Wireless network is based on radio waves so user can do his work while he is out of office. This thing can be done on large scale; users can share the information or have secure communication with their remote offices (All over the world) with the help of satellite signals. It is very useful in remote areas where the telecom is meager. It can be installed in very cheap and quick source for internet connectivity. In campus or university, wireless network is very helpful to share the information and resources very fast. Users can get the information regarding the research material from online library systems in very easy way. It is also very helpful in sharing of hardware for example printer, cameras etc. in these days Wireless enabled devices are also available (printer and Cameras).In these days wireless networks are using for emergency calls, for instance rescue and police department are using this network for quick sharing of information among their employees.

4 Wireless Network Architecture

Wireless network use the network resources and phenomena same like cable network. Difference is in medium, wireless network medium is Air and other is cable. In wireless network information coverts into the signal and transfer these signal in air to the receiver end. In this section we will discuss the main parts of the wireless network and how data conversion and transmission occur during communication. There are many devices are available, those support the wireless communication. These devices use the Radio waves for connectivity. There is some special configuration is required when you are planning to implement the wireless network. The figure shows the main components used in wireless network. Figure 2 showing the basic architecture of wireless network.

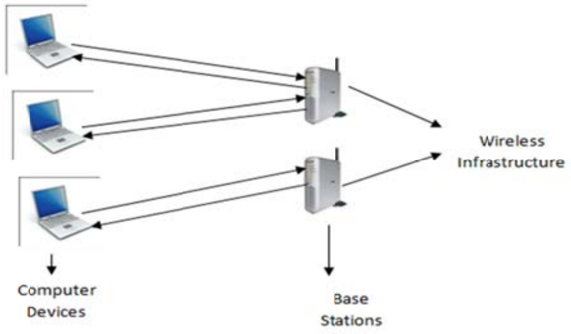


Figure 2: Basic Architecture of Wireless Network

Computer devices can be used by user. User may be a human or reboot. Users have ability to use wireless resources as well as connectivity. Person is user. For instance, being a user if you are travel consultant and your majority work is in travelling. You can access your office network by using public place wireless network. In other words you can start and stop using the wireless network if you are authenticated user of this network. Because in some place wireless network is paid, for example if you want to your wireless network in some airport you have to pay some money card to become authenticated user for the specific time period. After this you cannot use the Wireless network. Wireless network gives the many facilities to the users, so that user can change the places within range of wireless network signals. As earlier discuss that cellular phone is biggest example of wireless network. Latest technology in telecommunication, Cellular user, keep talking during walking and in the meantime phone is checking e-mails.

4.1 Computer Devices

The devices which are attached to a computer are behaved like a client in any network. These devices used for computers and end users. These are also used to create communication between computers within the range of network. Server machines are the part of this network. Cell phone and portable computers are also considered devices. Expert can easily convert running devices into wireless network. For instance, if you have laptop using cable network, after install wireless NIC in your laptop then you can easily use the wireless facility also. Every network device has built compatible drivers to easy use. For example, latest operating system like windows 7 and vista have ability to connect wireless network without any extra application for recognition of network.

4.1.1 Network Interface card

Network interface card (NIC) is basically used to create connectivity between computers and other network devices in any network. For the connectivity in wireless you need wireless NIC. In latest laptop and cellular phone have built Wireless NIC but it is easily available for external use. For example USB, PCMCIA, ISA and PCI wireless NIC are easily available in the market. There are some built-in standards in NIC which give the functions for connectivity. Wireless NIC has antenna. Wireless NIC is used for conversion of Electrical signal into radio waves and vice versa and spread into the wireless network.

4.1.2 Medium

Air is used as medium in wireless communication because radio waves (signal) travel in the air which cannot be seen or heard by human. It assumes these signal have no effect on human ears or eyes. It is the most important or it can be state that is the back bone of wireless network, because information are exchanging network devices through air. It behave same as when two person talking with each other. In this case if one person is far then other should speak loudly for clear hearing. This phenomena is used in wireless network if computer device is near to wireless access point then it have strong strength of signal and have good response. Any obstacle in the air can create problem like poor signal of wireless. So that smoke and snow have effect on wireless signal quality.

4.2 Wireless Network Infrastructures

Wireless infrastructure provides the connectivity between end-user with network devices and systems. There are four parts of the network infrastructure which are as follows:

4.2.1 Base Stations

It is an interface between wire network (Distribution System) and wireless network signals. Wireless NIC card is also installed in Base Station, and having same methodology which are installed on user end. It provides some services also i.e. Electronic Mail and Web services. It has different function as per requirement, for Example one Access point (AP) or more than one is also consider as a base station of Wireless network. For instance, in a building wireless network, user can connect with wireless network of building with nearest access point. Whenever a user is moving in building, its Wireless NIC connect with nearest access points so that user can remain connect with wireless network. Currently, there is a lot of development has been made in this technology so that Router and Gateways are the recent form of Base Station, which give the more functions. Router connect the any single wire (DSL or Local Area Network) and spread the signals and within range of router multiple user can start communication.

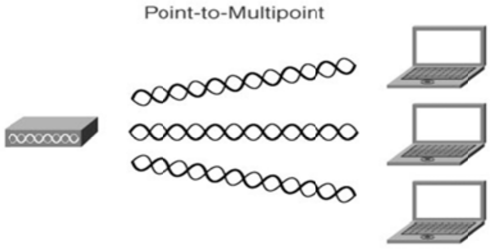


Figure 3: A base station might support point-to-point and Multi point Communication

4.2.2 Access Controllers

Initially when Wireless technology is developed, there was no security measurements defined. But after sometimes manufacturers defined some solution regarding security which is called Access Controller. It is a hardware device. It is used between the network (Wired) and access points. It works before the APs in the center to provide the secured communication between users and resources. In these days many access points also provide the access controlling functions as well. It can be implemented on many application. Network speed can be optimized because only authenticated users can use the wireless network. it is low cost, easy and open connectivity and centralize support are the major benefits of access controllers. It has many features to secure the network which are user authentication, Data Encryption and subnet roaming. It also used in bandwidth management.

4.2.3 Application Connectivity Software

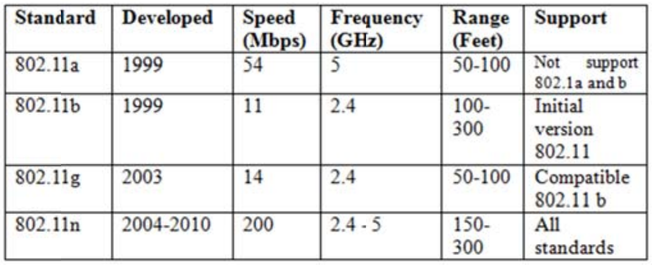
Internet browsing and electronic mail are the main functions which are performed very smoothly on wireless networks. Users need only application software to use these functions such as internet browser (Internet Explorer or Mozila) and e-mail software (MS outlook). Some time there is some special software application required to create communication between computer and main system (Server). It is acted as an interface. Data base connectivity (Data base server) and Wireless middle ware are the popular types of application connectivity software. Wireless middle ware software is used to connect user to the main server where application software reside. This is install on the specific computer (centrally), this is connected with existing cable network. Middle ware wireless network provides the reliable communication between user and servers.

4.2.4 Distributed and Management System

Wireless network systems have no wire for communication. On the other side if you have cable network as well as wireless network then it is possible within distribution system. Ethernet network mostly used the distributive technology. 10 Base T and 100 Base T is commonly used in Ethernet and wireless distributive system. Fiber Optic is also used, but its more expensive than twisted pair but it give very high speed of data transmission during communication. Network management is always required whether you have wired network or wireless network. Network management provides the reliable and smooth communication. People and management application software are the main part of network management. It also gives the secure network (Avoid unauthenticated user), configuration of devices, regular monitoring of network and generated the reports in the form of log files for network administrator. Management system is very helpful during solving the problem in network.

5 Methodology

Data transmission in wireless network in the form of waves (Radio). Mobile technology and TV based on this technique. Communication uses 2 way methodology. In this section we will design a wireless network. Wireless NIC converts the data (1 and 0) into waves and vice versa. Transmission rate of data frequency can be different, most probably it can be 2.4 to 5 GHZ. It is very high rate because others (Mobile, TV) work at very low rate. IEEE is defined few standards which are 802.11a, 802.11b, 802.11g and 802.11n. 802.11g is the higher transfer rate as compare to 802.11b. data transmission in 802.11is higher 5GHz and 54 Mb/s and less inference. 802.11g works at 2.4 GHZ it also support data rate of 802.11a. The latest standard is 802.11n and commonly used in wireless network. it can support 140 Mb/s if network have no congestion errors. It can be state that data transmission based on frequency rate (High Frequency allows more data). Table shows the standard of wireless network standards.



Standard	Developed	Speed (Mbps)	Frequency (GHz)	Range (Feet)	Support
802.11a	1999	54	5	50-100	Not support 802.1a and b
802.11b	1999	11	2.4	100-300	Initial version 802.11
802.11g	2003	14	2.4	50-100	Compatible 802.11 b
802.11n	2004-2010	200	2.4 - 5	150-300	All standards

Table 1: Wireless Network standards

For example, if you have running a wired network and you want to change or develop wireless network then you need few devices such as wireless NIC in computers, Wireless router having Ethernet Port and access points and repeater for wide range. Wireless NIC is used to conversion of data into waves and send to wireless router. Router receive the data and send to receiver end. On the other hand if router is connected with internet cable it receive the data from internet, convert into wave form and send to user. If a user wants to share any information in wireless network, this information comes to Wireless NIC first. NIC converts into wave (Radio) and send to receiver (Router or Computer) in wireless network. At the receiver side (Computer or Router) receive this wave and again translate into the same format of sender. Suppose if router is completed the receiving information and translation then this information is available for all users those connected to this router.

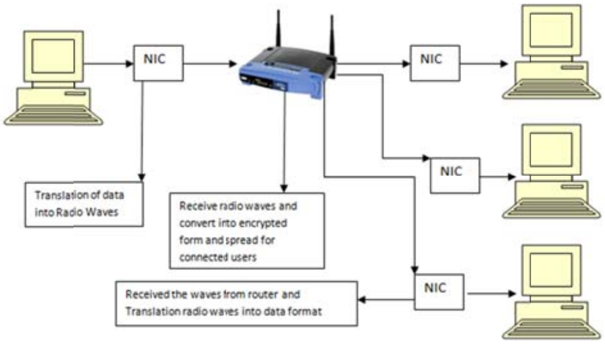

NIC
NIC
NIC
NIC
Translation of data into Radio Waves
Receive radio waves and convert into encrypted form and spread for connected users
Received the waves from router and Translation radio waves into data format

Figure 4: sharing information in Wireless network

6 Wireless Network Security

There are following steps which can helpful to make your wireless network secure and reliable. It helps you protect you PCs and other network devices.

Encryption

One of the best way to secure your communication in the wireless network is Encryption. In these days most of available wireless routers and APs have built in this methodology. By default whenever you install new Wireless Router this feature is turned off, it is recommended to turn on because it is very helpful to avoid unauthorized user to use network resources, which give the better performance. WPA (Wireless Protected Access) and WEP (Wired Equivalent Privacy) are types of encryption. WPA has more features and strong as compare to WEP.

Change the Identifier and password

Every Wireless Router has by default ID and administrator username/password from the manufacturers. It should be change before installation. Hackers can easy access the router and change configuration of router very easily.

MAC Filtering

It is the most secured methodology to secure your network. it is little difficult to configure than WEP and WPA. Hardware address is the unique for network devices. You can allow or deny the devices on basis of MAC address.

7 Conclusion

Main objective of this paper is defining the techniques for improving the performance and quality of Wireless network. Wireless networking is widely acceptable in all over the world. It has great benefits such as low cost, easy installation, reliable, secured and without cabling. This technology gives the user's mobility. It is very successfully using in public places, Organizations, institutes, Emergency department and airports. Paper focused on the working and user aspects in the wireless network. Aiming to improve the Wireless network performance, we present different techniques and precautionary measurements; by using these we can improve the performance and reliability of data in Wireless network systems. Encryption, MAC filtering and Change the Identifier and password are used for secured Wireless Network. Placement of Wireless devices (Wireless Routers/Access points, Repeater and adopter) also discussed and these should be placed on right location to get better Wireless network coverage and signal strength.

References

1 L. Bononi, M. Conti, and E. Gregori, "Runtime Optimization of IEEE 802.11 Wireless LANs Performance," IEEE Transactions on Parallel and Distributed Systems, vol. 15, January 2004.

2 L. Bononi, M. Conti, and L. Donatiello, "Design and performance evaluation of a distributed contention control (DCC) mechanism for IEEE 802.11 wireless local area networks," in Proceedings of First ACM International Workshop on Wireless Mobile Multimedia, Oct. 1998, pp. 59-67.

3 "Wireless Ethernet LAN (WLAN)", General 802.11a/802.11b/802.11g (FAQ)

4 10 tips to help improve your wireless
www.microsoft.com/athome/setup/**wirelesstips**.aspx

5 X.Y. Li, P.-J. Wan, O. Frieder, Coverage in wireless ad hoc sensor networks, IEEE Transactions on Computers 52 (6) (2003) 753–763

6 S. Meguerdichian, F. Loushanfar, G. Qu, M. Potkonjak, Exposure in wireless ad-hoc sensor networks, in: Proceedings of the ACM International Conference on Mobile Computing and Networking (MOBICOM_01), July 2001, pp. 139–150.

7 J. Lee, C. Chuang, C. Shen, Applications of short-range wireless technologies to industrial automation: a ZigBee approach, Venice, Italy, in: Fifth Advanced International Conference On Telecommunications, May 2009, pp. 15–20.

8 A. Zainaldin, I. Lambadaris, B. Nandy, Video over wireless ZigBee networks: multi-channel multi-radio approach, in: International Wireless Communications and Mobile Computing Conference, Crete Island, Greece, February 2008, pp. 37–43.

Experiences Managing a Parallel Mobile Ad-hoc Network Emulation Framework

Travis W. Parker
Contractor, ICF Jacob & Sundstrom, Inc.
U.S. Army Research Laboratory
travis.w.parker16.ctr@mail.mil

Abstract

Modeling a large mobile ad-hoc network is well suited to a cluster computing environment due to its parallel nature. Following the design of high-performance computing systems, uniform and scalable clusters optimized for wireless network emulation can be built using inexpensive commodity hardware. In this paper, we describe our experiences with managing a parallel mobile ad-hoc network emulation framework. We will discuss design differences in small and medium-to-large scale deployments, the use of cluster interconnects and bonded-channels with a network emulation framework, and the application of cluster management and job control solutions to clusters designed for network emulation frameworks.

1 Introduction

Large-scale wired or wireless network scenarios may contain hundreds to thousands of nodes. The Extendible Mobile Ad-Hoc Network Emulator (EMANE) provides a framework capable of emulating large wireless networks. To effectively emulate networks of this size requires considerable computational power. High-fidelity radio frequency propagation models require calculation of every possible signal path between all possible pairs of transmitters and receivers. These calculations may need to be repeated whenever there are changes in the positions of the nodes and environmental factors such as terrain and weather. This is well suited to cluster computing, as path-loss calculations can be computed in parallel.[1] Emulating the nodes of the network may be lightweight compared to RF modeling, but the number of nodes in the emulation network may outnumber the number of physical nodes in the cluster. Using virtual machines as emulation nodes permits efficient use of available resources.

Real-time emulation of wireless networks requires introducing as little latency as possible to the emulated wireless traffic. Latency can be significantly reduced by carrying emulation traffic over high-speed interconnects often used in cluster computing. Through proper hybrid EMANE deployment, it is possible to leverage both virtual machines and high-speed interconnects.

As emulation framework requirements grow, the cluster should be expandable to meet those requirements. A good cluster design should be scalable and uniform. As hardware is added or replaced, or software is upgraded, ensuring all nodes have the same system image becomes a challenge. The use of a cluster provisioning system allows new hardware and software to be deployed with minimal configuration changes. Likewise, a job control system designed for parallel computing solves the problem of starting and controlling the emulation and framework and related process

2 Overview of EMANE

The Extendable Mobile Ad-Hoc Network Emulator (EMANE) is an open-source network emulator designed to fully model the radio nodes and the emulate the wireless environment. The principal components of EMANE are the platform, the transport, and the event system. A platform consists of one or more Network Emulation Modules (NEMs) which model the wireless network from the perspective of the associated nodes. The MAC (Media Access Control, or data-link) and PHY (Physical, or RF communication) layers of the emulated radios are modeled by NEM plugins.

The transport connects the emulated radio to a network interface on the emulation node, and an Over-the-Air (OTA) manager relays the emulated RF signals to all NEMs in the emulation.

Events are generated by the event service and received by the NEMs and event daemons on the emulation nodes. Events typically modify or control the emulation environment. For example, location events indicate or change the position of the nodes in emulated space. This event can be used by the PHY plugin of the NEM to compute path-loss and also by the emulated node for positional awareness.[2]

2.1 EMANE Deployment

EMANE supports centralized, distributed and hybrid platform deployment. Centralized deployment places all NEMs on a single platform server. The transport on each node is connected over a network to the associated NEM on the platform, and the OTA manager is internal to the platform. If deploying EMANE on a single physical machine (possibly hosting virtual nodes), centralized deployment of a single platform on the host is possible. The raw transport can be used to capture traffic from network interfaces, or each virtual machine contains a transport and communicates with the platform on the host. Distributed deployment places a platform at each node to host the NEMs associated with that node. Communication between the transport and the NEM is internal to the node, and the OTA manager network carries the emulated wireless traffic between nodes. A hybrid deployment distributes the NEMs across some combination of nodes and platform servers.[3]

3 Cluster Design for EMANE

The basic cluster design consists of a set of compute nodes controlled by a head node. Nodes are connected by a conventional LAN network for management and optionally by a high-speed interconnect for computational traffic. [4]

When designing a cluster for use with EMANE, consider the EMANE functional components: the nodes of the emulated network, the NEM platforms, the OTA manager, and the event servers. In a centralized EMANE deployment, the compute nodes are the nodes of the emulation and the head node serves as the platform, the OTA manager, and the event server. In a distributed deployment, the NEMs are distributed to the compute nodes.

As the size of the emulated network increases, the platform and the event server may place an overwhelming load on the head node. To keep the head node responsive for scheduling and interactive sessions, one or more compute nodes should be assigned as dedicated event servers or NEM platforms. Therefore, three types of nodes are defined: the head node, the nodes of the emulated network, and compute nodes that actually perform the network emulation.

3.1 The Cluster Network and Interconnects

At a minimum, a cluster must have a management network through which the head node can communicate with the other nodes. Cluster management should always be on a dedicated network to avoid conditions where management of the cluster is not possible due to high network load. Conversely, computational traffic should be affected as little as possible by management overhead. Typically a cluster will have additional networks or interconnects for computational traffic.

EMANE generates two types of emulation-related traffic: OTA manager messages and event messages. OTA manager communication is between all platforms, whereas events are broadcast from the event server to the platforms and individual nodes of the emulation. Therefore, there needs to be logical or physical segmentation into three networks: Management, Event, and OTA manager. To minimize latency, if a high-speed interconnect is available, it should be used as the OTA manager.

3.2 Virtualization of Nodes

To emulate a network with more network nodes than the number of physical nodes in the cluster, a method is needed to provide each emulated node with an independent environment. This can be accomplished by hosting virtual machines on the cluster nodes.

In a distributed EMANE deployment, the NEM for each node is hosted inside the virtual machine. In a hybrid EMANE deployment, we deploy a platform server on each physical node, containing the NEMs for the virtual machines hosted on that node.

3.3 Design experiences

Cluster design and EMANE deployment to our small and medium-scale clusters differ based on functional requirements. In the case of the small cluster, the requirement was to address performance, uniformity and management while preserving as much of the original configuration as possible. Scalability was not a consideration. The medium-scale cluster is a development environment and test-bed for a design scalable by at minimum an order of magnitude. To address these issues, we implemented these following guidelines.

3.3.1 Small cluster

Our small-scale deployment is an upgrade of an existing MANE[5] emulation platform. The hardware consists of 1 head node, 8 compute nodes, and a single GPU-equipped server dedicated to event services and path-loss calculation. The compute nodes are eight-core servers capable of hosting 16 virtual machines each, for a total capacity of 128 emulated network nodes. This was designed to be a fully distributed EMANE deployment: each emulated node runs a transport, NEM, and event daemon.

A stacked pair of Cisco Catalyst 3750[6] switches are configured to provide three logical VLAN segments. The Catalyst switches are configured to provide three VLANs, plus outside connectivity to the cluster. Ports 1 through 10 of switch 1 are the management network and as such have no special configuration, placing them on the default VLAN. The management interface of each node connects to these ports.

Ports 1 through 10 of switch 2 are configured as the external network (VLAN 100). The external interface of the head node is connected to this VLAN, as is other hardware that must be accessible from the LAN.

The head node and compute nodes are equipped with two quad-port GigE NICs. The four ports on each NIC are bonded to form a 4Gb/sec channel.

Port 12 and ports 17-48 of switch 1 are the event network, and are assigned to VLAN 10. Each group of four ports starting at port 17 is configured as a bonded port-channel, and connected to one NIC of a compute node. The event server is connected to port 12. Ports 17 through 48 of switch 2 are the OTA manager network, and are assigned to VLAN 20. Each group of four ports starting at port 17 is configured as a bonded Port-channel, and connected to the remaining NIC of a compute node.

The head node is connected to the emulation networks to allow for the possibility of a NEM platform on the head node. Ports 13-16 of switch 1 and ports 13-16 of switch 2 are bonded to form an 8Gb/sec channel. Each switch is connected to a quad-port NIC of the head node. This channel is configured on the switch and the master host as a trunk with access to VLAN 10 and VLAN 20 traffic. The head node is configured to provide virtual VLAN interfaces sharing 8Gb/sec maximum bandwidth. (While the ports for the head node could be configured identically to the cluster hosts, this configuration provides the head node with extra bandwidth to both the event and OTA networks.)

Segregation of VLANs between switch 1 and switch 2 is important when the switch architecture is taken into consideration. The two switches forming the stack use a proprietary interconnect that provides 32Gb/sec throughput. By placing the management and event VLANs on switch 1, and the external and OTA VLANs on switch 2, the only traffic that will cross the interconnect is the trunked traffic to and from the head node. At 8Gb/sec, this load will not saturate the interconnect.

As this deployment is an upgrade of an existing, unmanaged deployment, some elements of the original cluster remain: The existing install of CentOS[7] 5.6 on the head node and event server was preserved, as was the existing shared storage and EMANE environment on the head node. Centos 5.6 is based on the Linux 2.6.18 kernel and lacks virtualization support[8]. For this reason Scientific Linux 6 images were customized and employed for the compute nodes and virtual machines[9].

3.3.2 Medium-scale cluster

In comparison, the medium-scale deployment was planned to be an EMANE emulation cluster from the onset. The hardware in this cluster consists of 2 head nodes (one active and one currently a cold-spare), 28 emulation platform nodes, and 2 compute nodes equipped with NVIDIA Tesla GPGPUs. Two independent Gigabit Ethernet switches provide a management network (to which all nodes are connected) and an event network (to which platform and compute nodes are connected). Infiniband adapters and a 32-port Infiniband switch provide a high-speed interconnect between platform nodes. A Panasas storage appliance is connected via Ten-Gigabit Ethernet to the management network to provide shared storage[10]. No special bonding or VLAN configuration of the management or event network switches is required.

A hybrid EMANE deployment is used on the medium-scale cluster. Each platform node can host 16 virtual machines and one platform containing the NEMs for the virtual machines on the node. This allows the platform direct access to the interconnect. This cluster is expected to emulate networks of up to 448 nodes.

For this cluster we chose Infiscale's CAOS Linux[11]. CAOS is a Redhat-based, HPC-targeted distribution that meets our criteria for a stable, cluster-ready operating system. A design goal for this deployment was to standardize on a single distribution for all nodes, and to have a single image for all managed nodes. A CAOS install onto the head node includes a deployable node image. Our goal is to maintain one image for platform, compute, and virtual nodes, so the base CAOS node image was customized with packages from the CAOS repository to enable virtual machine support, plus the NVIDIA drivers and CUDA SDK to enable OpenCL support[12]

3.4 Virtual machines and EMANE deployment

To use virtual machines as EMANE emulated network nodes, the virtual machines must have near-native performance, proper network connectivity and acceptable latency.

Our virtualization technology of choice is the KVM-enabled version of QEMU. Newer releases of the Linux kernel, coupled with virtualization-capable host hardware, allow QEMU to execute code in the virtual machine directly on the host CPU, providing true virtualization as opposed to CPU-level emulation.[13]

QEMU can be configured to emulate one or more Ethernet interfaces in the virtual machine. Virtual network interfaces are connected to VLANs, which can be connected to virtual interfaces created by the QEMU process in the host OS. This mechanism creates a channel between the virtual machine and the host. On the host side of the channel, a software bridge can be used to connect one or more virtual machines to each other and to the physical network interfaces of the host, allowing the virtual machines access to the network.[14]

Both of our EMANE clusters have three virtual interfaces per virtual machine (`eth0`, `eth1`, `eth2`) each connected a separate VLAN which is connected to a tap on the host. The host side of the `eth0` taps are added to a bridge with the physical `eth0` interface of the host, connecting the `eth0` interfaces of the virtual machines to the management network. The virtual `eth1` taps are handled in the same way (being bridged with the host event interface or bonded channel).

Connection of the virtual machine `eth2` interface depends on the type of EMANE deployment. In a fully distributed EMANE deployment, the NEMs are located at the emulation nodes, therefore inside the virtual machines. The virtual `eth2` interface is part of the OTA network, and is bridged to the other virtual `eth2` interfaces and the physical `eth2` interface. However, the VM-to-host channel, the kernel network bridge, and the physical interface all incur some degree of latency, and the resulting total latency may unacceptably delay OTA messages. To illustrate this point, Tabel 1 shows round-trip times across selected paths, measured using `mpong`[15]

	avg RTT	min RTT	max RTT
loopback	10.8	10	3210
GigE	187	86	5394
4xGigE (bonded channel)	70.1	51	1175
Infiniband	47.3	32	5220
VM to Host	118	104	5335
VM to VM (same host)	228.2	209	6706
VM to VM (over GigE)	349.1	284	5488

Table.1: Multicast RTT times (microseconds)

Non-Ethernet interfaces cannot be bridged with Ethernet interfaces[16]. This prohibits bridging the host-side tap interfaces to an Infiniband interface. QEMU does provide a workaround by which a QEMU VLAN can be encapsulated in multicast IP, but testing shows the resulting latency is equivalent to VM-to-VM over Gigabit Ethernet, negating the advantage of a high-speed interconnect.

The solution is a hybrid EMANE deployment. Moving the NEMs outside the virtual machine to a platform on the host OS moves the VM-to-host latency to the transport; the OTA latency is reduced to the latency of the OTA manager and associated network or interconnect. In this configuration, the virtual eth2 taps are bridged with each other but not with any physical interface. (In Linux, an IP address is assigned to the bridge's interface, not to the bridged interfaces, therefore the bridge interface can be used as the platform endpoint)

A script to start QEMU virtual machines with the proper interfaces, and `qemu-ifup`/`qemu-ifdown` scripts to configure the proper bridging are included in the appendices.

3.5 Cluster Provisioning with Perceus

Perceus[17] is an open-source cluster provisioning system provided by Infiscale. Perceus is included with CAOS and available for Redhat and Debian. The Perceus master runs on the head node and provides, in addition to DHCP and DNS services for the cluster, network booting of cluster nodes. Cluster nodes are PXE booted to a provisioning state, while virtual machines are started with the provisioning kernel and `initrd` (initial ramdisk image) loaded directly into memory. Nodes in the provisioning state contact the Perceus master and load a VNFS image over NFS, which contains a new kernel and compressed root filesystem. The VNFS image then executes entirely from RAM, requiring no NFS or SAN access for the operating system.

Each node should first be added to the Perceus database by management interface MAC address. The following commands will add a platform node, a compute node, and a virtual machine. (The convention we use for virtual machine MAC addresses is the format `00:01:00:vv:nn:nn`, where vv is the virtual interface index, in this case `eth0`, and `nn:nn` are the two bytes of the node number in big-endian format.)

```
perceus node add  xx:xx:xx:xx:xx:xx n0001
perceus node set group n0001 nodes
perceus node set vnfs <your vnfs image name>

perceus node add  xx:xx:xx:xx:xx:xx c0001
perceus node set group c0001 compute
perceus node set vnfs <your vnfs image name>

perceus node add  00:01:00:00:00:01 vm0001
perceus node set group vm0001 vm
perceus node set vnfs <your vnfs image name>
```

Perceus has the ability to create configuration files in the node filesystem before the new kernel is started, allowing the network interfaces to be configured in complex ways before the node boots. This is important for platform nodes that will host virtual machines. For this feature to work properly, every network interface of the node should be added to /etc/hosts with the node's hostname as an alias to the management interface. This will configure the IP addresses to be assigned to each interface. (Note that for platform nodes, we will assign IP addresses to the bridge interfaces, not the network interfaces.)

```
10.0.1.1 n0001-br0 n0001
10.1.1.1 n0001-br1
10.2.1.1 n0001-ib0

10.0.2.1 c0001-eth0 c0001
10.1.2.1 c0001-eth1

10.0.10.1 vm0001-eth0 vm0001
10.1.10.1 vm0001-eth1
10.255.10.1 vm0001-eth2
```

The ipaddr module allows Perceus to configure the network interfaces of the nodes before the VNFS is booted. The following lines configure the ipaddr module to create the proper bridges and assign IP addresses based on /etc/hosts entries. (Note the assignment of a fixed 10.255.0.1 address to br2 on each platform node, this is the bridge used for transport-to-NEM traffic and has no network connectivity.)

```
n* br0(TYPE=bridge&ENSLAVE=eth0):[default:NAME-
NIC]/255.255.0.0/10.0.0.1
br1(TYPE=bridge&ENSLAVE=eth1):[default:NAME-
NIC]/255.255.0.0 ib0:[default:NAME-
NIC]/255.255.0.0
br2(TYPE=bridge):10.255.0.1/255.255.0.0


c* eth0:[default:NAME-NIC]/255.255.0.0/10.0.0.1
eth1:[default:NAME-NIC]/255.255.0.0


vm* eth0:[default:NAME-NIC]/255.255.0.0/10.0.0.1
eth1:[default:NAME-NIC]/255.255.0.0
eth2:[default:NAME-NIC]/255.255.0.0
```

3.6 Job Management with SLURM

A compute cluster of any size can benefit from resource management. The three important aspects of resource management are allocation, control, and monitoring. Our criteria are that a resource manager must be:

Able to execute parallel jobs, so that the emulation can be started on all nodes and virtual machines.

Scriptable, as the EMANE transport, event daemon, and related processes such as GPSd must be started in the correct order.

Able to control child processes spawned by the job.

The Simple Linux Utility for Resource Management[18] fits our criteria and has the additional advantages of being lightweight, efficient, included in CAOS Linux, and having a single point of configuration. The latter is important with regard to maintaining a single VNFS image for the entire cluster: A single configuration file is shared by all head nodes and copied to the VNFS. SLURM nodes determine their role based on hostname.

SLURM provides two methods of starting jobs from the command line: srun and sbatch. srun executes the given command and arguments in parallel based on the options given. Output from the running processes is piped back to the shell. Srun returns when the remote process exits, and can optionally provided an interactive shell that will send input in parallel to all processes in the job.

sbatch is a job script submission utility. A shell script submitted with sbatch will be queued to run on the first node of the allocation when the required set of nodes is available. Commands to be run in parallel across the allocation are invoked from the script using srun. Each invocation of srun is considered a job step. Steps can be run concurrently in the background (via the "&" shell syntax). However, when the job script exits, all processes spawned from the script are terminated therefore, the script must transfer execution with "exec" or not exit while child processes are running.

SLURM passes the environment of the srun/sbatch command to the job processes, in addition to setting job-specific variables. The SLURM_PROCID and SLURM_STEPID variables are the most useful. SLURM_STEPID indicates which step of the job script is executing. SLURM_STEPID starts at zero and increments each time srun is invoked in the job script. If srun is invoked from the command line, the SLURM_STEPID is 0. SLURM_STEPID is not set in the environment of the job script. This allows a shell script to determine if it is running as a job script or the actual job. SLURM_PROCID indicates the parallel job index; starting at zero and incrementing with each parallel job spawned by that step.

Detecting these variables allows development of a single script for all use cases, and allows the job to generate the ID of the node. Consider the start_kvm script for starting virtual machines: This script could be invoked one of three ways: Directly from the shell, in parallel via srun, or submitted via sbatch. The script first checks for the SLURM_PROCID variable. If SLURM_PROCID is present but SLURM_STEPID is not, the script is being run by sbatch as a job script on the first node of the allocation, therefore, the script will invokes srun with itself as the command to run in parallel on the allocated nodes. If SLURM_PROCID is not set or SLURM_STEPID is set, the script is being invoked from the shell or is a parallel job. In either case the virtual machine should be started. The script attempts to derive the NODEID from SLURM_PROCID and falls back to the command line parameters if SLURM_PROCID is not set.

As an example, if the following command was executed:

```
sbatch -N16 -n128 start_kvm
```

SLURM would allocate 16 nodes to run 128 parallel jobs and spawn a single instance of start_kvm on the first node of the allocation. start_kvm would detect it is being run as a job script and srun itself, spawning eight instances of start_kvm on each of the 16 nodes in the allocation. These instances of start_kvm would detect they are being run in parallel and, based on SLURM_PROCID, start virtual machines with a NODEID of {1,2,3,4,5,6,7,8} on the first node, {9,10,11,12,13,14,15,16} on the second, and so on. The NODEID is used to generate the MAC addresses of the virtual interfaces, allowing Perceus to provision the nodes.

An emane_node to start the EMANE transport, event daemon, and related processes is included in the appendix. Some elements of the EMANE configuration depend on the IP address of the node, so the emane_node script determines the node ID from the hostname of the node, not by SLURM_PROCID.

4 Evaluation

To evaluate the performance of the small and medium scale clusters, we choose to measure the latency of delivering OTA manager packets to the NEM platforms.

4.1 Measurement of OTA Latency

OTA manager network latency was measured for a 16-node emulation in EMANE. Each node emits UDP broadcast traffic to the transport interface approximately once per second. The broadcast traffic is encapsulated and unicast by the transport to the NEM platform, where it is relayed to other NEMs on the platform and across the OTA manager network as UDP multicast traffic.

A packet logger is run on all NEM platform nodes, listening to the multicast group and port used by the OTA manager. The packet logger records the source IP address, receive UNIX timestamp and microseconds, IP header checksum, and IP ID of packets on the OTA. Time synchronization of all platform nodes is maintained using a Cengen-recommended method. [19]

For any given packet, the tuple of (source IP, timestamp, checksum) serves to identify the packet in each node's log. The absolute send time (as timestamp+microseconds/1000000) is determined per packet from the source node's log. For the source node, the latency is zero. For each other node, the send time is subtracted from the logged time to determine the latency. The absolute send time is translated to an elapsed time by subtracting the lowest timestamp recorded, and the resulting tuple of (time, node, latency) is output.

Figure 1 and Figure 2 are plots of time vs. latency for a distributed deployment and hybrid deployment with interconnect.

Figure 1 is a distributed deployment of 16 virtual machines across 8 nodes, with one NEM per virtual machine and one packet logger per virtual machine. The OTA manager network is quad-gigabit Ethernet bonded channel. 3475 distinct packets were logged with an average latency of 133 microseconds and a maximum latency of 774 microseconds.

Figure 2 is a Hybrid deployment of 16 virtual machines across 8 nodes, with one NEM platform and packet logger per node. The OTA manager network is IP-over-Infiniband. 2213 distinct packets were logged with an average latency of

17 microseconds, a maximum latency of 78 microseconds.

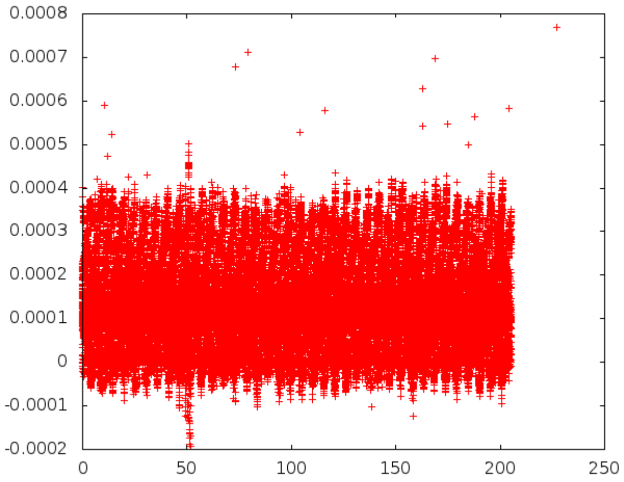


Figure 1: Distributed deployment,4xGigE, Tsend vs. Treceive-Tsend in microseconds.

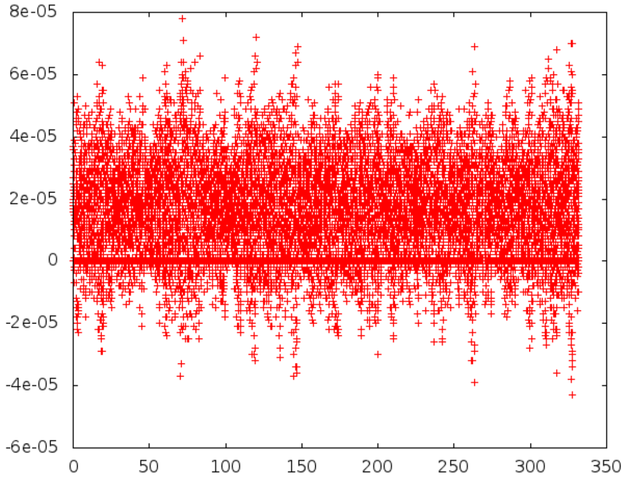


Figure 2: Hybrid Deployment, Infiniband, Tsend vs. Treceive-Tsend in microseconds.

5 Summary and Future Direction

We have presented guidelines for designing, configuring, and managing cluster systems capable of high-performance, low-latency wireless network emulations using the EMANE framework.

Future work will focus on the following:

- Develop a web-based front-end for configuring, submitting, and monitoring EMANE jobs. This will greatly simplify sharing the resources of a large cluster.

- Scale to a planned cluster of approximately 5000 cores and 500 GPGPUs across several hundred nodes. Ten-gigabit Ethernet interfaces interconnected via a fat tree of low-latency switches will serve as a shared event and OTA manager network. This is expected to support emulation of multiple networks with 5000-10000 nodes total.

References

[1] Cavalcante, A.M.; de Sousa, M.J.; Costa, J.C.W.; Frances, C.R.L.; Protasio dos Santos Cavalcante, G.; de Souza Sales, C. "3D ray-tracing parallel model for radio-propagation prediction," *Telecommunications Symposium, 2006 International* , vol., no., pp.269-274, 3-6 Sept. 2006
doi: 10.1109/ITS.2006.4433282
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4433282&isnumber=4433230

[2] Patel, K., Galgano, S. (2010). *Emulation Experimentation Using the Extendable Mobile Ad-hoc Emulator.* Retrieved from http://labs.cengen.com/emane/doc/emane-emulation_experimentation.pdf

[3] CenGen, Inc. *EMANE User Training*. Retrieved from http://labs.cengen.com/emane/doc/0.6.2/training/emaneusertraining-slides.0.6.2.20100301-1.pdf.

[4] Lehmann, T. (2009). *Building a Linux-Based High-Performance Compute Cluster*. Linux Journal. Retrieved from http://www.linuxjournal.com/magazine/building-linux-based-high-performance-compute-cluster

[5] US Naval Research Laboratory. *Mobile Ad-hoc Network Emulator (MANE).* Retrieved from http://cs.itd.nrl.navy.mil/work/mane/index.php

[6] Cisco Systems, Inc. *Cisco Catalyst 3750 Series Switches.* Retrieved from http://www.cisco.com/en/US/products/hw/switches/ps5023/index.html

[7] Centos Project. *The Community ENTerprise Operating System*. Retrieved from http://www.centos.org/

[8] *KVM*. Retrieved from http://www.linux-kvm.org/page/Main_Page

[9] *Scientific Linux*. Retrieved from http://www.scientificlinux.org/

[10] Panasas, Inc. *Panasas: Parallel File System for HPC Storage*. Retrieved from http://www.panasas.com/

[11] Infiscale, Inc. (2008). *CAOS Linux*. Retrieved from http://www.caoslinux.org/

[12] Nvidia, Inc .(2011). *CUDA Toolkit 4.0*. Retrieved June 2011 from http://developer.nvidia.com/cuda-toolkit-40

[13] Bellard, F. (2011). *QEMU open source processor emulator.* Retrieved June 2011 from http://wiki.qemu.org/
Main_Page

[14] *QEMU User Documentation.* (Sec 3.7). Retrieved June 2011 from http://qemu.weilnetz.de/qemu-doc.html#pcsys_005fnetwork

[15] Informatica, Inc. (2010). *Test Your Network's Multicast Capability.* Retrieved from http://www.29west.com/docs/TestNet/index.html

[16] http://www.linuxfoundation.org/collaborate/workgroups/networking/bridge#What_can_be_bridged.3F

[17] Infiscale, Inc. *Provision Enterprise Resources & Clusters Enabling Uniform Systems.* Retrieved from http://www.perceus.org

[18] Lawrence Livermore National Laboratory. (2011). *Simple*

Linux Utility for Resource Management. Retrieved June 2011 from https://computing.llnl.gov/linux/slurm/

[19] CenGen, Inc. *Emane 0.7.1 Documentation.* (Sec. 11). Retrieved from http://labs.cengen.com/emane/doc/0.7.1/html/emane.0.7.1.html#id370045

AISTC: A new Artificial Immune System-based Topology Control Protocol for Wireless Sensor Networks

Amir Massoud Bidgoli [1], **Arash Nikdel** [2]

[1]Department of computer engineering, Islamic Azad University, Tehran North Branch, MIEEE, Ph. d.
Manchester University, Tehran, Iran

[2] Department of computer, Science and Research Branch, Islamic Azad University, Khouzestan, Iran

Abstract—*Topology control protocols try to decrease average of node's transition radius without decreasing network connectivity. In this paper, we propose a new Artificial Immune System-based topology control protocol for wireless sensor networks named AISTC. In this protocol, proper transition radius can be determined using Artificial Immune System algorithm. This protocol is simulated and compared its functionality to some other protocols. Simulation results show high efficiency of the proposed protocol.*

Keywords: wireless sensor network, topology control, artificial immune system, network lifetime, energy consumption.

1. Introduction

The wireless sensor network (WSN) has emerged as a promising tool for monitoring the physical world. This kind of networks consists of sensors that can sense, process and communicate. Sensors can be deployed rapidly and cheaply, thereby enabling large-scale, on-demands monitoring and tracking over a wide range of applications such as danger alarm, vehicle tracking, battle field surveillance, habitat monitoring, etc [1].

Due to their portability and deployment, nodes are usually powered by batteries with finite capacity. Although the energy of sensor networks is scarce, it is always inconvenient or even impossible to replenish the power. Thus, one design challenge in sensor networks is to save limited energy resources to prolong the lifetime of the WSN [2].

A number of studies for reducing the power consumption of sensor networks have been performed in recent years. These studies mainly focused on energy efficient MAC protocols, data aggregated routing algorithms, and the applications of level transmission control. Power saving techniques can generally be classified in two categories: scheduling the sensor nodes to alternate between active and sleep mode, and adjusting the transmission or sensing range of the wireless nodes [2].

Topology control in sensor networks is coordination art of nodes by decision making about transition radius [3].Choosing appropriate topology for a sensor network has much effect on networks performance, especially considering power consumption and lifetime network.

In this paper, we propose a new topology control protocol based on Artificial Immune System, named AISTC. In this protocol, each node adjusts its transition radius using Artificial Immune System algorithm and considers the transition radius of its neighbors and the network status. The transition radius will be between minimum transition radius and maximum transition radius.

The remaining of this paper is organized as follow: Related works are explained in section 2. In section 3 problem definition is introduced. Artificial Immune System will be discussed in Section 4. Proposed protocol is explained in section 5. Simulation results are shown in section 6 and a final conclusion is discussed in Section 7.

2. Related Works

So far, many protocols have been introduced for topology control in sensor networks. Topology control protocols are divided into homogeneous and heterogeneous control topologies. In homogeneous control topology, all network nodes use the same transition radius and topology control problem is to find a minimum value for transition radius considering the network characteristic such as in heterogeneous control topology in which network nodes can have non uniform transition radius. In this group, protocols with information used for making topology are divided into three groups. First group are methods based on location. In this group, nodes are informed of their location and by using this information, a proper topology for network is made [4], [5]. Second group are methods based on orientation. In these methods, nodes don't have exact information of their location, but they can identify direction of their neighbors. Protocol CBTC[1] [6] is an example of these methods. Third group are methods based on neighbors. In these methods, nodes have limited information about their neighbors. This

[1] Cone Based Topology Control

information consists of ID number, and distance or quality of node's neighbors. Kneigh[2] [7] and XTC[3] [8] are examples in this group. RAA-2L[4] is another topology control protocol. In this protocol, each node chooses one of two transition areas R_S or R_W (R_W<R_S) [9]. If a node with transition area of R_W could communicate with a neighbor with transition area R_S, it chooses transition area node R_W, else it chooses transition area node R_S. In RAA-3L[5] , each node chooses one of three transition areas: R_t, R_S or R_W ($R_W < R_t < R_S$).

In [10], Cellular Automata is used for topology control, but it didn't consider the nodes relations and their residual energies. In our proposed topology control protocol, each node will consider the transition radius of its neighbors, its residual energy and network connectivity.

3. The Model And The Assumptions

In this section, we present the model and the assumptions used in this paper.

3.1. Adjustable transition radius

We assume that each node has adjustable transition radius that can be between a minimum and a maximum area. R_{min} is transition area with minimum power, R_{max} is transition area with maximum power and R_S is selective transition area of node. The value of R_T should be between the R_{min} and R_{max} ($R_{min} \le R_T \le R_{max}$). Value of transition area R_{min} and R_{max} will be calculated based on R_t. Value of transition area R_t is determined proportional to the network density [11]. When distance of both nodes is less than R_{max}, we assume they are neighbors. Both of neighbor nodes are in four different groups. Sets of A_{min}, A_S and R_{max} are obtained by (1). In (1), n_i is neighbor node number, and D_{ni} is the distance between current node and n_i.

$$\begin{cases} n_i \in A_{min} & \text{if } D_{ni} \le R_{min} \\ n_i \in A_S & \text{if } R_{min} \le D_{ni} \le R_S \\ n_i \in A_{max} & \text{if } R_S \le D_{ni} \le R_{max} \end{cases} \qquad (1)$$

Therefore:

$$A_{max} \cup A_S \cup A_{min} = All\ neighbor \qquad (2)$$

$$A_{max} \cap A_S \cap A_{min} = \{\} \qquad (3)$$

A_C consists of neighbor nodes that are in selective transition area or are accessible through nodes that are in selective node area. A_{min} is proper subset of A_C, because each node has minimum transition area A_{min}. Transition area and set of nodes are depicted for node n in Fig. 1. The main problem in this study is choosing minimum transition area R_T between R_{min} and R_{max} for each node without decreasing the network connectivity.

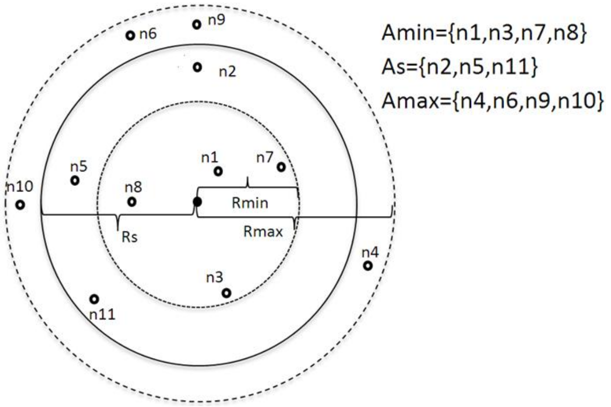


Figure1. Ttransition area and set of nodes

The transition radius of each node is coded in binary format and the required B bit is calculated using (4):

$$B = \lceil \log_2 (R_{max} - R_{min}+1) \rceil \qquad (4)$$

The transition radius of each node is calculated using (5):

$$R_T = R_{min} + (R_{max} - R_{min}+1)\ \Sigma_{b=1 \text{ to } B}\ 2^{b-1} a_{b-1} / \Sigma_{b=1 \text{ to } B}\ 2^{b-1} \qquad (5)$$

3.2. The cluster-based architecture

We introduce a cluster-based coverage control scheme in this paper, which is scheduled into rounds. In each round, firstly, the target area is divided into several equal squares. Then the node in each square having the largest energy will be chosen as the cluster-head, and the procedure of selecting the cluster-head is the same work in [15].

This cluster-based architecture is shown in Fig. 2. The nodes of cluster-heads are those asterisked ones. The black nodes represent the active ones which are working in the target area. And the red sensor nodes are these inactive ones in sleeping mode. The cluster-head has full control of the square and it will choose transition radius of nodes. In the next round, another sensor set will be turned on. It is done in a random way, so the energy consumption among all the sensors can be balanced well.

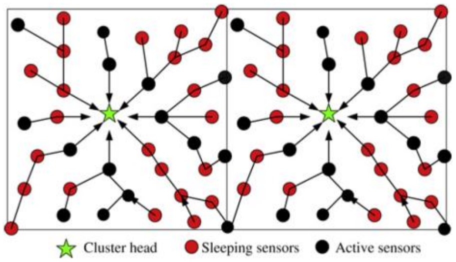


Figure2. The cluster-based architecture of AISTC protocol

3.3. Energy consumption analysis

For the brief of the energy consumption analysis, here we only consider the energy consumed by the transmission function, and do not include the power consumption of sensing and calculation.

[2] k-neighbors
[3] Extreme Topology Control
[4] Radius Adaptation Algorithm_2 Level
[5] Radius Adaptation Algorithm_3 Level

Define that the size of the monitoring area is A_{area} , the working sensor set is $n=\{\hat{n}_1, \hat{n}_2, \dots , \hat{n}_n\}$and the sensing radius set is $R=\{\hat{R}_{T1}, \hat{R}_{T2}, \dots, \hat{R}_{Tn}\}$, where $\hat{R}_{Ti}$ is the transition radius of node $\hat{n}_i$, and $\hat{R}_{Ti} \in [R_{min} , R_{max}]$.

According to different energy consumption models, the energy consumed by a node to deal with a transmission task is proportional to $R_T^{\ 2}$ or $R_T^{\ 4}$, where R_T is the transition radius of node [16]. In this paper, we take the transmission energy consumption as $u.R_T^2$, where u is the factor.

Thus, the coverage energy consumption of the sensor set, which is related to the sum of the sensor's transition radius squared, is defined as:

$$E_{total} = u. \sum_{i=1 \text{ to } n} \hat{R}_{Ti}^{\ 2} \qquad (6)$$

So, the energy consumption per area is shown as the following:

$$E_{total} / A_{area} = u. \sum_{i=1 \text{ to } n} \hat{R}_{Ti}^{\ 2} / A_{area} \qquad (7)$$

4. Artifical Immune Systems Models

AIS are distributed adaptive systems for problem solving using models and principles derived from the Human Immune System [13].

The capabilities of the AIS is mainly the inner working and cooperation between the mature T-Cells and B-Cells that are responsible for the secretion of antibodies as an immune response to antigens [14].

The different theories regarding the functioning and organizational behavior of the natural immune system (NIS) are discussed in literature. These theories inspired the modeling of the NIS into an artificial immune system (AIS) for application in non-biological environments [14].Many different AIS algorithm models have been built, including Classical View Models, Clonal Selection Theory Models, Network Theory Models, Danger Theory Models [14].

Artificial immune systems have been successfully applied to many problem domains. Some of these domains range from network intrusion and anomaly detection, to data classification models, virus detection, concept learning, data clustering, robotics, pattern recognition and data mining [14].

5. Proposed Protocol

In this section, we try to decrease average of node's transition radius without decreasing network connectivity. In proposed algorithm, at first the primary population of nodes transition radius are selected randomly. Then, the affinity rate of nodes is evaluated and based on this evaluation some of nodes are selected as memory cells and the transition radius of other nodes is mutated. The main loop of algorithm continues until the number of its repeats exceeds from threshold rate or the affinity rate of all nodes become better than threshold rate. Therefore, the proposed algorithm includes six steps as follows:

Pahse1. Problem and algorithm parameter initialization:

Step1: Initializing A_{min}, A_S and A_{max} sets for each node.

Step2: producing a transition radius mask and a transition radius mask operation for each node.

Step3: Initializing transition radius node, RT, for each node randomly.

Pahse2. Repeating main loop of algorithm until meeting termination criteria:

Step4: Calculating the affinity rate of nodes.

Step5: Selecting the nodes with more affinity rate as memory cells and mutating the transition radius of other nodes.

Step6: Checking the loop termination criteria and jumping to step 4.

5.1. Algorithm Details Description

In this section, we describe the proposed algorithm in detail:

Step1: Initializing A_{min}, A_S and A_{max} sets for each node.

At first, according to (8), the transition radius of each node is set between R_{min} and R_{max} .

$$R_T = (\hat{R}_{T1} , \hat{R}_{T2} , \hat{R}_{T3} , \dots , \hat{R}_{Tn}) \qquad (8)$$
$$\forall \hat{R}_{Ti} \in R_T : \hat{R}_{Ti} = R_{min} + (R_{min} - R_{max}) / \delta$$

Such a way that δ is a constant factor (e.g. $\delta=2$) which its rate can be determined regarding nodes density. According to (9), if δ is considered much more than R_{max}, $\hat{R}_{Ti}$ will almost equal R_{min}.

$$\text{If } \delta = \infty \Rightarrow \hat{R}_{Ti} \cong R_{min} \qquad (9)$$

Then according to (1), (2) and (3) as mentioned before, A_{min} , A_S and A_{max} sets for each node is created.

Step2: producing a *transition radius mask* and a *transition radius mask operation* for each node.

Then A_{min} , A_S and A_{max} sets are updated for each node. Whenever one node of A_S and A_{max} sets becomes a member of the set of another node, that will be removed from these sets. For this purpose, at first, we initialize A_C set by A_{min} set content and then the sets are updated according to (10):

$$A_C(N) \leftarrow A_{min}(N) \qquad (10)$$
$$\forall n_i \in A_{min}(N) \ \exists n_j \in A_{min}(n_i) \ AND$$
$$(n_j \in A_S(N) \ OR \ n_j \in A_{max}(N)) \Rightarrow$$
$$\begin{cases} A_{min}(N) = A_{min}(N) + n_j \\ A_S(N)=A_S(N) - n_j \ \ OR \ \ A_{max}(N)=A_{max}(N) - n_j \end{cases}$$

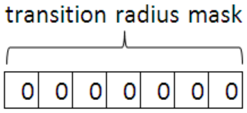
Then:

$$\forall n_i \in A_S(N) \ \exists n_j \in A_{min}(n_i) \ AND$$
$$(n_j \in A_{max}(N)) \Rightarrow$$
$$\begin{cases} A_S(N) = A_S(N) + n_j \\ A_{max}(N)=A_{max}(N) - n_j \end{cases}$$

Regarding to A_S and A_{max} condition, the node performs a transition radius mask ($Mask_{transition}$) and determines a transition radius mask operation ($Operation_{mask\_transition}$) with *OR/AND*. The method of determining transition radius mask and transition radius mask operator is calculated according to the four conditions:

- Both A_S and A_{max} sets are empty

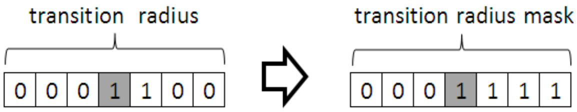
The transition radius of node is equal to A_{min}. So, radius mask is as below: (the mask operator is *AND*)



$$If\ A_S=\varphi\ and\ A_{max}=\varphi\ \Rightarrow \quad \begin{cases} Masktransition = 0 \\ Operation_{mask\_transition}='AND' \end{cases} \qquad (11)$$

- A_S set is not empty and A_{max} set is empty

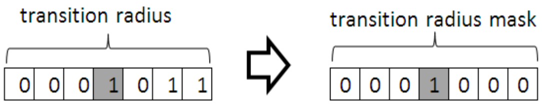
The node can select its transition radius between both R_{min} and R_T. The transition radius mask is as below: (the mask operator is *AND*)



$$If\ A_S \neq \varphi\ and\ A_{max}=\varphi\ \Rightarrow \quad \begin{cases} X = \lceil \log_2 R_T \rceil \\ Mask_{transition} = 2^X\text{-}1 \\ Operation_{mask\_transition}='AND' \end{cases} \qquad (12)$$

- A_S set is empty and A_{max} set is not empty

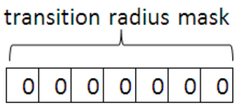
The node can select its transition radius between R_T and R_{max}. Mask of this transition radius is as below: (the mask operator is *OR*)



$$If\ A_S=\varphi\ and\ A_{max}\neq\varphi\ \Rightarrow \quad \begin{cases} X=[\log_2 R_T] \\ Mask_{transition} = 2^X \\ Operation_{mask\_transition}='OR' \end{cases} \qquad (13)$$

- Both A_S and A_{max} sets are not empty

The node can select its transition radius between A_{max} and A_{min}. This transition radius mask is as below: (the mask operator is *OR*)



$$If\ A_S\neq\varphi\ And\ A_{max}\neq\varphi\ \Rightarrow \quad \begin{cases} Mask_{transition} = 0 \\ Operation_{mask\_transition}='OR' \end{cases} \qquad (14)$$

Step3: Initializing transition radius node, A_T, for each node randomly.

In this step, according to (15), the transition radius node for each node is initialized randomly.

$$R_T = (R_T{}^{\wedge}{}_1\ , R_T{}^{\wedge}{}_2\ , R_T{}^{\wedge}{}_3\ ,\ ...\ , R_T{}^{\wedge}{}_n) \qquad (15)$$
$$\forall R_T{}^{\wedge}{}_i \in R_T : R_T{}^{\wedge}{}_i = random\ number\ between\ R_{min}\ to\ R_{max}$$

Then, according to (16), for each node, the *transition radius mask* is applied to transition radius of node by *transition radius mask operator* (*AND/OR*).

$$R_T = (R_T{}^{\wedge}{}_1\ , R_T{}^{\wedge}{}_2\ , R_T{}^{\wedge}{}_3\ ,\ ...\ , R_T{}^{\wedge}{}_n) \qquad (16)$$
$$\forall R_T{}^{\wedge}{}_i \in R_T :$$
$$R_T{}^{\wedge}{}_i \leftarrow R_T{}^{\wedge}{}_i\ Operation_{mask\_transition}(\text{AND/OR})\ Mask_{transition}$$

Step4: Calculating the affinity rate of nodes.

The process of calculating the affinity rate for each node N is as follows:

Whenever one node of A_C becomes a member of A_S or A_{max}, that node is removed from A_S or A_{max} and adds to the A_C set. See more details in (17):

$$\forall n_i \in A_C(N)\ \ \exists n_j \in A_C(n_i)\ \ And$$
$$(n_j \in A_S(N)\ Or\ n_j \in A_{max}(N)) \Rightarrow \qquad (17)$$
$$\begin{cases} A_C(N) = A_C(N)+n_j \\ A_S(N)=A_S(N)-n_j\ Or\ \ A_{max}(N)=A_{max}(N)-n_j \end{cases}$$

In (17), $A_x(y)$ shows A_x set of node y.

After updating the sets, the node determines the affinity of its selected transition radius regarding to neighbor's selected transition radius. For this purpose, the node considers a temporary TA_C set. As can be seen in (18), this set, at first, is equal to A_C.

$$TA_C(N) = A_C \qquad (18)$$

Then, according to (19), the node adds the A_S set of its neighbors to the same neighbor A_C set:

$$\forall n_i : A_C(n_i) = A_C(n_i) + A_S(n_i) \qquad (19)$$

After that, regarding the (20), the node updates TA_C set:

$$\forall n_i \in TA_C(N) \exists n_j \in TA_C(n_i)\ and$$
$$(n_j \in A_S(N)\ or\ n_j \in A_{max}(N)) \Rightarrow \qquad (20)$$
$$TA_C(N) = TA_C(N) + n_j$$

After updating TA_C set, the process of determining transition radius affinity of node is as below:

- If $A_S \subset TA_C$ and $A_{max} \subset TA_C$

At this situation, more closely the transition radius rate to R_{min}, more fit the transition radius. So:

$$affinity = \lambda_1 + \psi_1 * (my\text{-}node\text{-}A/max\text{-}node\text{-}A)(1/(R_T\text{-}R_{min}+\varepsilon)) \quad (21)$$

Where ε shows very small positive number, λ_1 shows the minimum acceptable rate for affinity of node and ψ_1 is selected as the affinity rate doesn't exceed a given limit.

- If $A_S \not\subset TA_C$ or $A_{max} \not\subset TA_C$

The node adds A_S set to TA_C set and updates TA_C set again by (20). Then, If $A_{max} \subset TA_C$ (evidently $A_S \subset TA_C$), so the transition radius will be fit and can be smaller. The details can be seen in (22):

$$affinity = \lambda_2 + \psi_2 * (my\text{-}node\text{-}A / max\text{-}node\text{-}A)(1/(R_T - R_{min} + \varepsilon)) \quad (22)$$

If $A_{max} \not\subset TA_C$, more closer the node transition radius to R_{max} , more affinity of it. So, affinity can be defined as (23):

$$affinity = \lambda_2 + \psi_2 * (my\text{-}node\text{-}A / max\text{-}node\text{-}A)(1/(R_{max} - R_T + \varepsilon)) \quad (23)$$

Where ε shows very small positive number, λ_2 shows the minimum acceptable rate for affinity node, and ψ_2 is selected as the affinity rate doesn't exceed a given limit.

In (21), (22) and (23), *my-nodes-A* shows all member of two A_S, A_{max} sets of node. The rate of *max-nodes-A* is calculated by (24). In this relation $A_x(y)$ shows the A_x set of node y.

For each node N :

$$max\text{-}node\text{-}A = max\,(a,b,c)$$

$$\begin{cases} a = max\,(|\,|A_S(n_j)| + |A_{max}(n_j)|\,|\ \ \forall n_j \in nebR_{max}) \\ b = max\,(|A_S(n_j)|\ \ \forall n_j \in nebR_S) \\ c = my\text{-}node\text{-}A = |\,|A_s(N)| + |A_{max}(N)|\,| \end{cases} \quad (24)$$

Step5: Selecting the nodes with more affinity rate as memory cells and mutating the transition radius of other nodes.

After calculating the affinity rate of the nodes, α percent of them (e.g. 50%) are selected as memory cells. It means that their transition radius doesn't mutate until next γ cycles (in the easiest mode, 1 cycle). For this purpose, the nodes are arranged in ascending order. Then α percent of nodes with more affinity rate are selected as memory cells.

After determining the memory cells, transition radius of other nodes is mutated. The ratio of mutation rate τ to affinity rate is inverses, as a result the node with more affinity rate will have less mutation and with less affinity rate, they have more mutation. The bits mutate and are selected randomly the selected bit will be inverted (zero change to 1 and vice versa).

The only operator in Artificial Immune System algorithm is mutation operator. The mutation rate is in reverse ratio to affinity rate. As a result the mutation rate for each node (τ_{node}) is in reverse ratio to that node affinity rate ($affinity_{node}$). τ_{node} is calculated for node according to (25).

$$\tau_{node} = \zeta \,/\, (\,affinity_{node} + \varepsilon\,) \quad (25)$$

Where ζ is a constant number that is calculated in the way that mutation rate doesn't become less than the determined level. Also ε is a constant number that should be selected properly in order that mutation rate doesn't exceed the determined level.

Fig. 4 shows the process of nodes mutation. In Fig. 4, the nodes with yellow transition radius are those that are selected as memory cells. As mentioned before, the nodes which are selected as memory cells don't mutate and their transition radius doesn't change. The nodes mutation rate is different as shown in Fig. 4.

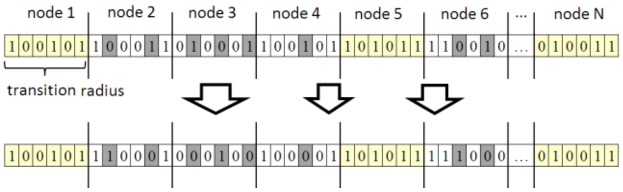


Figure4. Process of mutating the transition radiuses

The Fig. 5 shows a big binary number with four different mutation ranges. Regarding the binary rate come before and after mutation, we observe that the performance of mutation operator makes the small number bigger and big ones smaller very likely. Regarding the reverse ratio of mutation rate to affinity rate, the less node's affinity have the more node's mutation and if a number is big, It will become smaller and rice versa. The nodes with more affinity have less mutation and also less change.

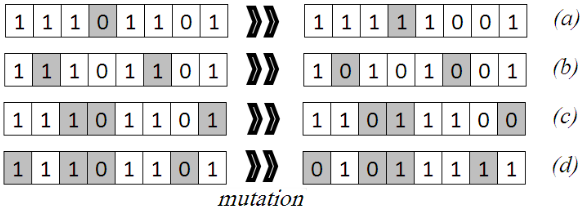


Figure5. A transition radius with four different mutation rates

Step6: Checking the loop termination criteria and jumping to step 4.

The main loop of algorithm (steps 4 and 5) continues until meeting one of the conditions stated bellow:

- The affinity rate of all nodes, become better than TA[6] threshold rate. And also this transition radiuses can provides the full connectivity of network.
- The number of performing the main loop of the algorithm exceed TC[7] threshold rate. Then, the final transition radius of nodes regarding the condition of their sets is determined. In the way that if the A_{max} set is not empty, R_{max} transition radius is selected. If A_S set is not empty, R_T transition radius is selected, otherwise R_{min} transition radius will be selected.

6. Simulation Results

In this section, our proposed protocol, AISTC, is simulated and compared to RAA-2L, RAA-3L [9] and homogeneous mode (HOM) [11] using NS2 simulator. We considered

[6] Threshold Affinity

[7] Threshold Cycles

1000×1000 m^2 area for these simulations. The number of nodes considered equal to 200, 300, 400, 500, and 600. Each node has a transition range between R_{min} and R_{max}. Transition ranges R_{min} and R_{max} are proportional to network density. Transition range R_{min} and R_{max} considered equal 87 and 136 (for 200 nodes), 69 and 108 (for 300 nodes), 59 and 93 (for 400 nodes), 54 and 84 (for 500 nodes), 48 and 75 (for 600 nodes) respectively. The ε parameter is equal 1, λ_1 and λ_2 parameters are equal 0.2, ψ_1 and ψ_2 are equal 0.3 and threshold ranges values, *TP* and *TS* considered 0.98 and 300 respectively. The energy model for these simulations is similar to energy model used in [12]. Three metrics are used for evaluations. These metrics are: average of transition area, average number of neighbors for each node, and Probability of complete connection of network nodes.

In the first experiment, we measured the average of transition area of network with 100 nodes for AISTC, RAA-2L, RAA-3L and HOM protocols. The result of this simulation is depicted in Fig. 6. As can be seen, AISTC has minimum average of transition area and HOM has maximum average of transition area.

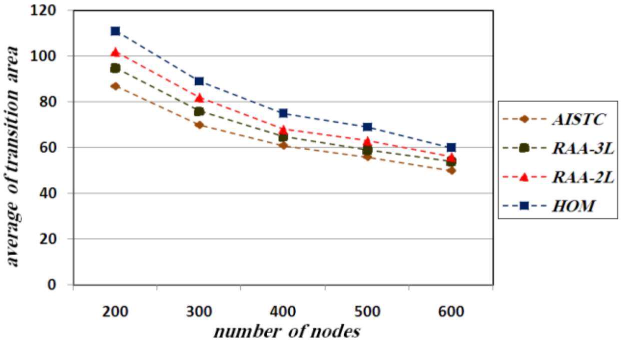


Figure6. Average of transition area Vs number of nodes

In Second experiment, average number of neighbor nodes for AISTC, RAA-2L, RAA-3L and HOM protocols is measured. The result of this experiment is depicted in Fig. 7.

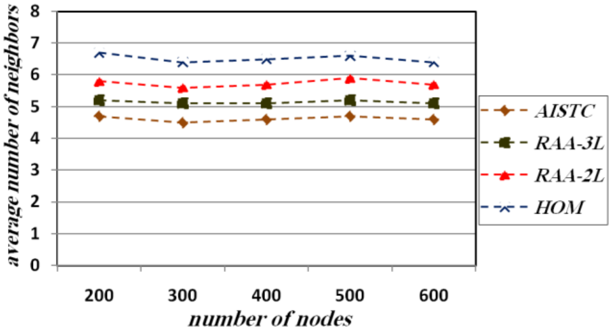


Figure7. Average number of neighbors

Protocol AISTC has minimum average number of neighbors compared to other protocols. Note that number of neighbors has a direct effect on interference between nodes and so, lower number of neighbors is better.

In the last experiment, network connectivity in AISTC is measured and compared to RAA-2L, RAA-3L, HOM and MAX-RANGE protocols. Note that, in MAX-RANGE, all of nodes have maximum transition radius. Network connectivity means ability of communicating with all of network nodes. For this purpose, we will define the concept of complete connectivity of network probability, P_C as (26):

$$P_C = \sum_{i=1 \text{ to } N_d} C_i / N_d \qquad (26)$$

$$C_i = \begin{cases} 1 & \text{if } mcp = N_n \\ 0 & \text{other} \end{cases}$$

In (26), *mcp* is the biggest component connected to network and N_n is the number of network nodes and N_d is the number of different configuration of network nodes. In this experiment, we suppose N_d is equal to 100. Probability of complete connecting of network nodes is depicted in Fig. 8 for AISTC, RAA-2L, RAA-3L, HOM and MAX-RANGE protocols.

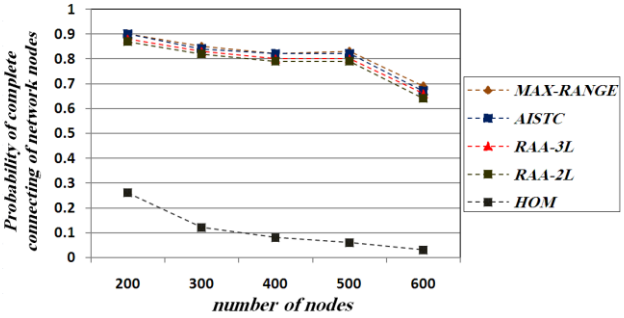


Figure8. Probability of complete connecting of network nodes

As can be seen in Fig 8, probabilities of complete connecting of network nodes for AISTC, RAA-2L, RAA-3L and MAX-RANGE are almost equal. So, network connectivity in our protocol is acceptable.

This result can show prominence of our considered mechanism. While maintaining network connectivity, they could decrease the average transition radius and the average number of network neighboring nodes and consequently it decreases energy consumption and interference between network nodes.

7. Conclusion

In this paper, we proposed a topology control protocol based on Artificial Immune System. In this protocol, nodes can select proper transition radius. Simulation results showed that the proposed protocol has some advantages compared to previous protocols. First advantage is minimum average of transition area and dynamic adjustment of the radius ratios, unlike previous protocols that should select radius ratios among predefined values. Second advantage is that our protocol has minimum average number of neighbors compared to existing protocols. So, the energy consumption in our protocol is less than others and the network lifetime will be prolonged. In addition, we showed that the network connectivity in our protocol is in the acceptable level.

8. References

[1] A. Jie Jia, C. Jian, C. Guiran, C. Zhenhua , “Energy efficient coverage control in wireless sensor networks based on multi-objective genetic algorithm” ,Elsevier , Computers and Mathematics with Applications 57 (2009) 1756_1766

[2] A. Jie Jia, C. Jian, B. Guiran, W. Yingyou, S. Jingping , “Multi-objective optimization for coverage control in wireless sensor network with adjustable sensing radius” , Elsevier, Computers and Mathematics with Applications 57 (2009) 1767_1775

[3] P. Santi, “Topology Control in Wireless Ad Hoc and Sensor Networks”, Wiley, 2005.

[4] V. Rodoplu and T. H. Meng, "Minimum energy mobile wireless networks", in: Proceedings of the IEEE Journal on Selected Areas in Communications, Vol. 17, pp. 1333-1344, 1999.

[5] N. Li, J. Hou and L. Sha., “Design and analysis of an mst-based topology control algorithm”, in: Proceedings of the IEEE Infocom, Vol. 4, pp. 1195- 1206, May 2005.

[6] R. Wattenhofer, L. Li, P. Bahl and Y. Wang, “Distributed topology control for power efficient operation in multihop wireless ad hoc networks”, in: Proceedings of the IEEE Infocom, Vol. 3, pp. 1388– 1397, 2001.

[7] D. Blough, M. Leoncini, G. Resta. and P. Santi, “The k-neighbors protocol for symmetric topology control in ad hoc networks”, in: Proceedings of the ACM MobiHoc 03, pp. 141–152, 2003.

[8] R. Wattenhofer and A. Zollinger, “XTC: a practical topology control algorithm for ad-hoc networks”. in: Proceedings of the 18th International Parallel and Distributed Processing Symposium, pp. 2-16, 26-30 April 2004.

[9] A.Venuturumilli and A. Minai., “Obtaining Robust Wireless Sensor Networks Throuh Self-Organization of Heterogeneous Connectivity”, Proceedings of the 2006 International Conference on Complex Systems (ICCS'06), Boston, MA, June 2006.

[10] M.R. Meybodi and S.Abolhasani , “ Usage of Learning Automata for Topology Control in Wireless Sensor Network” ,2008.

[11] D. Stauffer and A. Aharony, “Introduction to Percolation Theory”, London: Taylor & Francis, 1994.

[12] W. Heinzelman, A. Chandrakasan and H. Balakrishnan , “Energy Efficient Communication Protocol for Wireless Microsensor Networks”, Intl. Conf. on System Sciences, Hawaii, vol. 2, pp. 3005-3014 January 2000.

[13] Amir Massoud Bidgoli, Abdol Karim Javanmardi, Amir Masoud Rahmani,“Application of AIS algorithm for optimization of TORA protocol in ad hoc network”, in: IEEE 2010

[14] Andries P. Engelbrecht , “Computational Intelligence An Introduction” ,second edition, wiley 2007

[15] J. Jia, J. Chen, Y. Wen, G. Chang, An extensible core-control routing protocol in large scale ad-hoc networks, in: Proc. of the 6th International Conference on ITS Telecommunications, Chengdu, China, 2006, pp. 955_958.

[16] M. Lu, J. Wu, M. Cardei, M. Li, Energy-efficient connected coverage of discrete targets in wireless sensor networks, in: Proc. of the International Conference on Computer Networks and Mobile Computing, ICCNMC, Zhangjiajie, China, 2005, pp. 43_52.

NetQTM: Node Configuration In Network Setup By Quantum Turing Machine

Mehdi Bahrami[1], Peyman Arebi[2], Mohammad Bahrami[1]

[1]Department of Computer Engineering, Islamic Azad University, Booshehr Branch, Iran
[2]The Holy Prophet Higher Education Complex, Booshehr, Iran

Abstract - *The quantum Turing machine (QTM) has been introduced by Deutsch as an abstract model of quantum computation. In this paper we try to introduction the new transition function of a QTM can be used for any node configuration in the network. In this paper we introduce the fundamentals of NetQTM like a well-observed lemma and a completion lemma. The introduction of such an abstract machine allowing classical and quantum computations is motivated by the emergence of models of quantum computation like the one-way model. Furthermore, this model allows a formal and rigorous treatment of problems requiring classical interactions, like the halting[8] of QTM. Finally, it opens new perspectives for the construction of a universal QTM.*

Keywords: QTM, NetQTM, Quantom Turing Machin, Node Configuration

1. Introduction

A Turing machine (TM)[1] is a basic abstract symbol-manipulating machine which can simulate any computer that could possibly be constructed.

A Turing machine that is able to simulate any other Turing machine is called a Universal Turing machine (UTM, or simply a universal machine) [17, 18 and 19].

Studying abstract properties of TM and UTM yields many insights into computer science and complexity theory [2, 3, and 4].

Turing and others proposed mathematical computing models allow the study of algorithms independently of any particular computer hardware. This abstraction is invaluable [16].

2. Informal Description

A Turing machine consists of:

- A tape
- A head
- A table
- A state register

A tape is divided into cells, one next to the other. Each cell contains a symbol from some finite alphabet. The alphabet contains a special blank symbol (here written as 'B') and one or more other symbols [5,6].

The tape is assumed to be arbitrarily extendable to the left and to the right. Cells that have not been written to before are assumed to be filled with the blank symbol.

In some models:

- The tape has a left end marked with a special symbol
- The tape extends or is indefinitely extensible to the right

A head can read and write symbols on the tape and move the tape left and right one (and only one) cell at a time. In some models the head moves and the tape is stationary.

A table (transition function) [9] of instructions (usually 5-tuples but sometimes 4-tuples) for:

- The state the machine is currently in and
- The symbol it is reading on the tape tells the machine what to do [7].

In some models, if there is no entry in the table for the current combination of symbol and state then the machine will halt. Other models require all entries to be filled.

In case of the 5-tuple models:

(i) Either erase or write a symbol, and then

(ii) Move the head ('L' for one step left or 'R' for one step right), and then

(iii) Assume the same or a new state as prescribed.

In the case of 4-tuple models:

(ia) erase or to write a symbol or

(ib) move the head left or right, and then

(ii) Assume the same or a new state as prescribed, but not both actions (ia) and (ib) in the same instruction.

A state register stores the state of the Turing table. The number of different states is always finite and there is one special start state with which the state register is initialized.

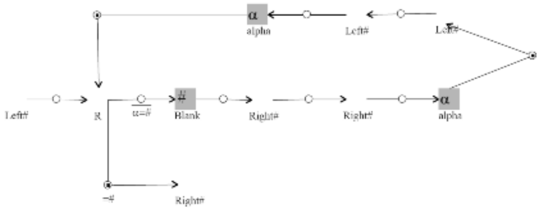
3. Formal Description

A (one-tape) Turing machine is a 7-tuple $M = \langle Q, \Gamma, b, \Sigma, \delta, q_0, F \rangle$ where

- Q is a finite set of states

- Γ is a finite set of the tape alphabet/symbols
- $b \in \Gamma$ is the blank symbol
- Σ, a subset of Γ not including b, is the set of input symbols
- $\delta : Q \times \Gamma \rightarrow Q \times \Gamma \times \{L, R\}$ is the transition function, where L is left shift and R is right shift.
- $q_0 \in Q$ is the initial state
- $F \subseteq Q$ is the set of final or accepting states

Example – Copy String



In 1985, Deutsch [10] proposed the first model of a quantum Turing machine (QTM), elaborating on an even earlier idea by Feynman [11]. Bernstein and Vazirani [12] worked out the theory in more detail and proved that there exists an efficient universal QTM (it will be discussed in Section 2.2 in what sense). A more compact presentation of these results can be found in the book by Gruska [14]. Ozawa and Nishimura [13] gave necessary and sufficient conditions that a QTM's transition function results in unitary time evolution. Benioff [15] has worked out a slightly different definition which is based on a local Hamiltonian instead of local transition amplitude [18].

The definition of QTMs that we use in this paper will be completely equivalent to that by Bernstein and Vazirani. Yet, we will use some different kind of notation which makes it easier (or at least clearer) to derive ana-lytic estimates like "how much does the state of the control change at most, if the input changes by some amount?" Also, we use the word QTM not only for the model itself, but also for the partial function which it generates.

4. NetQTM

To understand the notion of a quantum Turing machine (QTM), we first explain how a classical Turing machine (TM) is defined.

We can think of a classical TM as consisting of three different parts: a control C, a head H, and a tape T. The tape consists of cells that are indexed by the integers, and carry some symbol from a finite alphabet Σ. In the simplest case, the alphabet consists of a zero, a one, and special blank symbol #. At the beginning of the computation, all the cells are blank, i.e. carry the special symbol #, except for those cells that contain the input bit string.

The head points [24] can connect to one of the cells. It is connected to the control, which in every step of the computation is in one "internal state" q out of a finite set Q. At the beginning of the computation, it is in the initial state $q_0 \in Q$, while the end of the computation (i.e. the halting of the TM) is attained if the control is in the so-called final state $q_f \in Q$.

The computation itself, i.e. the TM's time evolution, is determined by a so-called transition function δ: depending on the current state of the control q ∈ Q and the symbol σ ∈ Σ which is on the tape cell where the head is pointing to, the TM turns into some new internal state q'∈Q, writes some symbol σ'∈ Σ onto this tape cell, and then either turns left (L) or right (R). Thus, the transition function δ is a map δ: Q × Σ → Q × Σ × {L,R}.

As an example, we consider a TM with alphabet Σ = {0, 1,#}, internal states Q = {q0, q1, qf} and transition function δ, given by

$$\begin{array}{rcl} q_0, 0 & \overset{\delta}{\mapsto} & q_1, 1, R \\ q_0, 1 & \overset{\delta}{\mapsto} & q_1, 0, R \\ q_1, 0 & \overset{\delta}{\mapsto} & q_1, 1, R \\ q_1, 1 & \overset{\delta}{\mapsto} & q_1, 0, R \\ q_1, \# & \overset{\delta}{\mapsto} & q_f, \#, R. \end{array}$$

We can define δ at these arguments in an arbitrary way. We imagine that this TM is started with some input bit string s, which is written onto the tape segment [0, L(s)-1].

The head initially points to cell number zero. The computation of the TM will then invert the string and halt. As an example, in Figure 2.1, we have depicted the first steps of the TM's time evolution on input s = 10.

A QTM is now defined analogously as a TM, but with the important difference that the transition function is replaced by transition amplitude.

That is, instead of having a single classical successor state for every internal state and symbol on the tape, a QTM can evolve into a superposition of different classical successor states.

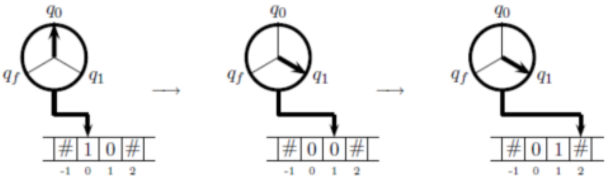


Figure 1: Time evolution of a Turing machine

For example, we may have a QTM that, if the control's internal state is q0 ∈ Q and the tape symbol is a 0, may turn into internal state q1 and write a one and turn right, as well as writing a zero and turning left, both at the same time in superposition, say with complex amplitudes

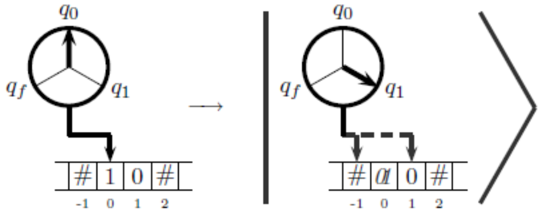


Figure 2: One step of time evolution of a quantum Turing machine

A symbolic picture of this behavior is depicted in Figure 2. This can be written as

$$q_0, 0 \overset{\delta}{\mapsto} \underbrace{(q_1, 1, R)}_{\frac{1}{\sqrt{2}}}, \underbrace{(q_1, 0, L)}_{\frac{-1}{\sqrt{2}}}.$$

Formally, the transition amplitude δ is thus a mapping from Q × Σ to the complex functions on Q× Σ × {L, R}. If the QTM as a whole is described by a Hilbert space HQTM, then we can linearly extend δ to define some global time evolution on HQTM. We have to take care of two things:

• According to the postulates of quantum mechanics, we have to construct δ in such a way that the resulting global time evolution on HQTM is unitary.

• The complex amplitudes which are assigned to the successor states have to be efficiently computable, which has the physical interpretation that we should be able to efficiently prepare hardware (e.g. some quantum gate) which realizes the transitions specified by δ .

Moreover, this requirement also guarantees that every QTM has a finite classical description, that there is a universal QTM (see discussion below), and that we cannot "hide" information (like the answer to infinitely many instances of the halting problem) in the transition amplitudes.

Consequently, Bernstein and Vazirani [4] define a quantum Turing machine M as a triplet (Σ,Q, δ), where Σ is a finite alphabet with an identified blank symbol #, Q is a finite set of states with an identified initial state q_0 and final state q_f 6= q_0, and $\delta : Q \times \Sigma \to \tilde{\mathbb{C}}^{Q\times\Sigma\times\{L,R\}}$ is the so-called the quantum transition function, determining the QTM's time evolution in a way which is explained below.

Here, the symbol $\tilde{\mathbb{C}}$ denotes the set of complex numbers that are efficiently computable. In more detail, $\alpha \in \tilde{\mathbb{C}}$ if and only if there is a deterministic algorithm that computes the real and imaginary parts of α to within 2^{-n} in time polynomial in n.

Every QTM evolves in discrete, integer time steps, where at every step, only a finite number of tape cells is non-blank. For every QTM, there is a corresponding Hilbert space

$$\mathcal{H}_{QTM} = \mathcal{H}_{\mathbf{C}} \otimes \mathcal{H}_{\mathbf{T}} \otimes \mathcal{H}_{\mathbf{H}},$$

where $H_C = C^Q$ is a finite-dimensional Hilbert space spanned by the (or-thonormal) control states q ∈ Q, while $\mathcal{H}_{\mathbf{T}} = \ell^2(T)$ and $\mathcal{H}_{\mathbf{H}} = \ell^2(\mathbb{Z})$ are separable Hilbert spaces describing the contents of the tape and the position of the head. In this definition, the symbol T denotes the set of classical tape configurations [20] with finitely many non-blank symbols, i.e.

$$T = \left\{ (x_i)_{i\in\mathbb{Z}} \in \Sigma^{\mathbb{Z}} \mid x_i \neq \# \text{ for finitely many } i \in \mathbb{Z} \right\}.$$

For our purpose, it is useful to consider a special class of QTMs with the property that their tape T consists of two different tracks an input track I and an output track O. [21 ,22] This can be achieved by having an alphabet which is a Cartesian product of two alphabets, in our case Σ = {0, 1,#} × {0, 1,#} [23]. Then, the tape Hilbert space H_T can be written as

$\mathcal{H}_{\mathbf{T}} = \mathcal{H}_{\mathbf{I}} \otimes \mathcal{H}_{\mathbf{O}}$, thus:

$$\mathcal{H}_{QTM} = \mathcal{H}_{\mathbf{C}} \otimes \mathcal{H}_{\mathbf{I}} \otimes \mathcal{H}_{\mathbf{O}} \otimes \mathcal{H}_{\mathbf{H}}.$$

5. Conclusion

In this paper, we have formally defined Network Configuration Quantum Turing Machine, based on QTM, and have given rigorous mathematical proofs of its basic properties. In particular, we have shown that the quantum complexity notions can use for every node configuration on the network and recognize by other node or transfer data by transfer function on the QTM properties.

6. References

[1] Paul Benio . Models of quantum turing machines. Fortschritte der Physik, 46:423, 1998.

[2] Gilles Brassard. Quantum computing: the end of classical cryptography? SIGACT News, 25(4):15{21, 1994.

[3] Marco Carpentieri. On the simulation of quantum turing machines. The-ory of Computer Science, 304(1-3):103{128, 2003.

[4] Yosee Feldman and Ehud Shapiro. Spatial machines: a more realistic approach to parallel computation. Communications of ACM, 35(10):60{73, 1992.

[5] Joel David Hamkins. Infinite time Turing machines. Minds and Machines,12(4):521{539, 2002.

[6] Toby Ord. Hypercomputation: computing more than the Turing machine. CoRR, Department of Computer Science, University of Melbourne, 2002.

[7] Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Always good Turing: Asymptotically optimal probability estimation. focs '03: Pro-ceedings of the 44th annual ieee symposium on foundations of computer science, vol 302, no 5644. pages 427{431, 2003.

[8] Alan Turing. Halting problem of one binary horn clause is undecidable. pages Ser. 2, Vol. 42, 1937.

[9] M. Ozawa and H. Nishimura, " Local Transition Functions of Quantum Turing Machines " , Theoret. Informatics and Appl. 34 379-402, 2000.

[10] D. Deutsch, " Quantum theory, the Church-Turing principle and the universal quantum computer" , Proc. R. Soc. Lond. A400,1985.

[11] R. Feynman, "Simulating physics with computers" , International Jour-nal of Theoretical Physics 21 467-488, 1982.

[12] E. Bernstein, U. Vazirani, " Quantum Complexity Theory" , SIAM Jour-nal on Computing 26 1411-1473, 1997.

[13] A. A. Brudno, “Entropy and the complexity of the trajectories of a dynamical system”, Trans. Moscow Math. Soc. 2 127-151, 1983.

[14] J. Gruska, Quantum Computing, McGraw-Hill, London (1999)

[15] P. Benioff, “Models of Quantum Turing Machines”, Fortsch. Phys. 46 423-442, 1998.

[16] A. S. Holevo, “Statistical Structure of Quantum Theory”, Springer Lecture Notes 67, 2001.

[17] R. Jozsa, M. Horodecki, P. Horodecki and R. Horodecki, “Universal Quantum Information Compression”, Phys. Rev. Lett. 81 1714-1717, 1998.

[18] A. Kaltchenko, E. H. Yang, “Universal compression of ergodic quantum sources”, Quantum Information and Computation 3, No. 4 359-375, 2003.

[19] G. Keller, Wahrscheinlichkeitstheorie, “Lecture Notes, Universit¨at Erlangen-N¨urnberg , 2003.

[20] J. Kieffer, “A unified approach to weak universal source coding”, IEEE Trans. Inform. Theory 24 No. 6 674-682, 1978.

[21] A. N. Kolmogorov, “Three Approaches to the Quantitative Definition on Information”, Problems of Information Transmission 1 4-7, 1965.

[22] K. Kraus, “States, Effects, and Operations: Fundamental Notions of Quantum Theory”, Springer Verlag, 1983.

[23] M. Li and P. Vitanyi, An Introduction to Kolmogorov Complexity and Its Applications, Springer Verlag, 1997.

[24] Mehdi Bahrami, Peyman arebi, "A Binomial Heap Algorithm for Self-Recognition in Exclusive Management on Autonomic Grid Networks", 2nd Int Conference on Computational Intelligence, Communication Systems and Networks - 14-N Parallel and Distributed Architectures and Systems, pp. 326 - 329 , 2010.

Q-learning Based Adaptive Zone Partition for Load Balancing in Multi-Sink Wireless Sensor Networks

Sheng-Tzong Cheng and Tun-Yu Chang
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, R.O.C.

Abstract - *In many researches on load balancing in Multi-Sink WSNs, sensors usually choose the nearest sink as destination for sending data. If all sensors in this area all follow the nearest-sink strategy, sensors around nearest sink called hotspot will exhaust energy early and this sink is isolated from network. In this paper, we propose a load balancing scheme for multi-sink WSNs. A mobile anchor with directional antenna is introduced to adaptively partition the network into several zones so the traffic load in the region can be assigned to the sink. Besides, to adapt to different data traffic pattern, we apply machine learning to mobile anchor and implement a Q-learning agent. Through interactions with environment, the agent can discovery a near-optimal control policy for movement of mobile anchor and achieve minimization of residual energy's variance among sinks, which prevent the early isolation of sink and prolong the network lifetime.*

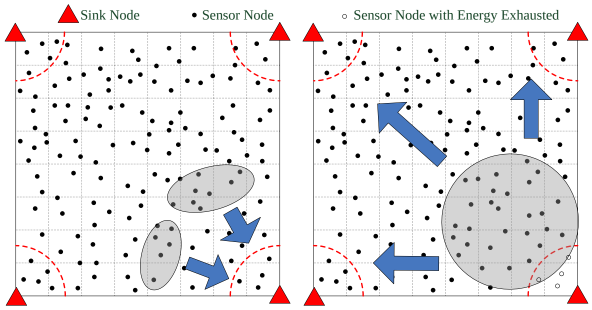
Keywords: Wireless Sensor Network, Multi-Sink, Load Balancing, Machine Learning, Q-learning

1 Introduction

The wireless sensor networks (WSNs) are widely used in a large variety of applications such as military, ocean and wildlife monitoring. The WSNs consist of a large number of low-cost devices called sensors, which monitor current status of environment and send sensing data to the sink node. Because of the limitations of the energy supply, storage space and computation of sensor nodes, the data transmitted between a sensor node and the sink node must been forwarded by other sensor nodes. The multi-hop WSNs with one sink have been developing for a long time, but many limitations exist in the kinds of architecture: robustness, scalability and reliability due to their complete dependence on the only one sink in large-scale WSNs. As all the sensors around sink exhaust energy, the sink will be isolated from the WSNs. It means that the WSNs lose its functionality. All these limitations make single WSNs infeasible in real applications. For this reason, the architecture of multi-sink WSNs is proposed.

The multi-sink wireless sensor network is a WSN with multiple sink nodes. Compared with single-sink, a multi-sink WSN has some advantages as follow: (1) it can avoid the breakdown of the whole networks in a single-sink WSN; (2) it can decrease the length of the communication path and prolong the lifetime of sensor networks; (3) it can balance the network traffic load and improve network performance. Therefore, lots of researchers have been working on energy efficient routing in multi-sink WSNs.

Many energy efficient routing algorithms [4][5][6] have been proposed to prolong the lifetime of sensor network. Sensors usually choose the nearest sink as destination for sending data. However, in WSNs, events often occur in specific area. If all sensors in this area all follow the Nearest-Sink strategy, sensors around nearest sink called hotspot will exhaust energy early. Algorithms that only consider sensor nodes will lead to early isolation of specific sink (EISS) problem in asymmetrical data generation environment. It means that this sink is isolated from network early and numbers of routing paths are broken.



(a) Asymmetrical traffic load

(b) Increase delay time and hop distance

Fig. 1. The Early Isolation of Specific Sink Problem

More specifically, sensors around specific sink will exhaust energy earlier due to a large number of data-forwarding. In Fig. 1.a, data generation rate in some area is higher than others. If all sensors inside the area send data to sink depend on the nearest sink strategy, sensors around sink in the lower right corner rapidly exhaust their energy due to the considerable times of data-forwarding. The lower right sink will be isolated from WSNs once these sensors near the sink exhaust battery. Furthermore, in Fig. 1.b, the EISS leads to that a lot of sensors switch the destination to farther sink. This situation considerably increases the energy consumption for overall network and accelerates the isolation rate for other sinks.

For solving the above-mentioned problem, we propose the Q-learning based adaptive zone partition (QAZP) scheme. By the way of QAZP scheme, the whole network is partitioned into numbers of zones for each sink, and the size of zones is adjusted for balancing power consume according to the residual energy of sensor nodes nearby each sink.

The rest of this paper is organized as follows. In section 2, we introduce some related work regarding energy efficient routing algorithms in multi-sink WSNs. In section 3, we propose the system architecture for QAZP scheme and formulate the decision making problem. In section 4, we elaborate the details of the learning agent implemented by QAZP scheme. In section 5, we compare our new QAZP scheme with the previous methods. We conclude our work in section 6.

2 Related Work

Several authors have developed data-centric routing algorithms in WSNs [1][2][3][4][5][6]. Unlike hierarchical routing algorithms, data-centric routing algorithms can support a large scale network with multiple sink nodes as well as multi-hop routing while sensing data communications. These algorithms are based on either the finding the nodes with minimal hop count or the computation the nodes and paths with least energy consumption to prolong the lifetime or increase capacity of the network.

In [1][2][3], sensors will choose the nearest sink in geometer as their destination for sending sensing data. Since that the sensing data can arrive at nearest sink through minimum hop distance, this protocol can achieve the purpose of least energy consumption for the whole network. However, the geographically asymmetric generation of events may lead to asymmetric energy consumption.

In [4], authors propose a PB (path bottleneck-oriented) and EC (energy cost-based) routing algorithm (PBEC) to optimize the balance of network-wide energy consumption. The authors investigate the effect of choosing different values for weight coefficients between PB and EC on the network lifetime performance. However, the tradeoff between PB and EC could not suit for kinds of environment, especially enduring to transmit sensing data when some nodes exhaust their energy.

In [5], authors propose the routing algorithms ELBR (Energy Level Based Routing) and PBR (Primary Based Routing) in multi-sink sensor networks. The energy level of nodes and energy cost of paths are defined to help routing packets: ELBR choose the path with the maximum energy level in order to transmit more times, PBR takes both the energy level and the energy cost of the routing path into consideration. Through the energy consumption is more balanced and the network lifetime is more prolonged, it still has high overhead while each sensor node acquires residual energy of its neighbors.

In [6], the authors present a MSLBR (multi-sink and load-balance routing) algorithm to balance the loads among the neighbors of sink nodes and prolong the network lifetime. Each node’s packets consider shortest communication hops from itself to a neighbor of one sink node in a round-robin fashion. With the destination selection strategy, the traffic load can be uniformly distributed among neighbors of sink nodes. However, it does not consider the location information about sensing node and neighbors of sink nodes so the sensor often selects a destination which is farthest and the overall energy consumption and data delay time will increase.

3 System Architecture and Problem Formulation

3.1 System Architecture

The system architecture of Q-learning based adaptive zone partition (QAZP) for load balancing in multi-sink wireless sensor networks we considered is shown in Fig. 2. The architecture typically consists of a task manager, a base station, a mobile anchor, several sink nodes and a large number of sensor nodes. Is has the following characteristics: (1) there are a large number of sensors with a unique identity in large scale WSNs and they have the same initial energy and communication range; (2) multiple sinks are deployed in WSNs, and all sinks have infinite amount of energy; (3) the existence of a controllable mobile anchor (MA), which is equipped with directional antenna and GPS device.

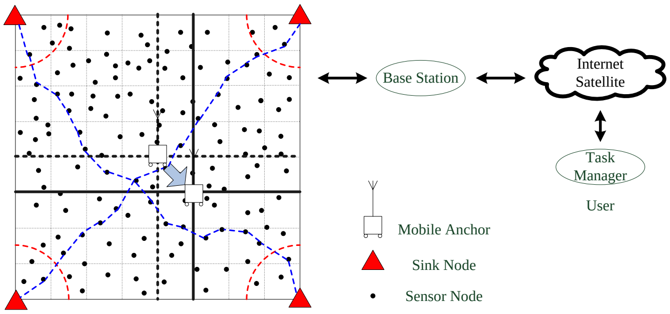


Fig. 2. System Architecture of QAZP

The sensor node establishments the several route paths to transmit sensing data to one of sink nodes. By the way of the MA is introduced to adaptively partition the network into several zones and the directional antenna are powerful enough to send beacon signals that can be heard by all sensors in WSNs, each zone is responsible for collecting sensing data to the sink. While the MA moves, a sensor node receives new broadcast message and then transmits follow-up sensor data to new assigned sink node. The system architecture of QAZP can adaptively distribute the traffic load among hotspots around sinks and avoid the EISS problem via the movement of the MA.

Moreover, we apply a learning agent to the MA for learning under unknown and stochastic environments: The movement of the MA is a reinforcement learning (RL) problem, and we resolve by a heuristic algorithm described in next.

3.2 Problem Formulation

The agent inside the MA may choose an action to let the MA be moved to vicinity or remain at current location. The MA can move a fixed distance toward eight different directions. By the movement of the MA, we adaptively partition the network into numbers of data collection zones for each sink. After a period of operation time in WSNs, the residual energy of hotspots may change into unbalance due to asymmetrical data generation. If the residual energy of hotspots around upper left corner sink is the highest, the agent shifts the MA to the lower right corner. Then the MA rebroadcasts sink-assignment packets to each sensor again from the new location. After the movement, the data collection zone of upper left corner sink is spread. Relatively, the other three sink's collection zone began to shrink.

We take into account the following two factors for definition of state: residual energy of hotspots for each sink and location of the MA in Fig.3. The state of *n* sinks WSNs in time *t* is defined as $s_t = \{\ E_{avg}(S_1), E_{avg}(S_2), \ldots, E_{avg}(S_n), X, Y\ \}$ where $E_{avg}(\ S_n\ )$ denotes the average residual power of one hop neighbors (hotspots) of sinks *n*. *X* and *Y* respectively denote the x-coordinate and y-coordinate of the MA.

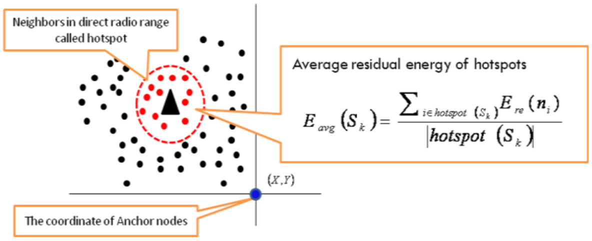

Neighbors in direct radio range called hotspot
Average residual energy of hotspots
$E_{avg}(S_k) = \frac{\sum_{i \in hotspot\ (S_k)} E_{re}(n_i)}{|hotspot\ (S_k)|}$
(X,Y)
The coordinate of Anchor nodes

Fig. 3. Average Residual Energy in Current System State

The reward function *r(s,a)* represent the immediate reward from environment due to a selected action *a* at state *s*. It is defined as

$$r(s_t, a_k) = -\sqrt{\frac{1}{N}\left[\sum_{i=1}^{N} E_{avg}(S_i)^2\right] - \overline{E_{avg}(S_i)}} \qquad (1)$$

where *N* is the number of sinks in WSNs, $E_{avg}(\ S_i\ )$ denotes the average residual energy among hotspots around sink *i*, and $\overline{E_{avg}(S_i)}$ denotes the average number among every $E_{avg}(\ S_i\ )$. Actually, $r(s_t,a_k)$ represents the standard derivation among every $E_{avg}(\ S_i\ )$. Moreover, this expression yields a negative quantity that results in lower magnitude values depicting a large reward and higher magnitude values representing a smaller reward. Specifically, the higher standard derivation represents that large differences of energy among hotspots, then the agent can only obtain less reward value from environment. On the contrary, it gets more reward value that represents that the residual energy of hotspots around sinks is well balanced.

The agent in MA learns to find an optimal policy through interactions with whole wireless sensor network and discovery optimal policy from this experience. The policy is to find the maximum return in states. The return is the sum of the rewards:

$$R_t = r(s_1, a_1) + r(s_2, a_2) + r(s_3, a_3) + \cdots + r(s_T, a_T) \qquad (2)$$

where T is a final time step. In infinite-horizon processes, a mechanism known as discounting is applied to control the rate at which rewards are accrued. The additional concept that we need is that of *discounting*. In particular, the agent maximizes the expected *discounted return*:

$$R_t = r(s_1, a_1) + \gamma r(s_2, a_2) + \gamma^2 r(s_3, a_3) + \cdots = \sum_{k=0}^{\infty} \gamma^k r(s_i, a_i) \qquad (3)$$

where γ is a parameter, $0 \le \gamma \le 1$, called the *discount rate*.

Finally, we introduce an evaluation function known as Q-function [8] to formulate the object of QAZP. It is defined as *Q(s,a)* which denotes the total discounted reward counting from the starting *(s,a)* state-action pair over an infinite time. The evaluation function is represented as

$$Q(s,a) = E\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right\} \qquad (4)$$

where *E{.}* denotes for the expectation operator and $0 \le \gamma < 1$ is a discounted factor which represents the importance of reward value in the future. The final goal of QAZP is to obtain an optimal policy: According to current state, the QAZP agent can choose an optimal action *a** which maximizes the evaluation function. In this Reallocation of the MA problem, the maximization of evaluation function also represents that minimization of energy consumption's variance among hotspots.

Next, the evaluation function can be expanded below,

$$Q(s,a)=E\left\{\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)\mid s_0=s,a_0=a\right\} \tag{5}$$

$$Q(s,a)=E\{r(s_0,a_0)\mid s_0=s,a_0=a\}+E\left\{\sum_{t=1}^{\infty}\gamma^{t}r(s_t,a_t)\mid s_0=s,a_0=a\right\} \tag{6}$$

$$Q(s,a)=E\{r(s,a)\}+\gamma\sum_{s'}P(s,a,s')\times E\left\{\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)\mid s_1=s',a_1=a'\right\} \tag{7}$$

$$Q(s,a)=r(s,a)+\gamma\sum_{s'}P(s,a,s')\times Q(s',a') \tag{8}$$

where *p(s,a,s')* denotes that taking action *a* in state *s* will transform into state *s'* with *p* probability. The result of this expansion means that the evaluation function of the current state-action pair can be represented as the sum of the current state-action's immediate reward and the expected value of evaluation function for all the possible next state-action pairs.

There is a standard equation to compute the optimal policy *π**. The equation is called Bellman Optimality Equation [9] which obtains the optimal action *a** by following two kinds of operations:

$$Q^*(s,a)=r(s,a)+\gamma\sum_{s'}\left\{P(s,a,s')\times\max_{a'\in A}Q^*(s',a')\right\} \tag{9}$$

The first operation is to find Q value in each *(s,a)* pair. The Q value is sum of immediate reward in *(s,a)*, and expected Q value for every possible next state-action pairs $Q_{t+1}(s,a)$ as follow:

$$\Delta Q_t(s,a)=\left\{r(s,a)+\gamma\max_{b}\left[Q_t(s',b)\right]\right\}-Q_t(s,a) \tag{10}$$

$$Q_{t+1}(s,a)=Q_t(s,a)+\alpha\Delta Q_t(s,a) \tag{11}$$

4 QAZP Scheme

The agent in MA learns Q-values through several interactions with environment. With the learned Q-values being stored in Q-table, the task of network partition is carried out by using the learned Q-values. Then we elaborate the details of the learning agent implemented by QAZP scheme. Fig. 4. shows the structure of the QAZP learning agent.

4.1 State Construction

Construct the states s_t and s_{t-1} by acquiring the residual energy of hotspots around sinks. We assume that powerful sink can directly send residual energy notification packets to the MA through one-hop communication.

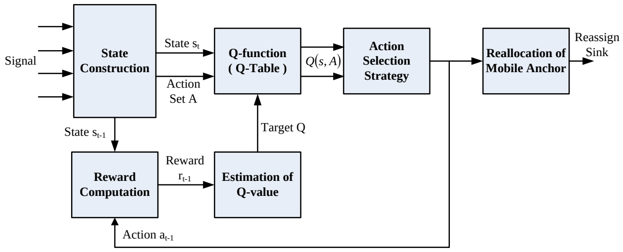


Fig. 4. Structure of Learning Agent

4.2 Reward Computation

The agent can detect the coordinates of the MA through GPS receiver. Based on this information, the state is constructed for Q-function block. Moreover, the residual energy of hotspots is used for reward calculation by Eq. (1). However, the reward calculated at current decision point is assigned to state-action pair at previous decision point.

4.3 Q-function

With the input of quantitative states which are fed into the Q-function called Q-table, the agent can obtain Q-values for all possible state-action pairs. Based on these Q-values, agent will decide a moving action.

4.4 Action Selection Strategy

Q values for possible state-action pairs can be obtained by executing the Q-function. However, if the Q values have not convergence, it means that the agent is still in learning procedure. In this situation, if the agent always chooses the best action, it probably will result in a local optimal action. For this reason, there is a trade-off between exploitation and exploration: To learn an optimal policy, the agent should try a number of wrong decisions in early learning to improve overall performance. However, after a long time, the policy will approach to a near-optimal policy. It means that the agent should not spend cost and time in training. There is a well-known strategy called *ε-greedy* [7] to resolve this trade-off between exploitation and exploration. In this paper we adopt the *ε-greedy* strategy to our action selection strategy block, in which the agent chooses the current optimal action with $1-\varepsilon$ probability, on the other hand, it learns the others action with ε probability where $0\le\varepsilon\le1$. Besides, the ε value will decrease over time, it means that the agent will explore as far as possible in the beginning of early learning. At the end of learning, agent will exploit the learned knowledge to execute the optimal action with high probability.

4.5 Reallocation of the MA

The agent guilds the MA in which direction to move. If the MA is reallocated, it rebroadcasts the sink assignment packet and repartitions the Multi-Sink WSNs into number of zones which equals to the number of sinks.

4.6 Estimation of Q-value

Based on these information: state s_{t-1}, selected action a_{t-1}, current state s_t, and reward r_{t-1}, we update the Q-values by Eq. (8). In any learning epoch, since the only one action is chosen for current state, it means that the Q-value of the chosen action pair is updated, while others remain.

Then detailed procedure is described as Table 1.

Table 1. Pseudo Code of QAZP

```
Program start
    Set the MA's location and spreading angle
    All sensors follow nearest-sink strategy temporarily
    Initialize $\alpha$ ,$\varepsilon$  and $Q(s,a)$
    Initialize s, $s_{t-1}$, a, $a_{t-1}$
    While ( all hotspot is alive and decision-making) do
        The agent collects system state from sinks
        $S_{t-1}$=s, $a_{t-1}$=a
        Select an action $a$ according to $\varepsilon$ -greedy strategy
        The agent obtain reward $r_{t-1}$
        Update $Q(s_{t-1},a_{t-1})$ depend on $s_{t-1}$, $a_{t-1}$, $s$ and $r_{t-1}$
        The selected action $a$ trigger anchor's movement
        If( $a$ == one of moving direction)
            Anchor is reallocated to a new position
            Rebroadcast sink-assignment packet
        Sensor choose a routing path to forward data
End
```

5 Performance Evaluation

The effectiveness of our proposed scheme called QAZP is validated through simulation. This section describes simulation environment, performance metrics and simulation results. The results are compared with performance of the Nearest-sink (NS) and Round-Robin (RR) [6] routing algorithms.

We implement our proposed scheme by C++ programming language. We partition simulation time into numbers of slots, the decision-making agent choose action at the beginning of each slot. The agent periodically collects system state from each sink, and Q-values are stored in table. As the simulation goes on, the Q-values are improved by employing expression (10) (11) and approach their true value. Parameter α is initialized to 1 and ε is initialized to 0.9, and they are linearly decremented until they reach 0 at the end of learning.

The simulation environment is set in Table 2. There are 600 sensors uniformly deployed in a square-shaped 100x 100 area. The communication range of sensor is set to ten units. At each round, sensors transmit one packet to sink with specific probability depended on different concentration model: *Linear* and *Complicated* concentration model, which are proposed by [10]. Sensors in hot area generate 1 packet per round with 1.0 or 0.6 probabilities, while others with 0.3 probabilities.

The performance evaluation of three strategies is based on two concentration model representing various data generation rate. We will show that our proposed scheme is able to adapt to kinds of concentration environment.

Table 2. Simulation Parameters

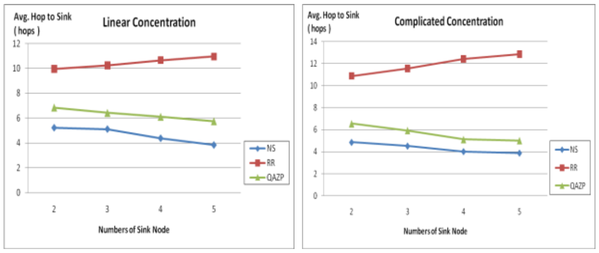
Number of sensors	600 to 800
Sensing area	100 x 100
Number of sinks	2,3,4,5
Initial energy of sensor nodes	5000 units
Energy consumption for transmission	1 unit per packets
Packet generation rate	Linear / Complicated concentration model
Communication range	10 units distance
Decision-Making Cycle	10 rounds
Moving distance per decision-making	10, $10\sqrt{2}$ units distance

Use the following metrics for comparing the performance of NS, RR and QAZP:

1) Average hop count: The average hop count from source to sink represents that the number of forwarding times for packets. The higher it is, the larger the aggregate energy consumption. Besides, the long distance also means high data delay time.

2) Network lifetime: The lifetime is defined from the deployment to the instant when the first hotspot exhausts battery. This is a good indicator for the expected lifetime of network as it shows how well the load balancing scheme avoid EISS problem.

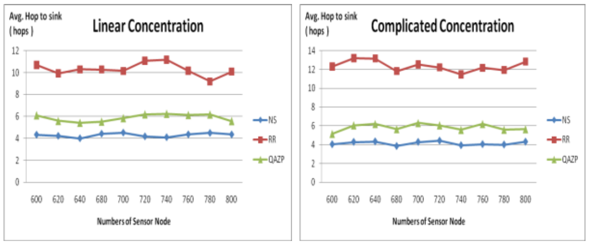
Fig. 5. shows the average hop count vs. numbers of sink node for different schemes. The number of sensor node is 600. The average hop count to sink of NS scheme is about four hops, it is the shortest among three schemes because sensors always choose the nearest sink; the average hop count to sink of RR scheme is about ten hops, it is the longest obviously because the round-robin way lead to that many sensors near a sink often choose another sinks which is much far away from them; the average hop count to sink of QAZP scheme is about 6 hops. The average hop count of QAZP scheme is between NS and RR. It means that QAZP scheme can balance load with little effect on hop distance. Sensors can choose sink which is as close as possible to them. Besides, we can observe the average hop count of RR scheme increase slightly as the numbers of sink node. It means that the number of sink node of RR scheme may be restricted; therefore RR scheme may be not suitable in more sink nodes' environment.



(a) Linear model (b) Complicated model

Fig. 5. Average Hop Count vs. Numbers of Sink Nodes

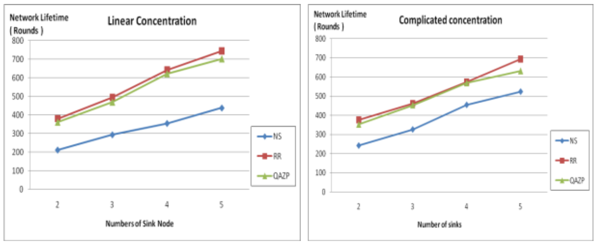
Fig. 6. shows the average hop count vs. numbers of sensor node for different schemes. The number of sink node is 4. We observe the average hop count varies inconspicuously because the sensing area is the same and the length of routing path is changeless.



(a) Linear model (b) Complicated model

Fig. 6. Average Hop Count vs. Numbers of Sensor Nodes

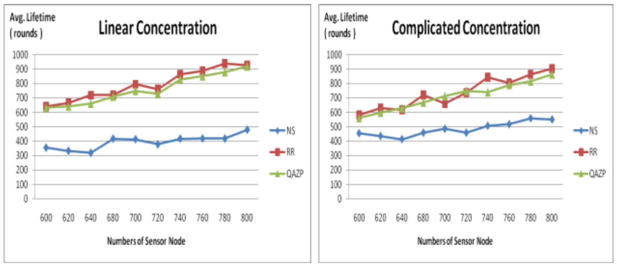
Fig. 7. shows the average lifetime vs. numbers of sink node for different schemes. The number of sensor node is 600. NS scheme derives least lifetime in both linear and complicated concentration model because it selects nearest node to transmit sensing data. However, it exists longer lifetime in complicated concentration model because mass sensor nodes gather mass data in linear concentration model. RR scheme has maximum lifetime in linear model because all hotspots exhaust energy equally, but it is susceptible in complicated concentration model. The average lifetime of QAZP scheme is close to RR scheme but more stable than RR scheme in complicated concentration model. It means QAZP scheme can balance the load under different traffic pattern environment.



(a) Linear model (b) Complicated model

Fig. 7. Average Lifetime vs. Numbers of Sink Nodes

Fig. 8. shows the average lifetime vs. numbers of sensor node for different schemes. The number of sink node is 4. We observe that the average lifetime increases as the numbers of sensor node because the data can select more nodes to transmit and then the interval of transmits is extended. As mention above, we can also observe the average lifetime of RR scheme in complicated concentration is more instable than QAZP scheme. It shows that the heuristic algorithms like QAZP scheme are more adaptive than steady algorithms.



(a) Linear model (b) Complicated model

Fig. 8. Average Lifetime vs. Numbers of Sensor Nodes

6 Conclusions

In this paper, we propose a load balancing scheme called QAZP. To resolve the EISS problem in Multi-Sink WSNs, the centralized the MA adaptively partitions the network into numbers of zones according to the residual energy of hotspots around sinks. Besides, we apply Machine Learning to the MA for adapting to any traffic pattern. Sensors around sinks are defined as hotspots, and source nodes can choose different hotspots as their destination. The residual energy of hotspots and hop distance are used to choose the path for routing a packet. From performance evaluation, we show that the proposed QAZP scheme prolongs the WSNs' lifetime under asymmetrically data generation environment. Besides, data packet can achieve sink through shorter path than RR scheme, it means that overall energy consumption is lower than RR.

7 Acknowledgement

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 99-3113-P-006-002.

8 References

[1] Soyturk, M. and Altilar, T., ”A Novel Stateless Energy-Efficient Routing Algorithm for Large-Scale Wireless Sensor Networks with Multiple Sinks” in Proc. of the IEEE Annual Wireless and Microwave Technology Conference, 2006.

[2] Hyunyoung Lee, Klappenecker, A., Kyoungsook Lee, and Lan Lin ”Energy efficient data management for wireless sensor networks with data sink failure” in Proc. of the IEEE International Conference on Mobile Adhoc and Sensor Systems Conference, Nov. 2005.

[3] Yuichi Kiri, Masashi Sugano, Masayuki Murata, "Self-Organized Data-Gathering Scheme for Multi-Sink Sensor Networks Inspired by Swarm Intelligence," saso, pp.161-172, First International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2007), 2007.

[4] Min Meng, Xiaoling Wu, Hui Xu, Byeong-Soo Jeong, Sungyoung Lee, and Young-Koo Lee, “Energy efficient routing in multiple Sink sensor networks,” The fifth International Conference on Computational Science and its Applications, pp. 561-566, 2007.

[5] Yunyue Lin, Qishi Wu, ”Energy-Conserving Dynamic Routing in Multi-Sink Heterogeneous Sensor Networks.” 2010 International Conference on Communications and Mobile Computing (CMC), pp. 269-273, 2010.

[6] Chunping Wang, Wei Wu, "A Load-balance Routing Algorithm for Multi-sink Wireless Sensor Networks," in International Conference on Communication Software and Networks, 2009. ICCSN 2009.

[7] L. P. Kaelbling, M. L. Littman, and A. P. Moore, “Reinforcement learning: A survey,” Journal of Artificial Intelligence Research, vol. 4, pp. 237–285, 1996.

[8] C. J. C. H. Watkins, and P. Dayan, “Technical note: Q learning,” Machine Learning, vol. 8, no. 3, pp. 279-292, 1992.

[9] R. Bellman, Dynamic Programming. Princeton, NJ: Princeton Univ. Press, 1957.

[10] T. Suzuki, M. Bandai and T. Watanabe, “DispersiveCast: Dispersive Packets Transmission to Multiple sinks for Energy Saving in Sensor Networks,” in Proc. of Personal, Indoor and Mobile Radio Communications (PIMRC) 2006.

Beyond LTE: Next Generation Multiple Access Technology with intelligent lower layers

Rajarshi Sanyal **Ramjee Prasad** **Ernestina Cianca**

Brussels, Belgium
sanyalrajarshi@yahoo.com

Aalborg University, Denmark
prasad@kom.auc.dk

University of Rome, Italy
cianca@ing.uniroma2.it

Keywords : AFCRN;OFDMA;LTE;8PSK;QAM;Coodination Processor

Abstract: The idea of WWWW, World Wide Wireless Web, has started from 4G technologies. With all the evolving technologies, we aim to achieve higher data speed and bandwidth, low latency, enabling the interworking for next generation intelligent applications. While we aim to actuate higher data speeds in 4G , the overall mobility and location management processes remain the same as the previous generation. We believe however, that in order to realize smart networks, the network layers should be more intelligent. We aim to realise by rendering intelligence to layer 2 and to directly involve it in mobility and location management processes. Hence we make the application layer lighter and craft more room for the intelligence required for the next generation applications. The proposed smart network can serve as the technology of tomorrow for green-field mobile operators and for mobile networks with wide coverage area and serving large demography with high population density. In this paper, we strive to discuss the implementation proposed Multiple Access technology for Mobile Network and unfurl the processes leading to handover and mobility management without direct involvement of layer 7.

Introduction

'Simplicity is the ultimate sophistication'.
Leonardo da Vinci

In this paper we propose to focus on the Novel Multiple Access Technology and propose a global mobile network architecture. We describe the handover and mobility management processes related to this novel access technology. The paper is organized as follows: Section 2 overviews the addressing mechanism proposed in [1]; Section 3 presents the proposed global mobile network architectures, frame formation defining also the handover procedure. Section 4 discusses the error estimate and correction scheme for the addressing. In section 5, we draw the conclusions.

1. The addressing mechanism

The Multiple Access Schemes deployed by the different access technologies are agnostic to the identity of the users. The symbols in the complex plane depict some digital values (bits depending on the multiple access scheme) which are essentially used for conveying user information through the physical channel. The identification of the user in the network is solely dependent on the capabilities of the application layer.

With our proposed methodology, there is a fundamental shift from the basic access mechanism. Here, each user will be represented by a symbol pertaining to a phase angle on the complex plane which has a total of 8 symbols in the outer ring. So 8 symbols will represent 8 users in the complex plane..

We use circular 8PSK/ QPSK with two circles. The outer circle having 8 symbols on 8PSK and the inner circle having 4 symbols on QPSK There will NOT be a dedicated channel for carrying the signalling information. The main Channel is subdivided in multiple channels, each of which will have 8 symbols per frequency band for addressing the user and 4 symbols for user data. The symbols in each frequency will have a constant phase and amplitude (implies fixed coordinates) w.r.t the time advance.

The number of frequency band (ie, the carriers) will depend on the Erlang figures catered by the specific network.

Below is a synopsis of the evolution of the Multiple Access Schemes used by various technologies.

The constellation diagram derived from the various technologies is provided below. Starting from the basic GSM to HSDPA and LTE, we see that the fundamental principle of all these technologies at the lower layers remain the same. As the access technology evolves, it caters to more data speed by squeezing more symbols on the complex plane. Hence the distance between the symbols decrease and so are the chances of error. Hence newer methods of error corrections are required. The Layers 5 to 7 become more intelligent to cater to advanced Mobility Management , Location Management , High Bandwidth and presence related applications.

More intelligence in the application layer necessitates the devices and the network elements to be more complex, and hence increases the cost and the complexity of network processes, design and operations.

In the constellation diagrams in following Fig.1 for LTE , we see that the symbols are responsible for carrying user information . There is a fundamental and philosophical drift from the prevailing technologies, as we see in our proposed access scheme. Here a certain category of the symbols residing at the outer ring (propagated at very low frequency, proportional to the call attempt rate of a mobile network), are directly employed for identifying / addressing the user in the network plane. As the frequency is low, hence symbol error factor and loss is relatively decreased.

The other 4 symbols in the inner ring and propagated at the real data rate following QAM, the same followed today by various mobile technologies.

So with this mechanism, the same time slot is used for control and data, with different Multiple Access schemes, once by 8 PSK for addressing and then with 4 QAM for data traffic.

LTE (16 QAM)

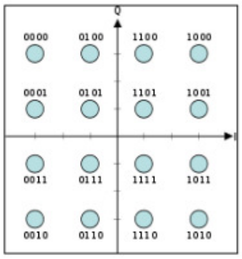


Fig 1. The Constellation diagram for LTE

Our Proposed Multiple Access Scheme

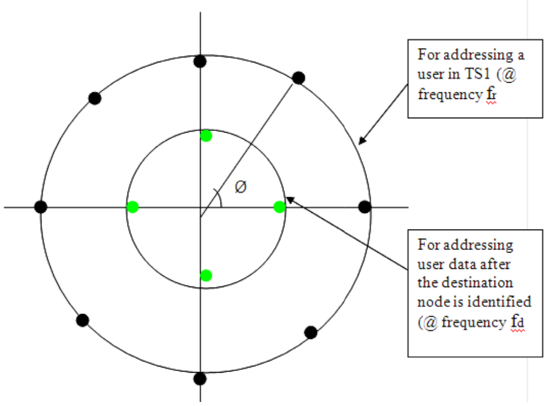

For addressing a user in TS1 (@ frequency fr
Ø
For addressing user data after the destination node is identified (@ frequency fd

Fig 2. Constellation Diagram for the proposed Novel Multiple Access Process

The proposed Multiple Access Methodology as in Fig 2., is a blend of 8 PSK (for outer ring) and QPSK (for inner ring) as the modulation technique and TDMA/FDD as the multiple access technique. the constellation is formed comprising the 8 symbols for user traffic in the outer ring and the 4 symbols in the inner ring dedicated for the purpose of addressing the users.

8-PSK used in 3G is used for conveying information from all users. In the proposed scheme, we allocate a specific symbol for a specific user of the channel by employing 8-PSK for the outer ring. The channels are also sliced up in time slots. So by virtue of this, the users are identifiable directly at the physical layer, if by some means we can render the intelligence to the user to clamp on to the specific time slot and to scan whether there are any incoming symbols located at the specific coordinate of the complex plane, which is hard coded for the user during the provisioning process.

Time slotting for addressing in network plane [ref1]

A user is identified in the network with respect to the symbol coordinate in the complex plane. Each user is allocated a specific Time Slot and a Frequency Sub band (AFCRN).

The E.164 address of a mobile station is linked to the Time slot, AFCRN and symbol coordinate allocated to the particular subscriber during provisioning. For a mobile roaming network , the address of the mobile node will be linked to these parameters in the Home Location Register. The number of time slots in the TDM frame and the scanning frequency primarily depends on the below factors.

1. The refresh frequency of a time slot and the number of time slots in a frame depends on the number of call attempts per second made in the network and the data rate offered per time slot when the time slot is used for traffic for a particular user.

2. It also depends on the max number of users that the network can address at a given point of time.

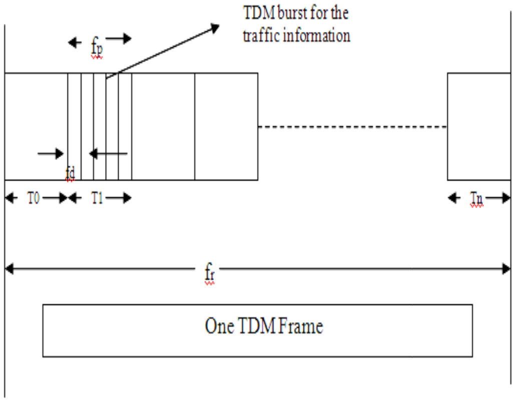


Figure 3 : The time frame structure

Let there be T0 to Tn time slots per frame
Let the call attempts per second in the given network = X
The data rate required for the traffic burst = fd
Refresh frequency for the given frame= fr
fd >> fp » fr
Now T0 to Tn makes one frame
Lets say there are ‘m’ frequency bands (AFCRN)
So the Total available time slots are = m x n
Say tr is the time required per frame

tr = 1/ fr

So m x n call attempts can be made in tr time period

Hence Call Attempts per second = m x n / tr

A = m x n x fr (A implies the permissible call attempts per second).

Data rate

The data rate will be bursty in nature, which means that the data for a given user will arrive at a frequency of fr and followed by a gap of time period

t g = tr x (1 – 1/n) = tr (n-1)/n. (where tg = time gap between 2 TDM bursts for a given time slot for data transfer)

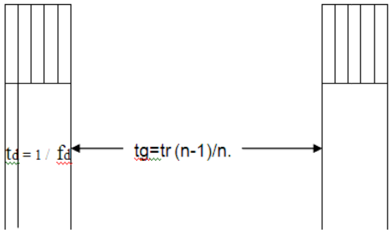


Figure 4 : Time Gap between the bursts

Relation between the CAPS and tg and td.

Let tg max be the max be the maximum time gap allowed between 2 consecutive TDM bursts for carrying the data for a given user.
Let there be k number of time slots for the data transfer within the time slot allocated for addressing
So tn = k x td
Now, tg=tr (n-1)/n
= tr [(A/m.fr) – 1] / (A/m.fr) = k x td [(A/m.fr) – 1] / (A/m.fr)
So, tg **max = k x td [(A/m-max.fr) – 1] / (A/m-max.fr)**

3. Architecture, Frame formation and addressing

Frame formation in the Downlink and Uplink

The Access points are also referred as Time Slot grabbers. In a location area, it receives the call attempts made by the Mobile stations in vicinity. This mean, it looks for information arriving at specific timeslots. The information is mainly the symbol depicting a called user. From the different users the timeslot obtained are aggregated by a Mux before the coordination processor.

If a timeslot is already busy in the pre call establishment activities, the Mux discards a second call request in the same timeslot towards the same or different user depending upon the symbol information.

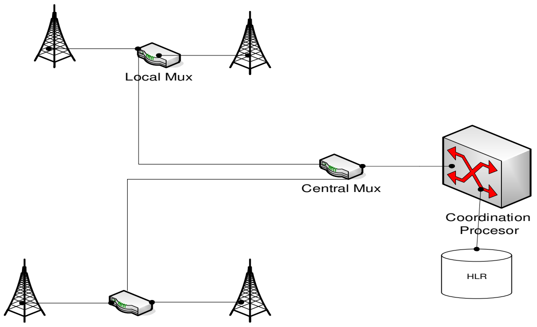


Figure 5 : The Multiplexer Unit in the Access Network

The Coordination processor replaces the conventional time/space switch of a mobile network. So at one end, it interfaces with the Aggregate Mux and access points both in the forward and reverse channels. On the other hand, it interacts with the core network elements.It sits between the access and the core network. It replaces the functionality of a modern day space and time switch of a mobile network. Fig 5 explains the high level functionality of the coordination processor.

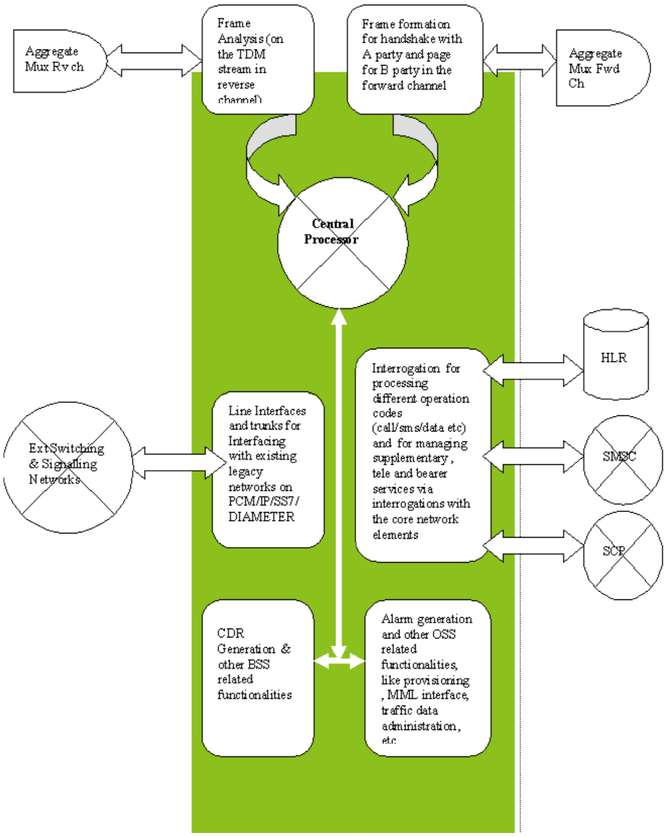

Aggregate Mux Rv ch
Frame Analysis (on the TDM stream in reverse channel)
Frame formation for handshake with A party and page for B party in the forward channel
Aggregate Mux Fwd Ch
Central Processor
HLR
Ext Switching & Signalling Networks
Line Interfaces and trunks for Interfacing with existing legacy networks on PCM/IP/SS7/ DIAMETER
Interrogation for processing different operation codes (call/sms/data etc) and for managing supplementary , tele and bearer services via interrogations with the core network elements
SMSC
SCP
CDR Generation & other BSS related functionalities
Alarm generation and other OSS related functionalities, like provisioning , MML interface, traffic data administration , etc

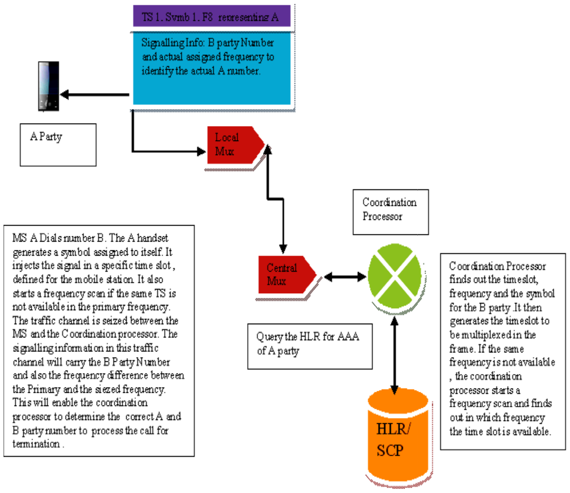
Fig 6. High level functionality of Coordination Processor

As in Fig.5 , there can be several MUXes before the coordination processor. But just before the coordination processor, there is an intelligent central mux which aggregates all the time slots, and also understands the information to reject based on the circumstances detailed below.

Short description of the frame formation in the uplink and downlink and the interaction between the Mobile Station and the Coordination Processor.

Fig 7. Illustrates how the frame is formed in the uplink (Call Initiation and response)

Figure 7 :Frame Formation in the Uplink


TS 1. Svmb 1. F8 representing A
A Party
Local Mux
Coordination Processor
Central Mux
Query the HLR for AAA of A party
HLR/ SCP

Signalling Info: B party Number and actual assigned frequency to identify the actual A number.

MS A Dials number B. The A handset generates a symbol assigned to itself. It injects the signal in a specific time slot , defined for the mobile station. It also starts a frequency scan if the same TS is not available in the primary frequency. The traffic channel is seized between the MS and the Coordination processor. The signalling information in this traffic channel will carry the B Party Number and also the frequency difference between the Primary and the siezed frequency. This will enable the coordination processor to determine the correct A and B party number to process the call for termination .

Coordination Processor finds out the timeslot, frequency and the symbol for the B party .It then generates the timeslot to be multiplexed in the frame. If the same frequency is not available , the coordination processor starts a frequency scan and finds out in which frequency the time slot is available.

Uplink

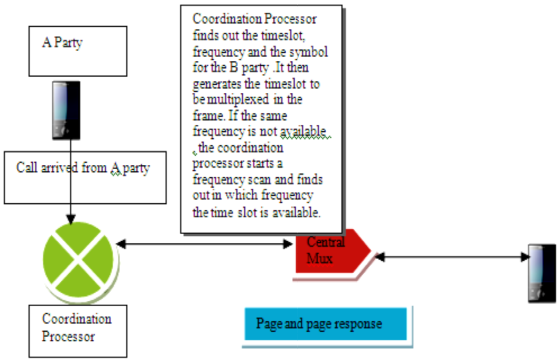
1. When the MS is provisioned, it is assigned a symbol, Time Slot and a primary frequency. All these 3 parameters are known to the Mobile station. Hence whenever the mobile station attempts to place a call, it tries with the allocated primary frequency first. If it is not available, because some other user assigned with the same time slot is already using the frequency, then the handset starts a frequency scan with the scanning algorithms [1].

2. The mobile station then includes the specific symbol value in the given Time Slot and forms a frame using the Time slot information. It synchronises with the access point and synchronises with the Frame before inputting the Time Slot.

3. The frames pertaining to all the frequency channels are aggregated by the Central Mux and finally fed to the Coordination processor.

4. If the page towards the MS B and page response is successful towards the B party, the A party receives a response, after which the MS of A Party seizes the time slot to convey the traffic information.

Fig 8. Below illustrates how the frame is formed in the Downlink (page and page response)



Coordination Processor finds out the timeslot, frequency and the symbol for the B party .It then generates the timeslot to be multiplexed in the frame. If the same frequency is not available , the coordination processor starts a frequency scan and finds out in which frequency the time slot is available.

Figure 8 : Frame Formation in the Downlink

Downlink

- In case of the downlink, the coordination processor determines the parameters of the B party after carrying out the AAA. The parameters are mainly the symbol number, Time slot and frequency number which is provisioned in the HLR.

- The coordination processor handles all the calls, so it keeps a track of all the active outgoing calls and the time slots and frequency in use

- So for placing a call towards a particular B party, it first starts a frequency scan to check which frequency has the Time slot free.

- When that time slot is found to be free in a particular frequency, the coordination processor multiplexes the Time slot in the frame with the specific symbol and transmits the same to the access point, through a central distributer.

- The coordination processor then awaits a response message from the B party. This implies that it expects the same symbol (of the B party) to come back in the same Time slot and the same frequency.

- On the receiver end, the Mobile station always listens to the frames of different frequency channels When it finds a symbol allocated to itself in a given time slot, it responds back in the reverse direction by injecting the same symbol (assigned to itself) , in the specific time slot and the same frequency. If the same frequency is not available in that instant of time, the call will not mature.

- After the Coordination processor receives the information back, it understands that it is a response message from the B party. It distinguishes between a response message and a new call request, as it maintains the call context immediately after generating a page signal in the downlink and starting a timer to await exactly the same information as generated, but in a latter time frame to come back.

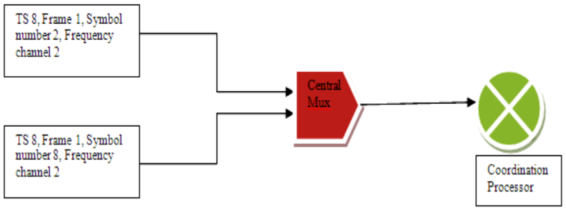

TS 8, Frame 1, Symbol number 2, Frequency channel 2
TS 8, Frame 1, Symbol number 8, Frequency channel 2
Central Mux
Coordination Processor

Fig 9 : Conditions for symbol rejection in the time slot for the downlink.

As illustrated in Fig 9, it may happen that 2 MSs initiate the call towards two different called numbers (B Party), who share the same Time Slot in the same frequency sub band. Both the calls should mature. In this scenario, the Coordination Processor determines randomly to respond to one call initiation in the same Frequency band. For the other mobile station, it responds at a different free frequency band by multiplexing the same symbol (of A Party) in the same time slot. So as a general principle, the mobile stations listen to time slots (in the downlink) which holds the same coordinate as allocated to itself.

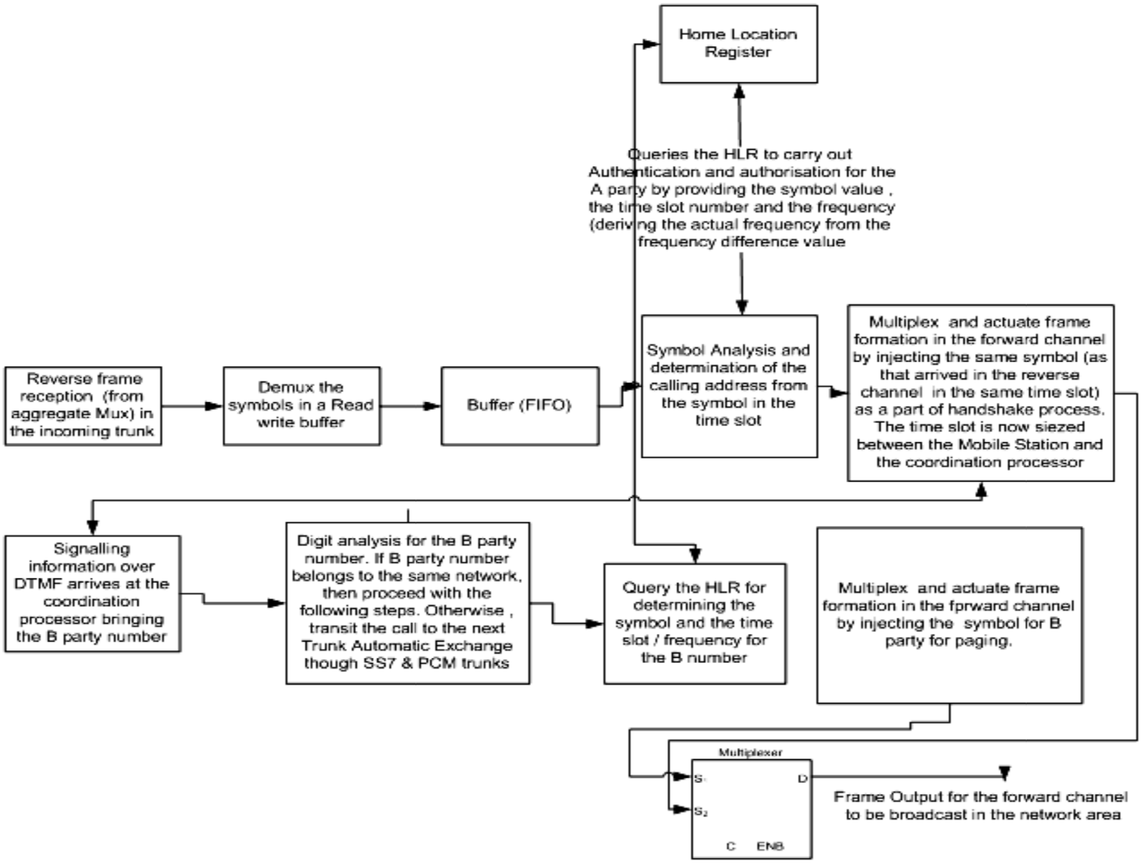

Home Location Register
Queries the HLR to carry out Authentication and authorisation for the A party by providing the symbol value , the time slot number and the frequency (deriving the actual frequency from the frequency difference value
Reverse frame reception (from aggregate Mux) in the incoming trunk
Demux the symbols in a Read write buffer
Buffer (FIFO)
Symbol Analysis and determination of the calling address from the symbol in the time slot
Multiplex and actuate frame formation in the forward channel by injecting the same symbol (as that arrived in the reverse channel in the same time slot) as a part of handshake process. The time slot is now siezed between the Mobile Station and the coordination processor
Signalling information over DTMF arrives at the coordination processor bringing the B party number
Digit analysis for the B party number. If B party number belongs to the same network, then proceed with the following steps. Otherwise , transit the call to the next Trunk Automatic Exchange though SS7 & PCM trunks
Query the HLR for determining the symbol and the time slot / frequency for the B number
Multiplex and actuate frame formation in the fprward channel by injecting the symbol for B party for paging.
Multiplexer
S
S2
D
C
ENB
Frame Output for the forward channel to be broadcast in the network area

Fig 10. Frame formation aided by Coordination Processor

Mobility and Location Management

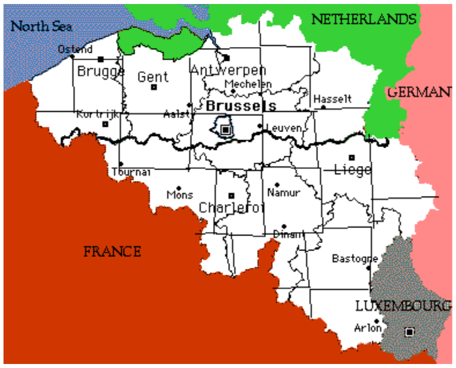


Fig 11. Network Area of Belgium segmentised in large sized square shaped cells

1. High Level description

For a global sized network, reusing the frequency is mandatory for a higher spectral efficiency. Frequency reuse also implies that we have consider an important called handover
In a present GSM network, we can say that more than 10% of bandwidth resource is consumed due to signalling. Also the complexity in the application level processes also need to be considered.

We thrive to handle frequency allocation and reuse though a process which is more dependent on the intelligence in the handset and then the physical layer.

Proposal

The geographical area is distributed in multiple cells. The cells are square shaped (Fig. 11) ,each having a diagonal distance from the centre of around 35 Kilometers . The 35 Kms restriction is due to retaining the timing advance value of max 63. As timing advance changes every 550 metres , hence the max cell size is 550 x 63 = 35 Kms.

There is a central Access point with a coordination processor in each cell.

Or there may be a central system with virtual coordination processors taking care of each cell.

The entire network area can be represented as chess board. Each block represents a square shaped cell. The cells do not have any cell ID. So the network is agnostic to the exact location of the mobile station. The white blocks are allocated AFCRNs from 1 to 64, while the black ones are allocated AFCRNs from 65 to 124.

Each mobile station is fitted with 2 transreceivers. There is an active part, which is active in cell which serves the mobile station.. The dormant part keeps on scanning the adjacent cells and is mainly responsible for actuating the handover procedure. It only scans the available timeslots (in its domain) in the subsequent frequency range. Hence while the active one operates in the AFCRN domain 1 to 64, the dormant transreceiver operates in the range 65-124.

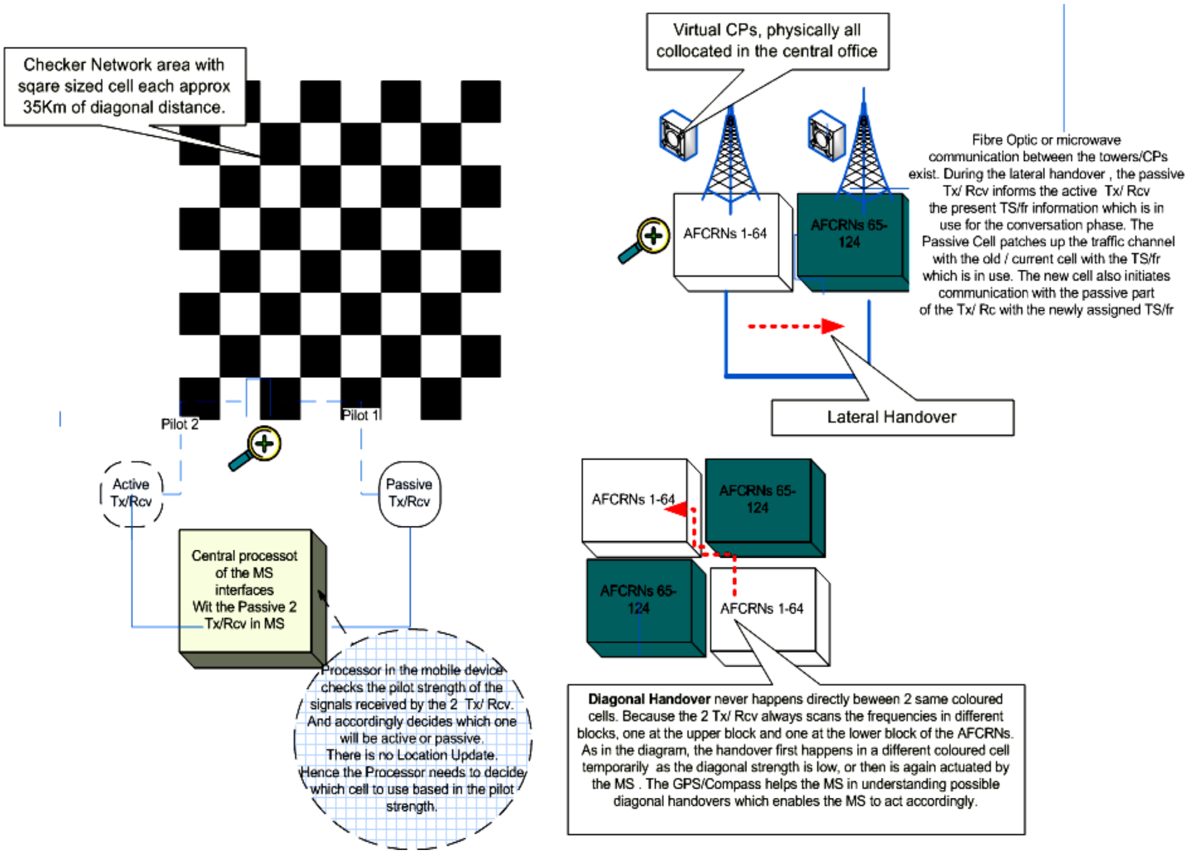
A central processor in the Mobile Station monitors and compares the signal strength as perceived by the 2 transreceivers. The MS also has a inbuilt GPS unit, and if in open air (with LOS with the satellite) can also determine the direction and speed of motion. It also has an electronic compass to analyse the direction and interpret the direction of motion if the LOS with the satellite is not available. Analysing the signal strength of the 2 received signal (and also gathering intelligence from the GPS / compass unit if available), the handset decides to initiate the handover,

In such a case, the passive Tr/Rcv informs the time slot and the frequency which is in use by the active Tr/Rcv. The coordination processor serving the cell establishes a channel (TS) , the same value which was intimated by the passive Tr/Rcv. This happens exactly when the active Tr/Rcv releases the channel so that the CPs of the adjacent white and the black boxes can establish channel on the same TS / Freq which is currently in use between the CP of the white Box and the MS.

Now the Physical path of the call is the Passive Tr/Rcv of the MS ----to--- CP/Access point of Black cell ------to -------CP /Access point of the white cell which was already in conversation phase directly with the MS before the handover.

After the handover, the passive TR/RCX in the MS becomes the primary one, …….and the TR/RCX which was primary before becomes dormant.

The process of handover is described in details in the following diagrams.



Checker Network area with sqare sized cell each approx 35Km of diagonal distance.

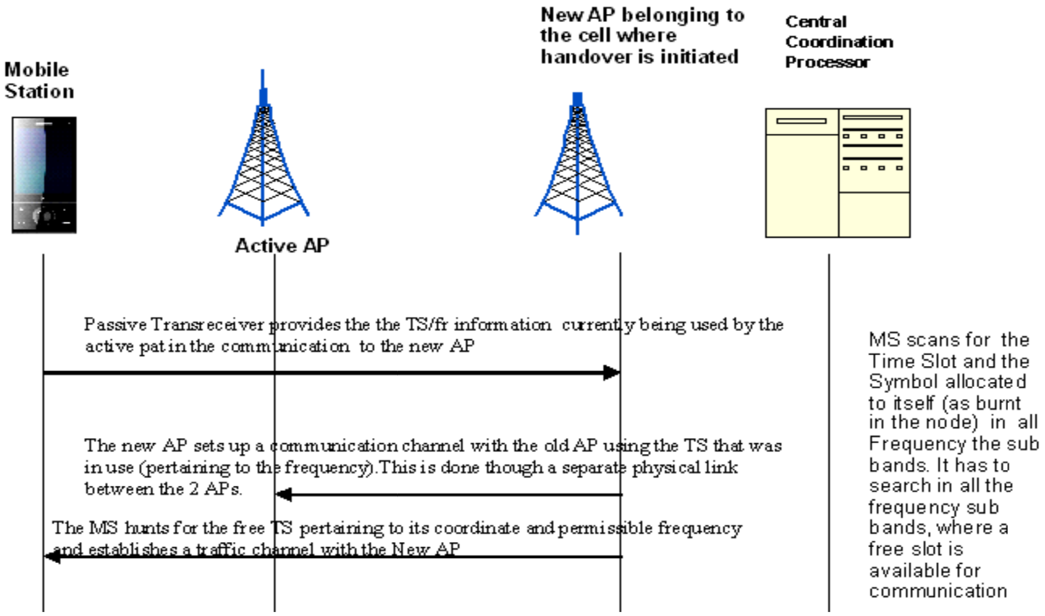
Virtual CPs, physically all collocated in the central office

Fibre Optic or microwave communication between the towers/CPs exist. During the lateral handover , the passive Tx/ Rcv informs the active Tx/ Rcv the present TS/fr information which is in use for the conversation phase. The Passive Cell patches up the traffic channel with the old / current cell with the TS/fr which is in use. The new cell also initiates communication with the passive part of the Tx/ Rc with the newly assigned TS/fr

Processor in the mobile device checks the pilot strength of the signals received by the 2 Tx/ Rcv. And accordingly decides which one will be active or passive. There is no Location Update. Hence the Processor needs to decide which cell to use based in the pilot strength.

Diagonal Handover never happens directly beween 2 same coloured cells. Because the 2 Tx/ Rcv always scans the frequencies in different blocks, one at the upper block and one at the lower block of the AFCRNs. As in the diagram, the handover first happens in a different coloured cell temporarily as the diagonal strength is low, or then is again actuated by the MS . The GPS/Compass helps the MS in understanding possible diagonal handovers which enables the MS to act accordingly.

Fig 12 : Handover Initiation by the Mobile Station


Mobile Station
Active AP
New AP belonging to the cell where handover is initiated
Central Coordination Processor

Passive Transreceiver provides the the TS/fr information currently being used by the active pat in the communication to the new AP

The new AP sets up a communication channel with the old AP using the TS that was in use (pertaining to the frequency).This is done though a separate physical link between the 2 APs.

The MS hunts for the free TS pertaining to its coordinate and permissible frequency and establishes a traffic channel with the New AP

MS scans for the Time Slot and the Symbol allocated to itself (as burnt in the node) in all Frequency the sub bands. It has to search in all the frequency sub bands, where a free slot is available for communication

Fig 13 : Call flow – Inter cell Handover

4. Error Estimate and Correction

Error Correction for the addressing symbols due to Channel Impairment

In our proposed technology, the addressing part is directly handled by the lower layers, namely layer 2 of the OSI stack. The errors in the physical channel can distort the coordinate of the symbols in the complex plane, leading to wrong addressing. Hence any error at access channel implies that the addressing related functionality is impacted, which means the rate of success of any call events like call maturation, SMS, data session setup etc is jeopardised. So it is extremely important to ensure that we employ an appropriate and a robust error correction mechanism to ward off this kind of situation, which can in effect endanger the practicality of such a Multiple Access Technology.

For the addressing part, there are a total of 8 symbols per time slot in the outer circle. When a transmitter (either a MS or AP) transmits a symbol to the receiver it will transmit odd number of times the same symbol. The frequency of transmission of this symbol is fp which is determined according to the Call Attempt per second in a mobile network. Hence the frequency is low (compared to the frequency of traffic data propagation) and hence is less prone to error. We propose a 2 step process to the error correction mechanism. The process starts with a repetitive code as explained. In such codes each bit of the message is repeatedly sent an odd number of times so that a simple majority decode rule can be followed to recover the original message. For example, the repetition code may send each symbol three times and then decide to decode and determine the given symbol. So in general, if each bit is repeated 2*n + 1 times then the code can tolerate up to n errors. Repetition codes are not very efficient as they increase the size of a given message by a factor of 3 or 5 say, depending on the repetition parameter. The redundancy added by a repetition code is just additional copies of the message itself, whereas more sophisticated codes add a smaller more intelligent amount of redundancy that can specifically pin point errors and correct them. More compact and intelligent forms of redundancy come at a cost of additional message processing for both the sender and the receiver. A definite advantage of repetition codes is simplicity both in encoding, and particularly decoding.So during any call processing when there is an attempt by the network to reach the mobile node, the network will transmit the same symbol pertaining to the specific coordinate in the complex plane and the fixed time slot (allocated for the target mobile node) .The receiver(mobile node) will calculate the error between the symbols by calculating the Euclidean distance between the reference symbol and the received one. If the mean error is less than Δe (error) , then it will respond back to the network.So if the error is less than Δe for each incoming code, the target mobile node responds back with the same symbol repeated the same number of times.

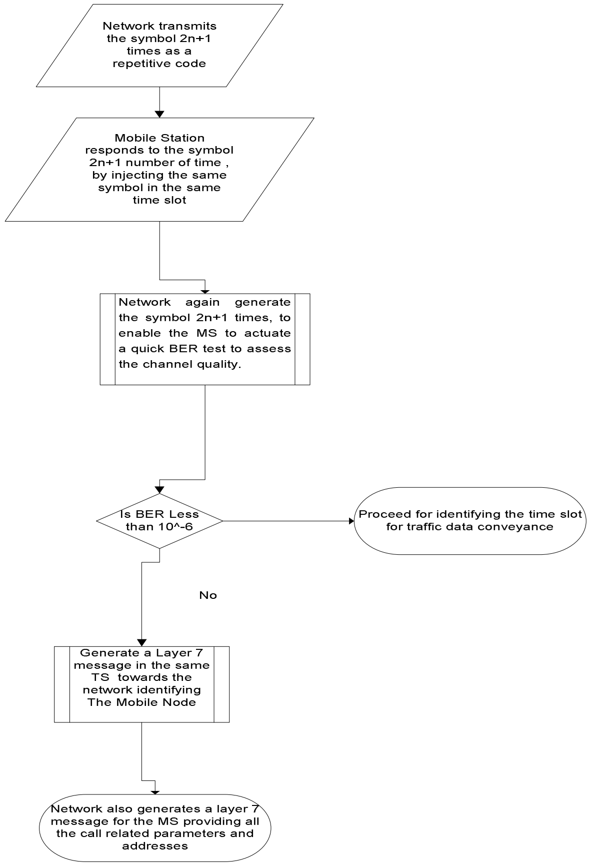

Network transmits the symbol 2n+1 times as a repetitive code
Mobile Station responds to the symbol 2n+1 number of time , by injecting the same symbol in the same time slot
Network again generate the symbol 2n+1 times, to enable the MS to actuate a quick BER test to assess the channel quality.
Is BER Less than 10^-6
Proceed for identifying the time slot for traffic data conveyance
No
Generate a Layer 7 message in the same TS towards the network identifying The Mobile Node
Network also generates a layer 7 message for the MS providing all the call related parameters and addresses

Fig 14 : Error correction for addressing

Now after the mobile node completes responding back with the same symbol 2n+1 times ,the Network again generate the symbol 2n+1 times, to enable the MS to actuate a quick BER test to assess the channel quality.If the BER is 10 ^-6 to 10^-10, then there is no further coding action is required and the network / Mobile node can conclude that the identification process is completed. The Mobile Node and the Network registers this process as completed for the given activity and related to the particular mobile node. However if the BER is more that 10^-6, then the Mobile Node generates a layer 7 Identification message with the IMSI parameter written. It is transmitted in the same time slot towards the network. The network has to respond to this message with the information regarding the call request (with the appropriate parameters). This response message shall be accepted and processed by the Mobile Node. The mobile node and the network thus reserves the time slot for the transition towards a traffic channel. The algorithm that will be used is in Fig 14..

In the following section, we try to understand the symbol error on 8PSK for a SNR of 15 d B. As the data rate for addressing is related to Call Attempts , it is much lower than the actual traffic late. This should keep the error rate comparatively lower.

Calculation for Symbol Error (Ps) and Bit Error (Pb) for 8 PSK with SNR = 15 dB for addressing a mobile node

We denote the coefficients *{sij}* as a vector **s***i* = (*si*1, . . . , *siN*) which is called the **signal constellation point** corresponding to the signal
si(*t*). The **signal constellation** consists of all constellation points *{***s**1, . . . , **s***M}*

The signal constellation for MPSK has
$si1 = A\cos[\ 2\pi(i-1)M\ ]$
$si2 = A\sin[\ 2\pi(i-1)M\ ]$
M] for i = 1, . . .,M.

The symbol energy is $Es = A$^2, where A is the amplitude

Hence SNR for symbol is γs = Es/No= *A*^2/No …..(No=> Noise)
γb ≈ γs / log2M

Let γb = 15 Db, where , γb = SNR for bits.

For the addressing synbols of the outer ring of the constellation diagram , we follow 8PSK.

Hence for 8PSK, γs = (log2^8) · 10^(15/10) = 94.87.
Now , Probaility of symbol error,
$Ps \approx 2Q(\sqrt{2}A/No \times \sin(\pi/M)) = 2Q(2\gamma s \sin(\pi/M))$

Where where the *Q* function, *Q*(*z*), is defined as the probability that a Gaussian random variable *x* with mean 0 and variance 1 is bigger than *z*:

$$Q(z) = p(x > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Substituting this into above equation yields
$Ps \approx 2Q\sqrt{}189.74 \sin(\pi/8)$= **1.355 · 10^−7**.
and using (6.3) we get *Pb* = *Ps*/3 = **4.52 · 10^−8.**

Conclusion

We believe that the future networks should be designed such that most of the prosaic functionalities of a mobile networks be shifted to the lower layers. The higher layers may take part in more intelligence activities to realise the next generation intelligent services which will require considerable processing power.

This will make it possible to realise a more optimised network , lesser resources to operate and to cater the expansion of the subscriber base. The overhead of signalling mainly in the Access domain will be reduced substantially resulting in more efficiently managed Mobile Stations requiring less battery power. The Network Elements constituting a network should also be engaged in lesser application layer activities and hence less requiring less processing power eventually lessening the network set up cost. The authors are engaged in this long drawn research effort to realise such future generation networks.

Reference

[1] Rajarshi Sanyal , Ernestina Cianca, and Ramjee Prasad , "Rendering Intelligence at Physical Layer for Smart presented at IARIA conference on wireless networks 2010 and published at IEEE digital library. http://www.computer.org/portal/web/csdl/doi/10.1109/ICN.2010.16

[2] Rajarshi Sanyal , Ernestina Cianca, and Ramjee Prasad , "Colour Pixel Multiple Access : The Multiple Access Technology for future generation wireless networks," International Conference on Wireless Communications , 2009 – World Congress Computer Science, Computer Engineering and Applied Computing, 2009 (http://www.world-academy-of-science.org), Las Vegas, USA. pp. 2-4.

[3] ETSI/3GPP , Digital cellular telecommunications system (Phase 2+); Numbering, addressing and identification ,(3GPP TS 03.03 version 7.8.0 Release 1998).

Analysis of Handover over WLAN for VoIP

Angela Chacón, and Edward Paul Guillen
Telecommunications Engineering Department, Military University Nueva Granada, Bogota, Colombia
gissic@umng.edu.co, edward.guillen@unimilitar.edu.co

Abstract— *The growing use of smartphones has allowed users to handle multiple networks in a small equipment, but in VoIP services with QoS, the mobility is limited to the cell network operator although a WiFi service could offer that kind of service. The Handover process means real mobility for users on wireless LAN networks in an automatic method for association within a correlated wireless system. This paper shows a brief overview on handover proposed schemes and gives a description about QoS required for VoIP with the schemes that reduce the effect caused by the handover process in WLAN.*

Keywords **- Handover, Voice over IP, Wireless LAN, Quality of Service.**

I. Introduction

Wireless network technology based on the standard 802.11 has become a worldwide solution for solving communication needs in multiple applications. Some of the applications that have gained momentum are those related to the communications that not only transport LAN data, but also mainly IP-based voice.

One of the problems of real-time applications like VoIP over WLAN is the effects on service quality by the process of handover within and between WLAN subnets during a VoIP session.

The latency and the jitter are greatly impacted when the control of the mobile node is handed over from one access point (AP) over to another one. This poses a challenge to providing and preserving QoS for VoIP users in WLAN environments. [3]

In general, voice quality in WLANs has been analyzed usually either using cost functions [1], Handover based on physical layer [15][16][17], parameters of QoS based on E-Model [5][6], MADM methods [1][10], needless handover and Hysteresis [12], THOA Algorithm[14], scanning solutions[18][19], between other.

In this paper, we describe a research proposal with the effect of the mobility handover on VoIP communications. Our work focuses on the impact of handoff mechanism on the objective quality of the voice traffic. The remainder of the paper is organized as follows. In section II we show the system architecture for the Handover process. In section III, we describe the meaning of the handover process and its classification. In section IV we discuss the QoS requirements of VoIP. In section V we describe the proposal methods and algorithms of handover. Finally we conclude this paper in the section VI.

II. System Architecture

Nowadays there is the trend that devices are equipped with multiple network interfaces, and it is possible to make transfers in a transparent (seamless) among these interfaces. [1][13] Great efforts had been made in architecture design, standardization and coverage of access networks. Another important aspect is the development of sophisticated policies that help the Mobile Node and Mobile Node (MN) to decide when to perform handover between networks.

Handover refers when a Mobile Node is connected to a network access point and the communication is transferred to another one. The handoff can be classified by different properties or characteristics of the network as described below, but the most common factor is the network types involved. [1] According to the network types, the handover is classified as horizontal or vertical, as illustrated in Figure 1.

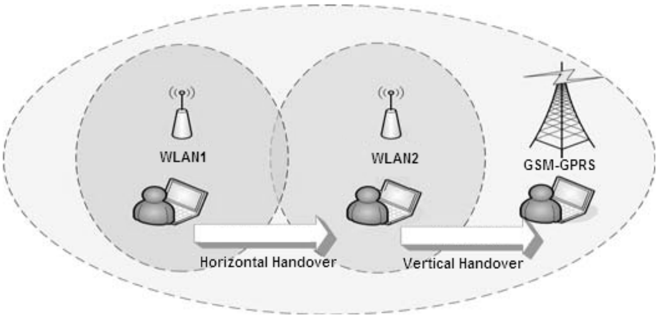


Fig. 1. System Architecture [1]

III. Handover and classification

Handover is the mechanism that allows users to move within a network or from a network to another one, without losing connectivity [1] [2]. The 802.11 standard does not define the handover process but it does describe basic processes as the reassociation. Reassociation occurs when the user is moving and passes from one AP to another within network. It is often not transparent for the user.

The Table I describes the classification of the Handover according to various factors, namely, number of connections, type of frequency, type of network, administrative domain, need for Handover and user control allocation.[1]

TABLE I
HANDOVER CLASSIFICATION

Factor	Classification
Type of Network	Horizontal Handover, Vertical Handover
Type Frequency	Inter-Frequency Handover, Intra-Frequency Handover.
Number of Connections	Hard Handover, Soft Handover, Softer Handover
Administrative Domain	Intra- Administrative Handover, Inter-Administrative Handover
Need for Handover	Voluntary Handover, Mandatory Handoff
User Control Allocation	Proactive Handover, Passive Handover

According with the last classification, the description of each of them is as follow:

A. Type of Network

The Horizontal Handover takes place when the transfer process is done within the same technology. On the other hand, in the Vertical Handover the transfer process takes place between different technologies.

B. Type of Frequency

The Inter-Frequency is the process of Handover that a mobile terminal done through the APs with different operational frequency. The Intra-Frequency is the done process with the same operational frequency.

C. Number of Connections

Hard Handover: The mobile terminal keeps a single link with the network removing the old link before to establish a new link.

Soft Handover: In order to eliminate the disconnection time, the soft Handover maintains at least two links to the network connection simultaneously.

Softer Handover: This type of connection also maintains two links, but the mobile terminal switches on radio links that belong to the same access point [3].

D. Administrative Domain

When the mobile terminal moves between different networks, supporting the same or different types of network interfaces, the domain could be Inter-administrative or Intra-administrative; the first one refers to networks managed by different administrative domains, the second case refers to networks managed by the same administrative domain.

E. Need for Handover

The mandatory Handover is necessary to avoid disconnections, the mobile terminal transfer the connection to another access point, for example, when the mobile terminal moves with a constant displacement.

F. User Control Allocation

Proactive Handover allows to the mobile terminal decide and configure the Handover process; the decision taken by the mobile terminal could be based over specific preferences of the user. The passive Handover is most common in 1G, 2G and 3G Wireless Systems. The user has no control over the process of Handover.

IV. QoS REQUIREMENTS

Real-time applications such as voice over IP are becoming a better available alternative than the traditional telephony. The implementation of such technology over wireless systems is a solution that has been widely studied, especially in terms of the quality of service.

The parameters of QoS as codec, capacity of the network, the delay and jitter are considered crucial to determine a good evaluation for the QoS of VoIP. Thus, Table II shows the requirements specified in the G.113 [4], that deals with the more important codecs for voice over IP and their corresponding bandwidth requirements.

TABLE II
CODEC REQUIREMENTS

Algorithm	Coding	Bandwidth (Kbps)
PCM	G.711	64
ACELP	G.723.1	5.3
MP-MLQ	G.723.1	6.3
ADPCM	G.726	32
LD-CELP	G.728	16
CS-ACELP	G.729	8

The delay and jitter are including in the WLAN when a mobile terminal is moving and request the process of Handover to another AP in the network, [3]

Abderrahmane propose [3] two issues relate to time delay and jitter: (1) Signaling for call set up, tear down and other call control communications will be delayed; (2) the jitter in the voice traffic/bearer channel will cause delay.

The voice delay requirements are showed in Table III, according with the recommendation G.113 [4].

TABLE III
DELAY REQUIREMENTS

Delay	Acceptable Quality
0 to 150 msec	Most Applications
150 to 400 msec	International Connections
> 400 msec	Public Network Operation

A. Handover over WLAN for VoIP

The Handover process over WLAN has two important descriptions: the vertical handover and horizontal handover, according with these mechanisms, the quality of service for voice traffic will be affected by the number of connections.

The vertical handover has the advantage of allowing the 'make before break' approach [3], which means that the mobile terminal maintains at least two links to the network connection simultaneously, making a seamless handover from one network to the other. On the other hand, the horizontal handover allows the 'make before break' approach. The mobile terminal keeps a single link with the network, removing the old link before establishing a new link, causing quality degradation with small interruptions in the communication.

In the experimentation done in [3], the setup consisted in a wireless scenario with two APs 802.11g and two Windows-based laptops equipped with 802.11g LAN card. One of these laptops was used as a server, and the other laptop as a client. The server generated LAN traffic to the client. This scenario is illustrated in Figure 2.

The study focuses on the impact of the Handover mechanisms over the quality of voice traffic. This research took into account the SNR (Signal-to-Noise Ratio), the throughput and the jitter included on the WLAN when the user is associated with another access point. It also performed the correlation between these parameters in three experiments.

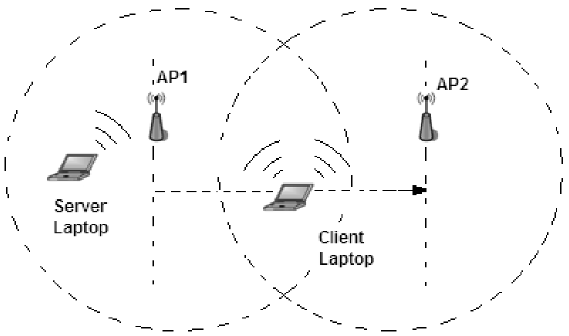


Fig. 2. Experiment Setup

- Experiment 1: Consists the effects of handover process on the variance of the SNR (Signal-to-Noise Ratio) during the transfer process. [3] The result shows that a SNR variation decreases when the mobile terminal moves away from the AP1 while it increases when the mobile terminal approaches to AP2.

- Experiment 2: Consists the effects of handover process on the throughput. The results indicate that the throughput dropped during transfer; this drop corresponds to a difference of 450 Kbps due the reduction of bandwidth related with the packet loss experienced. [3]

- Experiment 3: Consists the effects of the handover process on the variation of jitter and their relationship with the variation of throughput. [3] The results indicate that the jitter presents a large peak of 100ms while the handover occurs in a short period of 5ms, the throughput increases to 1Mbps.

B. QoS based on E-Model

The E-model is a computational model for estimating the subjective quality of a VoIP call [5]; it is standardized by the ITU-T as G.107 [6]. The voice quality can be estimated using this standard; it calculates a rating factor R that is an additive combination of five factors as follows:

$$R = R_0 - I_s - I_d - I_e + A \qquad (1)$$

Where:

R_0, Basic signal-to-noise ratio.

I_s, Impairments simultaneous to voice encoding.

I_d, Impairments due to network transmission.

I_e, Effects of equipment (e.g. low bit rate).

A, Advantage factor.

The congestion and signal quality degradation are the effects caused by the packet losses related with the handover process, the propagation and coverage.

It is possible with the E-Model to get the transmission rating factor R as a function of the packet loss for each of the voice codecs. Furthermore, the R factor represents the quality of the transmission, and can be converted into the more commonly known metric MOS (Mean Opinion Score), which comes from statistical surveys of quality graded from 1 to 5 as bad quality to the excellent quality [5].

For VoIP call, the G.107 [6] recommendation gives the following rates:

TABLE IV
G. 107 VOIP CALL RATING

R value (lower limit)	MOS (lower limit)	User Perception
90	4.34	Very Satisfied
80	4.03	Satisfied
70	3.6	Some users dissatisfied
60	3.1	Many users dissatisfied
50	2.58	Nearly all users dissatisfied

V. CRITERIA AND DECISION ALGORITHMS HANDOVER PROCESS

In order to ensure the properly QoS to a user that consumes a service provided by a network, it is important to provide and maintain an optimal bandwidth and low delay, but these parameters do not reflect the satisfaction user. [3][4] The satisfaction of the users is based to their own perceptions, thus, the satisfaction is subjective and making decisions are based on the level of satisfaction, [5] the best way to offer a service that meets the needs of the user consist to allow configure their own preferences.

Both, the vertical and horizontal handover process handled three phases: discovery, handover decision and execution.

In the discovery phase, the mobile phone equipped with various network interfaces determines which networks are available and the characteristics they have, such as bandwidth, SNR and jitter. In the decision phase, networks are evaluated according to their characteristics and select the most optimal. Quality and cost of service, security, power requirements, speed and proactive handover, are criteria to take into account when making decisions on networks. In the execution phase, the handover process is performed to transfer the link to the new base station and channel assignment. [1][5]

A. Cost Functions

The cost functions implements the policies model described by IETF [7] and it is widely used in heterogeneous networks to make the decision process of handover. The model consists in three entities, Figure 3: (PR, “Policy Repository”), the PR contains information such as user preferences, the strength of the signal and cost of availability of access networks.

The PR can obtain information through measurements of the environment and is responsible for delivering the policy settings applied to the entity making the decision based on the policies called the PDP (“Policy Decision Point”). A policy decision point PDP is the body control evaluates the access networks through a policy decision.

The policy decision made based on the parameters received from the PR. If the PDP decides that a handover should be done, it tells this to the PEP (“Policy Enforcement Point”) and it runs the Handover. The entity PEP ("Policy Enforcement Point") receives policy decision PDP and executes the handover transparent to the user [1].

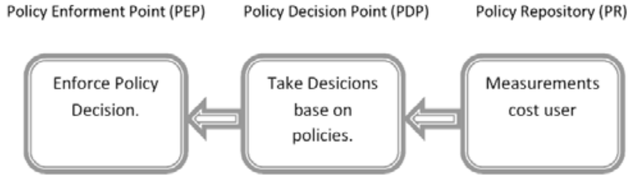

Policy Enforment Point (PEP)
Policy Decision Point (PDP)
Policy Repository (PR)
Enforce Policy Decision.
Take Desicions base on policies.
Measurements cost user

Fig. 3. Policies Model

The cost function defines rules for optimal handover decisions and it was first defined by Helen Wang in 1999 [8] and this proposal system was based in policies.

The cost function takes the user preferences and calculated a value for each network. The network interface with the lowest calculated value was considered the most optimal network to use.

The researches at the University of California at Los Angeles (UCLA) presented in 2004 another cost function [9]. This feature was implemented in a seamless handover architecture called Universal Seamless Handoff Architecture (USHA); this can be applied in a system of vertical handover, the network with the highest value is considered the best.

- *Absolute Cost Function.*

This cost function calculates an absolute scale value, the coefficients requires to be obtained from a tight-fitting function, i.e. through testing.

The absolute cost function S_i is showed in (2) [1][2]:

$$S_i = \sum_{j=1}^{k} w_j f_{ji} \qquad 0 < S_i < 1 \qquad \sum_{j=1}^{k} w_j = 1 \qquad (2)$$

Where,

k, number of weights.

f_{ji}, normalized functions for the parameters j and network i.

The parameters j can be the cost of using the (E), the binding capacity (C) or energy (P). If these parameters are used, replacing the cost function in (2) is obtained:

$$S_i = w_e f_{e,i} + w_c f_{c,i} + w_p f_{p,i} \qquad (3)[1][2]$$

In order to meet wit the introduction of the coefficients, the restrictions are showed as following:

$$\alpha_i \ge 0, M \ge \beta_i, \gamma_i \ge 0$$

Where, M is the maximum bandwidth demand required by the user. The values of the coefficients αi, βi and γi can be obtained through a lookup table or a special function and the normalized function for the parameters shown below:

$$f_{e,i} = \frac{1}{e^{\alpha_i}} \qquad f_{c,i} = \frac{e^{\beta_i}}{e^{M}} \qquad f_{p,i} = \frac{1}{e^{\gamma_i}} \qquad (4)[1][9]$$

- *Relative Cost Function.*

This cost function calculates a relative scale value and each parameter is calculated of the same way without any coefficient. If the parameter value has a positive effect like the bandwidth, the parameter provides the value $\ln(1/x)$, but if this parameter has a negative effect like the cost, the parameter provides the valued $\ln(x)$. The relative cost function is showed as follow:

$$f_i = w_b \ln \frac{1}{\beta_i} + w_p \ln P_i + w_c \ln C_i \qquad (5)\ [1][9]$$

f_i cannot handle scenarios where the cost for a network interface is zero because uses a natural logarithm and it is undefined for valued zero. S_i does support this. [1]

B. MADM (Multiple Attribute Decision Making) Methods

The MADM (Multiple Attribute Decision Making) methods are a group of four vertical handoff decision algorithms and they are the best known. [1] [10] MEW (Multiplicative Exponent Weighting), SAW (Simple

Additive Weighting), TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), and GRA (Grey Relational Analysis). All of the four algorithms allow different attributes such as bandwidth, delay, packet loss rate, and cost for the vertical handoff decision. [10]

The fuzzy MADM (Multiple Attribute Decision Making) is described in [10][11] as a method that consist in two steps: The first step is to convert the fuzzy data into a real number and the second step is to use classical MADM methods to determine the hierarchical order of the candidate networks. SAW (Simple Additive Weighting) and TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) are proposed as two classical MADM methods in [11].

- *SAW* (Simple Additive Weighting)

The overall score of a candidate network is determined by the weighted sum of all the attribute values. [10] The score of each candidate network i is obtained by:

$$A^*_{SAW} = \arg\max_{i\in M} \sum_{j=1}^{N} w_j r_{ij} \qquad (6)[10][11]$$

There is a sum of the normalized contributions from each metric r_{ij} multiplied by the importance weight assigned w_j of metric j. N is the number of parameters, and M denotes the number of candidate networks.

- *TOSIS (Technique for Order Preference by Similarity to Ideal Solution*

The selected candidate network is the one which is the closest to the ideal solution, while the farthest from the worst case solution. The ideal solution is obtained by using the best values for each metric; it is showed in (7).

$$A^*_{TOP} = \arg\max_{i\in M} c_i^* \qquad (7)\ [10][11]$$

Where, c_i^* denote the relative closeness of the candidate network i to the ideal solution.

- *GRA (Grey Relational Analysis)*

The selected network is based on Analytic Hierarchy Process (AHP) and Grey Relational Analysis (GRA). AHP decomposes the problem of selection of the network into several sub problems and each assigned a weight.
GRA is then used to rank the candidate networks and select the most important or better. A normalization process is required to negotiate with the cost-benefit metrics. [10]

$$A^*_{GRA} = \arg\max_{i\in M} \Gamma_{0,i} \qquad (8)\ [10][11]$$

Where, $\Gamma_{0,i}$ is the GRC (Grey Relational Coefficient) of network i [10]. The GRC is used to calculate the score or value of each network to describe the similarity between the network and the network ideal candidate, the selected network is more similar to the ideal.

- *MEW (Multiplicative Exponent Weighting)*

The network is defined as the network with the best values in each metric. For a benefit metric, the best value is the largest. For a cost metric, the best value is the lowest. [10][11] The value ratio R_i between network i and the positive ideal is calculated by:

$$R_i = \frac{\prod_{j=1}^{N} x_{ij}^{wj}}{\prod_{j=1}^{N} (x_{ij}^{**})^{wj}} \qquad (8)[10][11]$$

Where, $0 \le R_i \le 1, w_j$ is a positive power for benefit metrics x_{ij}^{wj} and it is convenient to compare each network with the score of the positive ideal network A^{**} [11].

The results in [10] show, according to the comparison in terms of bandwidth, that the conversational traffic class receives approximately the same bandwidth in any of the four decision algorithms used.

Taking into account the different traffic test used (conversational, streaming, interactive and background) and the relationship of the parameters evaluated in each case, [10] determined that MEW, SAW, TOPSIS provide performance similar to the four classes of traffic. GRA provides a slightly larger bandwidth and also a minor delay for the interactive and background traffic.

C. Needless Handover and Hysteresis

The evaluation in [12] about the effects over quality of service in wireless networks was determined over the analysis of the horizontal handover process. Basically this process is divided in two parts: The first one is to determinate whether the handover is necessary and the AP selection destination; the second one is the evaluation of handover at layers 2 and 3, which takes into account the millisecond delay in the handover process (Layer 2 between 50 ms and 400 ms, and Layer 3 up to 300 ms depending on the network.). The mobile terminal is considered in such Handover technique as the best handover server. The transfer decision is based only on S_1 signal strength of a candidate AP.

Thus,

If, $S_i > S_0 + h$, then execute the handover process.

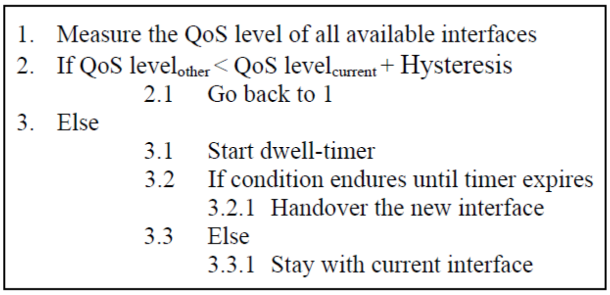
The mechanism consists to make a handover to a new AP when the signal strength has improved over certain h value. The hysteresis approach described, has the implicit drawback of producing a certain amount of unnecessary handovers when the ratio $S_i > S_0 + h$, it is still possible if the signal strength from the current AP is high.

In order to improve the last technique, SSHT (Relative Signal Strength with Hysteresis and Threshold) is evaluated.

This technique describes the implementation of the handover process if $S_i > S_0 + h$ and furthermore if $S_o > T_{cs}$. It means that the probability of having a handover in the path is not only subject to the level of hysteresis, but also set a threshold level that defines power of the current access point.

The Dwell Timer is an algorithm that ensures that the candidate network be stable. This algorithm is used with the hysteresis margin and only helps to optimize the handover system [1][13].

The dwell-timer waits a short time in order to verify if the candidate network QoS level persists, if the candidate network maintains the high level of QoS the network is reported as stable and could take the handoff, otherwise it is unstable and would not be a good candidate for the transfer process [13]. The operation of this algorithm is shown in the Figure 4:



```
1. Measure the QoS level of all available interfaces
2. If QoS level_other < QoS level_current + Hysteresis
        2.1   Go back to 1
3. Else
        3.1   Start dwell-timer
        3.2   If condition endures until timer expires
              3.2.1  Handover the new interface
        3.3   Else
              3.3.1  Stay with current interface
```

Fig. 4. A dwell-timer algorithm [13]

D. THOA (Trust-assisted Handover Decision Algorithm) Algorithm

The basic idea of THOA (Trust-assisted Handover Decision Algorithm) method is to have reliability coefficients considered in a making handover decisions. The candidate network selection is made from a list of networks to avoid attempts of risk handover.

THOA is implemented in both, the mobile terminal and the network, due to the separation of confidence information and the handover decision process. In [14] handover algorithm is based on trust assisted (THOA) using indicators of trust networks as a preliminary trial factor for deciding to handover. Trusted information of a candidate network is used by the mobile terminal to evaluate if the authentication required can be satisfactorily completed by the candidate networks.

According to the results in [14], the conclusion shows that THOA reduces the number of unnecessary handover attempts, reviewing the relationship of trust with a network, before activating the handover process. Each of these early reviews ensures that any attempt to handover will be effective.

In terms of quality and bandwidth, THOA offers lower bandwidth compared to MHOA method, which provides more bandwidth in the process of handover.

THOA can offer lower bandwidth due the fact that it selects an access point with a lower quality of service in search of reliable information. In this way, THOA may offer a faster handover but with less compromise to quality service.

E. Handover Based on Layer 2

The handover scheme proposed based MAC layer [15] is capable scan the available channels in a smooth way with the aim of generating a low impact applications in the upper layers, also it is used a mechanism adaptive to limit the search frequency of channels, primarily to adjust dynamically activation threshold for channels operational. This method is implemented on 802.11b devices and made different experiments to evaluate performance; each of these experiments was repeated for 10 times and Table V shows the average results.

Thus, the smooth handoff and greedy smooth handoff can significantly reduce the consecutive packet loss, both smooth and greedy smooth handoff, test the channels continuously, therefore, the delay and lost packets demonstrate the same pattern; the full scan handoff is used to emulate the firmware-based handoff. The firmware-based scheme has a better performance than the full scan handoff because hardware solution has less overhead than software solution. Therefore, if these handoff schemes are implemented by hardware, the packet loss and delay can be further reduced.

TABLE V
AVERAGE RESULTS OF THE SCHEMES

Scheme	Total Loss Packets	Max Delay (ms)
Full Scan	50.3	384.4
Smooth	6.2	48.1
Greedy	4.9	33.8
Firmware	24.6	102.2

The study proposed in [16] focuses in the impact of handover process to delay-sensitive applications, taking into account that in a hard handover process the mobile terminal cannot maintain simultaneous communication with a new AP and AP candidate; this implies that the data transfer is interrupted during the handover process.

In this way, the approach is as follow:

If, T is the total time of connection and A is the total delay of the handover process, then $T \leq A$ and it appears as a primitive process if $T = A$.

HAMS (Handoff Accurate Measurement Strategy), the method proposed in [17], suggested that the parameter that really affects performance is T and not A, since the higher the downtime, the greater the chance of loss, duplication, or delay of packets.

According to the testing, the value T is 6.907ms and the value A is 48.748ms, this indicates that the handover process takes a value of 41.841ms, i.e. after the scanning cycle and before the re-association, the mobile terminal exchanged three packets of data with other stations. It is likewise clear that technology tested was 802.11b, but it indicates that the technique can be applied to other wireless technologies higher rate.

F. Scanning Solution

The proposal [18] analyzes the handover methods for large 802.11 networks with goal of a handover process faster and smoother, the study in this type of networks focuses on improving the scanning process of the access point.

Based on the above mentioned the proposal performed a series of experiments to evaluate the critical parameters of transfer through extensive analysis of the data acquisition. Basically it determined two key features: the response time of scanning and the hidden information of the access point. The solution presented was renamed as D-Scan that uses the two features mentioned above to improve the scanning of point access. The scanning of the candidate access point is the principal cause of delay; therefore it is considered that a thorough scan before handover process (pre-scan) could be less satisfactorily for many applications. This study demonstrates that the delay time is of 50 ms to receive 99% of responses scan; these results were based on a field study on 802.11 networks in areas Hong Kong, such as universities, streets with wireless coverage, etc.

Likewise concluded that the pre-scan takes a long time to find the optimal access point, this time is between 500ms to 1s, it affects substantially the performance of applications used on the network.

D-Scan is activated by the current quality link of an AP, basically performs a detection of this quality standard, if the current link has a low quality then need a handover process, i.e. RSSI<HANDOFF_THRESHOLD, otherwise the network card starts to make a deep scan. This analysis pretends to find a certain amount of APs with a RSSI level greater than 75dBm, if it is not possible to find an access point that meets the requirements, and then the scan is passed to the next channel in order to cover the total frequency. [19]

VI. Conclusion

In this paper we have analyzed different methods proposed for the operation of the handover process. First, it takes into account the quality of service parameters for voice over IP as a starting point in the decision to handover, thus, it is important that parameters such as the throughput, jitter, and SNR, meets the standards recommended in the standard. In this way, it is possible proceed to the next level of handover process, where the cover, power level and transfer time serves as stabilization parameter through a hysteresis level supported by an algorithm such as the d-well timer.

Other methods such as cost functions based on policies set out three main parameters to make decisions for handover, the parameters are: cost, capacity and energy, applying these proportions to a relative and an absolute scale it is possible to identify the optimal network.

MADN methods propose additional parameters such as bandwidth, delay, packet loss rate, getting similar results for the four proposed methods according to the bandwidth and the type of traffic transmitted.

The description given for the MAC layer-based method and scanning solutions handle response time measurements, thresholds and conditions decision as most of the methods proposed, but represent advances that are taken into account in the process of handover stabilization even to be applied in other wireless environments.

References

[1] G Juan Guillermo Gómez T. José Lisandro Abadía S., Andrés Felipe Millán C. Estudio de Algoritmos de Handover para Redes Inalámbricas Heterogéneas, Grupo de Investigación COMBA I+D, Consorcio I2COMM- Univ. Santiago de Cali, Cali, Valle, Colombia, 2008.

[2] Chen L.-J., Sun T., Chen B., Rajendran V., and Gerla M., A smart decision model for vertical handoff, In Proceedings 4th ANWIRE International Workshop on Wireless Internet and Reconfigurability, Athens, Greece, 2004.

[3] Abderrahmane Lakas and Mohamed Boulmalf, Study of the Effect of Mobility Handover on VoIP over WLAN. - 2007.

[4] ITU, "Tranmision impairments due to speech processing," ITU, Geneva, Switzerland, ITU-T Rec. G. 113, (05/2002).

[5] Alfonso Fernandez Duran Eugenio Carrera del Pliegol and Jose I. Effects of Handover on Voice quality in wireless convergent networks - 2007.

[6] ITU-T."The E-Model, A Computational Model In Use In Transmission Planning".G.107, March 2003.

[7] R.Yavatkar, D. Pendarakis and R. Guerin, "A Framework for Policy-based Admission Control", RFC 2753, January 2000.

[8] Helen J. Wang, "Policy-Enabled Handoffs Across Heterogeneous Wireless Networks," Master Thesis, Computer Science Division, University of California, Berkeley, USA, 1998.

[9] Chen L.-J., Sun T., Chen B., Rajendran V., and Gerla M., A smart decision model for vertical handoff, In Proceedings 4th ANWIRE International Workshop on Wireless Internet and Reconfigurability, Athens, Greece, 2004.

[10] Wong Enrique Stevens-Navarro and Vincent W.S. Comparison between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks, Department of Electrical and Computer Engineering The University of British Columbia, Vancouver, Canada. - 2006.

[11] W. Zhang, "Handover Decision Using Fuzzy MADM in Heterogeneous Networks," in Proc. IEEE WCNC'04, Atlanta, GA, March 2004.

[12] Mariano Molina-García Alfonso Fernandez Duran, Raquel Perez Leal, and José I. Alonso Method to Assess the Effect of the Horizontal Handover Decision on Voice Quality in Wireless Convergent Networks. - 2009.

[13] Gustav Nyberg, Seamless Mobility – SEMO: A Policy-Based Prototype for Handovers in Heterogeneous Networks, Master Thesis, Department of Computing Science, UMEA University, Sweden, 2006.

[14] Mo Li Kumbesan Sandrasegaran, Tracy Tung Trust-Assisted Handover Decision Algorithm in Hybrid Wireless Networks - 2007.

[15] John Fitzpatrick Seán Murphy and John Murphy An Approach to Transport Layer Handover of VoIP over WLAN - 2006.

[16] Yong Liao Lixin Gao Practical Schemes for Smooth MAC Layer Handoff in 802.11Wireless Networks - 2006.

[17] Francisco A. González Jesús A. Pérez, and Victor H. Zárate HAMS: Layer 2 Handoff Accurate Measurement Strategy in WLANs 802.1. - 2005.

[18] Changqing Xu Jin Teng, Weijia Jia Enabling faster and smoother handoffs in AP-dense 802.11 wireless networks, Science Direct, El Sevier. - 2010.

[19] Kyoungnam Kwon Chaewoo Lee A Fast Handoff Algorithm using Intelligent Channel Scan for IEEE 802.11 WLANs – 2004.

Modified Decoding Algorithm for Irregular Repeat Accumulate Codes in Wireless Sensor System

Meng Zhang, Hao Liu, Tao Wang, Chao Chen, Fuqing Huang, and Jiafeng Zhu
National ASIC Research Center, School of Electronic Sci. and Eng.
Southeast University, Nanjing, 210096, China

Abstract – *IRA (Irregular Repeat Accumulate) codes can be well utilized in wireless sensor systems, computer storage systems due to their approaching to Shannon limit by means of prominent decoding algorithm and etc. Belief propagation algorithm is the best decoding algorithm for IRA codes, but there is great hardware complexity with implementation. Minimum-sum algorithm modifies belief propagation algorithm at the cost of decoding performance. To utilize both normalized factor and offset factor, a modified low complexity algorithm is proposed to improve decoding performance better. Simulation results show that if BER requests to be 10^{-5}, the encoding gain of the proposed decoding algorithm can be 0.5dB and 0.4dB better than normalized algorithm and offset algorithm respectively, and 0.1dB worse than belief propagation algorithm only.*

Keywords: IRA Codes, Belief Propagation Algorithm, Low Complexity, Normalized Factor, Offset Factor

1 Introduction

In the field of error correcting codes of wireless communication system and computer storage system, many scholars wanted to find out good codes with the deeper research, i.e. Turbo code and low density parity check (LDPC) code[1]. Therefore, seeking error correcting codes of linear encoding and decoding complexity and performance approaching to Shannon limit has been a hot issue in recent years. Turbo codes are a kind of good codes reaching Shannon limit, but with a more complex decoding algorithm. The decoding algorithm of LDPC codes is simple and its performance almost approaches to Shannon limit[2], but it has square encoding complexity. In 1998, Divsalar Doliner and Jin Hui etc proposed repeat-accumulate (RA) codes. In 2000, irregular repeat-accumulate (IRA) codes were proposed[3] [4].They proved that the binary IRA codes and irregular LDPC codes can get the same superior performance, with far less encoding complexity than LDPC codes. More and more scholars have carried out research in IRA codes recently[5~7]. So IRA codes have started to be applied in wireless sensor systems[5][8] and computer storage systems[9~11].

Belief propagation (BP) algorithm is the best decoding algorithm of IRA codes, but BP algorithm is very complex in implementation because it refers to hyperbolic tangent function and its inverse function. Look-up table is used in conventional method to implement those complex functions. Therefore, it is complex to implement by BP algorithm. Minimum- sum algorithm modifies belief propagation algorithm, but some decoding performance is lost. Normalized algorithm and offset algorithm are common modified minimum-sum algorithms. In this paper, an optimal low complexity decoding algorithm is proposed to improve decoding performance better by making the best of both normalized factor and offset factor.

2 Decoding Algorithm for IRA Codes

The IRA codes can be expressed by Tanner chart, shown in figure 1, which has information nodes, check nodes and parity nodes where f_i is the ratio of information nodes linked to check node c_i. Define the reliability information from check nodes to information nodes m[c, u], check nodes to parity nodes m[c, x], the edges linked node c to check nodes N(c).

Define the reliability information from check nodes to information nodes m[c,u], check nodes to parity nodes m[c,x], linked node c to check nodes N(c).Then

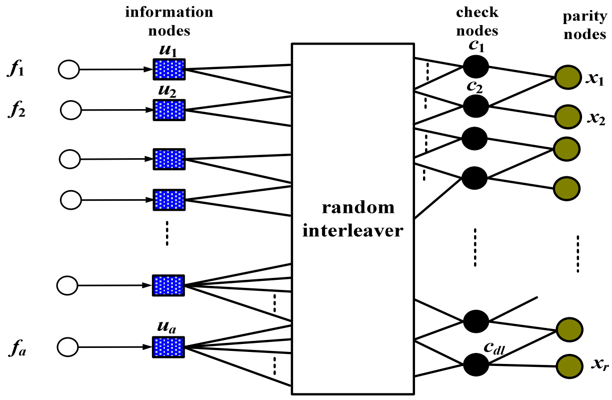


Figure 1 The Tanner chart of IRA codes

the iterative step of m[c, u] and m[c, x], in belief propagation decoding algorithm can be illustrated as:

$$m[c,u]=2\arctan h\left[\prod_{j=1}^{a-1}\tanh\left(\frac{m[u_j,c]}{2}\right)\bullet\prod_{j=1}^{2}\tanh\left(\frac{m[x_i,c]}{2}\right)\right]$$

$$u_j\in N(c)\setminus u, x_i\in N(c) \qquad (1)$$

$$m[c,x]=2\arctan h\left[\prod_{j=1}^{a}\tanh\left(\frac{m[u_j,c]}{2}\right)\bullet\tanh\left(\frac{m[x',c]}{2}\right)\right]$$

$$u_j\in N(c), x'\neq x \qquad (2)$$

We know belief propagation decoding algorithm is complex because of the complex function. Minimum-sum decoding algorithm is the simplified BP decoding algorithm like LDPC codes, which have the best decoding performance and great complexity. The equation m[c,x] and m[c,u] are as follows in minimum-sum decoding algorithm[1] [2].

$$m[c,u]\approx\prod_{j=1}^{a-1}\operatorname{sign}(m[u,c])\prod_{i=1}^{2}\operatorname{sign}(m[x',c])\bullet\min\left(|m[u_j,c]|,|m[x_i,c]|\right)$$

$$u_j\in N(c)\setminus u, x_i\in N(c) \qquad (3)$$

$$m[c,x]\approx\prod_{j=1}^{a}\operatorname{sign}(m[u_j,c])\operatorname{sign}(m[x',c])\min\left(|m[u_j,c]|,|m[x',c]|\right)$$

$$u_j\in N(c), x\neq x' \qquad （4）$$

Define m[c, u] and m[c, x] in minimum-sum decoding algorithm L_2 and the corresponding equations in BP decoding algorithm L_1. It can prove that the sign of L_2 is the same as that of L_1, but the amplitude of L_1 is smaller than that of L_2, namely $|L_1|<|L_2|$. So minimum- sum algorithm has a loss of performance.

There are two ways to reduce $|L_2|$ to $|L_1|$ by equation (5) and equation (6) like LDPC codes, leading to two decoding algorithms, normalized algorithm and offset algorithm respectively.

$$L_3=\alpha L_2 \qquad （5）$$

where normalization factor $0<\alpha<1$.

$$L_3=\operatorname{sign}(|L_2|)\max(|L_2|-\beta,0) \qquad （6）$$

where offset factor $\beta>0$.

In order to get the best decoding performance, normalization factor of equation (5) can be expressed by $\alpha=E[|L_1|]/E[|L_2|]$, while offset factor of equation (6) can be expressed by $\beta=E[|L_2|]-E[|L_1|]$. The parameters can be solved by Monte Carlo method and can be derived through the theory of probability too.

3 Modified Low Complexity Decoding Algorithm for IRA Codes

The modified decoding algorithm in this paper is proposed by adjusting multiplicative normalized factor and by adjusting additive offset factor at the same time. Therefore, the enhanced decoding performance of normalized algorithm and offset algorithm is limited. Taken the hardware complexity into account, a modified low complexity decoding algorithm is proposed using both normalized factor and offset factor. Because normalized factor and offset factor are a single multiplicative and additive parameter respectively, m[c, u] and m[c, x] can be expressed by

$$L_3=\max\left(\alpha|L_2|+\beta,0\right) \qquad (7)$$

where α is normalized factor, β is offset factor. The two factors can be solved to make the mean square error between $|L_3|$ and $|L_1|$ minimal.

The mean square error is:

$$\begin{aligned}\Delta(\alpha,\beta)&=E[(|L_3|-|L_1|)^2]=E[(\alpha|L_2|+\beta-|L_1|)^2]\\&=\alpha^2E[|L_2|^2]+\beta^2+E[|L_1|^2]+2\alpha\beta E[|L_2|]-2\alpha E[|L_1|]-2\beta E[|L_1|]\end{aligned} \tag{8}$$

when $\Delta(\alpha,\beta)$ is minimal,

$$\begin{cases}\partial\Delta(\alpha,\beta)/\partial\alpha=0\\ \partial\Delta(\alpha,\beta)/\partial\beta=0\end{cases} \tag{9}$$

That is to say:

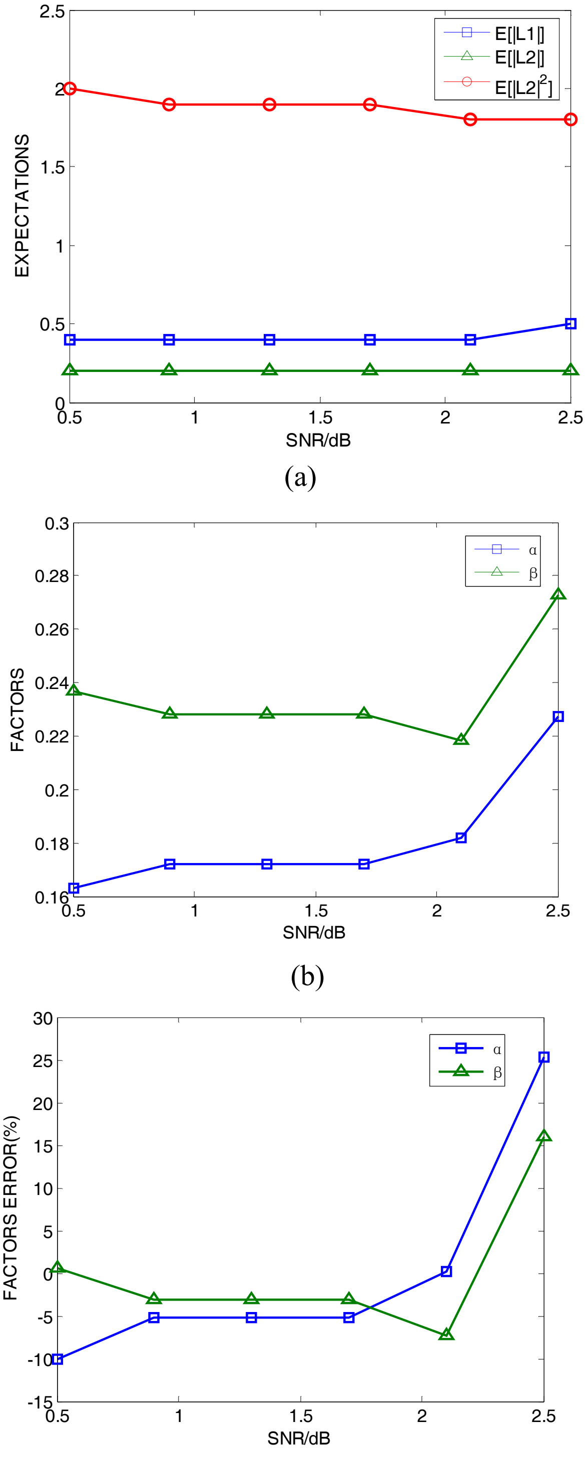
$$\begin{cases}\alpha E[|L_2|^2]+\beta E[|L_2|]=E[|L_1|]\\ \alpha E[|L_2|]+\beta=E[|L_1|]\end{cases} \tag{10}$$

We can get:

$$\begin{cases}\alpha=\dfrac{E[|L_1|](1-E[|L_2|])}{E[|L_2|^2]-(E[|L_2|])^2}\\ \beta=\dfrac{E[|L_1|]\{E[|L_2|^2]-1+E[|L_2|](1-E[|L_2|])\}}{E[|L_2|^2]-(E[|L_2|])^2}\end{cases} \tag{11}$$

4 Performance Simulation of Modified Low Complexity Algorithm

The IRA codes whose length is 2000 and coding ratio is 0.5 (a=8, f_3=0.452, f_{10}=0.148, f_{16}=0.075, f_{20}=0.3, f_{26}=0.125) to be modulated by BPSK through AWGN channel to get the performance of the proposed algorithm are simulated in the wireless sensor system, and iterative number is 30. At first, mathematical expectations $E[|L_1|]$, $E[|L_2|]$ and $E[|L_2|^2]$ need solving, which change with different SNRs. By Monte Carlo method, these three expectations can be calculated, shown in figure 2(a). Then normalized factor and offset factor α,β can be solved, shown in figure 2(b). We can know that α and β change according to different SNRs, and the errors to their average values are shown in figure 2(c). We find the errors aren't distinguished in figure 2(c), only about 20%.



(a)

(b)

(c)

Figure 2 (a) Mathematical expectations E[$|L_1|$], E[$|L_2|$] , E[$|L_2|^2$]. (b)Normalized factor α and offset factor β; (c) Errors of normalized factor and offset factor

To make the optimal algorithm simpler, we make $E[|L_1|], E[|L_2|], E[|L_2|^2]$ with differ- rent SNRs fix on their average values. And the simulation curves for decoding performance are shown in figure 3.

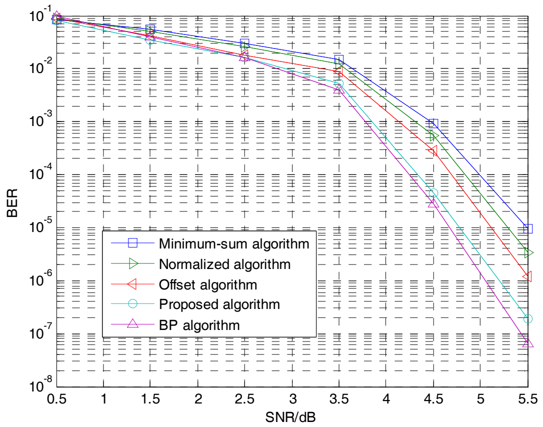


Figure 3 Performance simulation of IRA codes' several algorithms

From figure 3, we can know that if the wireless sensor system needs BER to be 10^{-5}, the encoding gain of the proposed algorithm can be 0.5dB and 0.4dB better than normalized algorithm and offset algorithm respectively, and only 0.1 dB less than BP algorithm. So the encoding gain of IRA codes is distinct in wireless sensor system. The hardware implementation complexity of these algorithms can be illustrated in Table I. BP algorithm is the most complex among these algorithms because complex functions, tanh and arctanh. What's more, hardware complexity will be greater because of the use of many multiplications. Normalized algorithm and offset algorithm are simpler than BP algorithm very much, because these algorithms won't refer to complex functions, and only a few of multiplications and adders. The proposed algorithm is a little more complex than normalized algorithm and offset algorithm, but its performance is excellent.

Table I Comparison table of the hardware complexity of decoding algorithm for IRA codes

algorithm	tanh	arctanh	multipli c-ation	Adder
BP	a+2	a+2	3(a+2)	0
Minim-sum	0	0	0	a+1+log$_2$(a+2)
Normalized	0	0	2	a+log$_2$(a+2)
Offset	0	0	0	a+2+log$_2$(a+2)
Purposed	0	0	2	a+1+log$_2$(a+2)

5 Conclusion

Making the best of the ideas of normalized algorithm and offset algorithm, a optimal low complexity algorithm is proposed according to minimum mean square error rule. Simulation results show that the encoding gain of the proposed algorithm can be much better than normalized algorithm and offset algorithm respectively, and a little worse than belief propagation algorithm. The proposed algorithm can be used in the wireless sensor systems, achieving good performance with low hardware complexity.

6 Acknowledgment

This research work was partly supported by the Natural Science Foundation of Jiangsu under Grant No.BK2010167, BK2010166, BK2010165 and the National Natural Science Foundation of China under No. 61006029 and No. 60871079.

7 References

[1] P. N. Hossein, F. Faramarz. 'Results on punctured low-density parity-check code and improved iterative decoding technique'. IEEE Transactions on Information Theory, 2007,53,(2),pp.599- 614

[2] R. S. Mohammad, H. B. Amir, P. Daniel. 'Low-density parity-check lattices: construction and decoding analysis'.IEEE Transactions on Information Theory, 2006, 10,(52),pp.4481-4495

[3] M D. Divsalar, H. Jin, R. McEliece. 'Coding theorems for Turbo-like code'. Proceedings of the 36th Annual Allerton Conference on Communication Control and Computing, Monticello, USA, 1998,pp.201~210

[4] Jin Hui, A. Khandeka, R. McEliece. 'Irregular repeat accumulate codes'. Proceedings of the 2nd International Symposium on Turbo codes and Related Topics, Brest, France, 2000. pp. 1-8

[5] Eckford, A.W., Adve, R.S., 'A Practical Scheme for Relaying in Sensor Networks Using Repeat-Accumulate Codes'. 40th Annual Conference on

Information Sciences and Systems, 2006.

[6] Hwang S O, Myung S, Lee H. 'Partial Prallel Encoder for IRA Codes'. Electronics Letters, 2010, 46,(2),pp.135- 137

[7] Y.S Ye, X.C. Liu, H. Cho. An energy- efficient single-hop wireless sensor network using Repeat-Accumulate codes, International Conference on Communications, Circuits and Systems, 2008. ICCCAS 2008.

[8] Mao-Ching Chiu. 'Bandwidth-Efficient Modulation Codes Based on Nonbinary Irregular Repeat-Accumulate Codes'. IEEE Transactions on Information Theory, 2010, 56,(1), pp. 152-167

[9] Yang Han, Ryan W.E., 'Performance of a Structured IRA Code on a Perpendicular Recording Channel with Media Noise', IEEE Global Telecommunications Conference, pp. 271-276, Nov. 2007.

[10] Han Yang, Ryan, William E. 'Performance of a structured IRA code on a perpendicular recording channel with media noise'. IEEE Global Telecommunications Conference Proceeding, GLOBECOM 2007,pp.271-276, 2007.

[11] K. Kenta, M. Shinya, S. Tomoharu, S. Kohichi, 'Design of irregular repeat accumulate codes with joint degree distributions'.IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2006, E89-A, (11), pp. 3351-3354

Performance Comparison of AODV and AOMDV routing protocols

S. R. Biradar[1], Sawal Kishore Singh[2], Subir Kumar Sarkar[3]
[1]MITS, Lakshmangrah – 332311, INDIA
[2]National Institute Of Technology, Patna-800005, INDIA
[3]Jadavpur University, Kolkata-700032, INDIA
[1]srbiradar@gmail.com, [2]srbiradar@gmail.com, [3]su_sarkar@hotmail.com

Abstract— *Ad hoc networks are a new wireless networking paradigm for mobile nodes. It enabling devices to create and join networks on-the-fly anywhere, anytime. The military, emergency rescue, disaster relief operations are the main applications of ad hoc networks. Unlike traditional wireless networks, ad hoc networks do not rely on any fixed infrastructure. Instead, hosts rely on each other to keep the network connected. In this paper we compared single path and multipath routing protocols over mobile ad hoc wireless networks using ns-2 simulator. Simulation results are presented by varying number of source, pause time and node mobility.*

Keywords— AODV, AOMDV, Routing, ns-2, Simulation

1. Introduction

Multipath routing protocols for MANETs main objectives are to provide reliable communication and to ensure load balancing as well as to improve quality of service (QoS) of MANETs. The main objectives of MANET are providing a solution to the user connected anytime, anywhere to a network. This kind of network is spontaneous, self-organized and self-maintained. The primary challenge of routing in such an environment is to cope with node mobility, which leads to frequently changing network topology. Several routing protocols [1, 2, 3, 4] have been proposed in the literature for ad hoc networks. Most of these protocols assume cooperative network settings, where nodes are willing to receive, and transmit packets. The key idea behind all these protocols is to establish and maintain loop-free paths to respective destinations while achieving a good balance among important performance metrics.

Multipath routing protocols establish multiple routes from source to destination. The main advantage of discovering multiple paths is that the bandwidth between links is used more effectively with greater delivery reliability. It also helps during times of the network congestion. Multiple paths are generated on demand or by using a proactive approach and are of great significance because routes generally get disconnected quickly due to node mobility. The main objectives of multipath routing protocols are to provide reliable communication and to ensure load balancing as well as to improve quality of service (QoS) of ad hoc and mobile networks. Other goals of multipath routing protocols are to improve delay, to reduce overhead and to maximize network life time. The many multipath routing protocols were proposed for MANETs [5]. They can be classified according to different criteria. Based on the number of paths that are discovered from a route request, the routing protocols are divided into single path [2, 5] and multipath [7, 8]. On-demand multipath routing protocols [6, 12] establish multiple paths to a given destination in a single-route discovery phase. Multiple paths to a destination provide better fault-tolerance to path breaks. In the case of on-demand multipath protocols a new route discovery (which typically requires a network-wide flooding of a route request message) is necessary only when all paths to a given destination break. Thus, they provide an (overhead) efficient means to recover from routing failure compared to single-path on-demand routing protocols.

Another feature of the routing protocols is the number of the discovered paths that are actually used for sending data. Some protocols use only a single path for the communication, while others distribute the data through different channels. The ad hoc network is more prone to both link and node failures due to expired node power or node mobility. As a result, the route used for routing might break down for different reasons. To increase the routing resilience against link and node failures, one solution is to route a message via multiple disjoint paths simultaneously. Thus, the destination node is still able to receive the message even if there is only one routing path exist. The multipath approach takes advantage from the large and dense networks. The route discovery process in the multipath protocols may be initiated either when the active path collapses or when all known paths towards the destination are broken [9]. The route discovery may stop when a sufficient number of paths are discovered or when all possible paths are detected. Multipath routing protocols can be node-disjoint [10] or link-disjoint [11] if a node (or a link) cannot participate in more than one path between two end nodes.

Multiple paths can be used as backup route or be employed simultaneously for parallel data transmission. The multiple paths obtained can be grouped into three categories:

Disjoint: This group can be classified into node-disjoint and link-disjoint. In the node-disjoint multipath type, there are no shared nodes between the calculated paths that links source and destination. The link-disjoint multipath type may share some nodes, but all the links are different.

Inter-twisted: The inter-twisted multipath type may share one or more route links.

Hybrid paths: The combination of previous two kinds.

Many disjoint multipath routing techniques [13, 14, 15, 16, 17, 18, 19, 20] have been proposed for ad hoc networks, which have focused on improving the reliability of routing using path disjointness. In [13] proposed a node-disjoint multipath routing protocol for traffic load-balancing. They introduce a correlation factor for a set of multiple paths between source and destination, which measures the disjointness of paths in the set. The routing algorithm selects the set of multiple paths with minimum correlation so as to minimize the interference between transmissions in the individual paths. Saha et al. [15] proposed a maximally zone-disjoint multipath routing, which computes a set of zone-disjoint shortest paths for traffic load-balancing. Disjoint multipath source routing proposed in [17], statically multiplexes the data traffic over multiple disjoint paths at all nodes on the primary path. The stability based multipath routing algorithm proposed in [16] computes a set of stable independent (disjoint) multiple paths, which can be used for a longer time to recover from the path breakage.

Multipath protocols [12, 21, 22] based on distance-vector routing scheme have been proposed for ad hoc networks. Other multipath routing protocols have been also proposed for ad hoc networks [9, 23, 24]. Multipath Source Routing (MSR) [23, 25] extends DSR's route discovery and route maintenance phases to compute multiple node-disjoint paths. It also proposes a mechanism to distribute load over multiple paths, based on the RTT measurement. SMR [14] finds maximally disjoint multiple paths and uses a per-packet allocation scheme to distribute data packets on to multiple paths. This enables the effective utilization of network resources and avoids nodes from being congested. SMR computes only two paths to each destination. All the above protocols are based on the source routing protocol DSR. They cannot scale to large networks because source routing requires every data packet to carry full path to the destination. AODV-BR [21] calculates multiple paths without any extra control overhead. In this protocol, neighboring nodes hear the route reply transmissions by being in promiscuous mode, and store a route to the destination through the neighbor that transmitted the reply packet. The newly discovered paths are called a backup path. But, effectively nodes on the primary path contain only single path to the destination. It is the neighboring nodes who store backup paths. When a node on the primary path moves away due to mobility, it loses connection to its immediate downstream node on the path. Then, the node broadcasts the future data packets that it receives for that destination, assuming that any of its neighbors would have stored a backup path to the destination. The node also sends a route error packet to the source node, informing the route disconnection. Maxemchuck [26] proposed a routing mechanism called dispersity routing for store and forward data networks. It discusses the ways of splitting data and dispersing them over multiple paths to achieve smaller average and variance in delay. The popular link-state protocol OSPF [27] can find multiple paths of equal cost. In [28], Ad hoc On-demand Multipath Distance Vector (AOMDV) [12] is a multipath routing protocol based on AODV [29], which can compute both node-disjoint and two segment link-disjoint paths. It uses a notion of advertising-hop count to form an invariant that ensures loop freedom. Further, in large networks two segment link-disjoint paths also take considerable amount of time to recover the route break, as route error has to traverse multiple hops to inform the route disconnection to the node that has alternate path(s).

Ad hoc On-demand Distance Vector Multipath (AODVM) is also a multipath routing protocol based on AODV. It proposes a routing framework to provide robustness to route breaks. The protocol computes node-disjoint paths between source and destination. Through simulation results, it shows that only a few such multiple paths can be found in the network and hence they cannot provide much robustness. In order to increase robustness to route breaks, AODVM assumes the existence of a set of reliable nodes in the network and place them at the junctions of the link-disjoint multiple paths. A mechanism to identify such reliable nodes in the network would be a good addition to the protocol.

In [18] proposed a stable node-disjoint multipath routing, which applies the path accumulation feature of DSR and AODV. But, this path accumulation feature requires the route request packet to carry the full path it has traversed. This requirement increases the size of route request packet, particularly in large networks where paths between nodes are longer. Disjoint multipath routing [19] proposed by Abbas and Jain tries to reduce the effect of path diminution problem in finding node-disjoint multiple paths. As this routing technique also requires the route request packets to carry the traversed path, it suffers from the same disadvantage as the previous protocol. In [30], Ducatelle et al. propose a hybrid multipath routing based on ant colony optimization framework for traffic load-balancing. Multipath fresnel zone routing [31] proposed to take the capacity of intermediate nodes into consideration for selecting disjoint multiple paths. It evaluates the capacity and the transmitting cost of different intermediate nodes, and formulates end-to-end paths of different capacity and cost. Then the protocol forwards the traffic through these different paths, by adjusting the amount of traffic on each path based on path capacity and congestion conditions.

Roy et al. [32] compared the two disjoint multipath techniques that use omni-directional and directional antennas, respectively. They showed through simulations that directional antennas help in computing multiple paths efficiently, when compared to omni-directional antennas. Fault tolerant routing proposed in [33] uses a path estimation mechanism for selecting a reliable route. This algorithm relies on destination nodes sending feedback to source nodes, about the packet delivery capacity of multiple paths, which may increase the control overhead in the network. Also, there is some protocol complexity involved in implementing the path estimation mechanism. All the disjoint multipath routing techniques discussed above compute maximally disjoint multiple paths, whose availability is very less due to the disjointness constraint imposed in the path selection. Hence, disjoint multiple paths cannot provide efficient fault-tolerance towards route breaks. Also, these techniques involve considerable delay and overhead in selecting disjoint multiple paths. The multipath routing mechanisms proposed in [34, 35] use path redundancy for ensuring confidentiality to data transmission over ad hoc networks, which are not closely related to fault-tolerant multipath routing techniques

2. Multipath routing protocols

Ad hoc On-demand Multipath Distance Vector (AOMDV) [11, 12] is an extension to the AODV protocol for computing multiple loop-free and link-disjoint paths. AOMDV is a reactive hop-by-hop routing protocol which finds node/link-disjoint multiple paths. In AOMDV when a source has packets to send to a destination and finds no routes in its routing table, it invokes a route discovery by broadcasting RREQ packets. Route discovery in AOMDV is similar to AODV. A RREQ packet in AOMDV includes all fields as that in AODV. Besides, it includes an additional field called the last hop, i.e., the neighboring node of the source. This information and the next-hop information, i.e., the node from which to receive the RREQ, are used to achieve link-disjointness (shown in Figure 1) for reverse paths to the source. A node may receive multiple duplicate RREQ packets. For each packets received, it examines if an alternate reverse path to the source can be formed such that loop-freedom and link-disjointness are preserved.

Upon establishing a reverse path to the source, an intermediate node checks if it has any valid path to the destination. If so, it generates RREP packets, including a forwarding path not used in any previous RREPs for this RREQ, and sends the RREP back to the source through the reverse path. Otherwise, it checks if it has forwarded this request before and forwards it if not. Upon receiving an RREQ packet, the destination tries to form a reverse path to the source. Then, it generates an RREP packet for each RREQ copy that arrives through a loop-free path to the source. Multiple RREPs intend to increase the probability to find multiple disjoint paths.

Upon receiving an RREP packet, an intermediate node checks if it can form a loop-free and disjoint path to the destination using the same rule as when it receives an RREQ packet. If not, it drops the RREP. Otherwise, it checks if there is any reverse path to the source that has not been used to forward an RREP for this route discovery. If so, it chooses one of the unused reverse paths to forward the RREP. Otherwise, it drops the packet.

Loop-freedom is guaranteed by satisfying the following sufficient conditions:

Sequence rule: For each destination, multiple paths maintained by a node should have the same sequence number, i.e., the highest known destination sequence.

For the same destination sequence number: a node never advertises a route shorter than one already advertised, and never accepts a route longer than one already advertised. More details and a proof of correctness are available in [12].

Route maintenance is also very similar to AODV. Upon link breakage of the last path to the destination, a node generates or forwards an RERR packet.

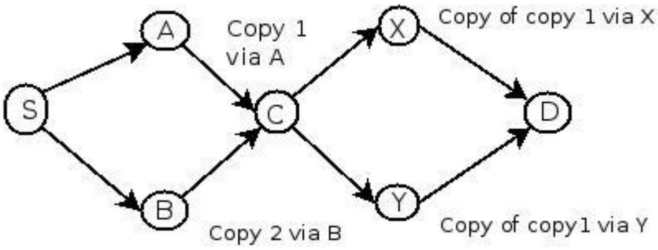


Figure 1 Route discovery process in AOMDV

3. Performance comparison of single path and multipath routing protocols

Simulation Environment: Network simulation ns-2 is with distributed coordination function of IEEE 802.11 for wireless LANs is used. The radio model uses characteristics similar to Lucent's WaveLAN radio interface. The bit rate is set to 2Mbps and the radio range is limited to 250 meters and an error free wireless channel model is used.

Movement Model: The random waypoint model is used to model node movements. Simulation time is 1000 seconds while the pause time varies from 0 seconds (continuous movement) to 1000 seconds (no mobility) [0, 50, 100, 250, 500, and 1000 seconds]. We vary the node speed 1m/s, 10m/s and 20m/s to compare the protocols performance for low and high mobility. Nodes speed uniformly distributed in the ranges of 0 to 1 m/s, 0 to 10 m/s and 0 to 20 m/s.

Network size and communication model: The network size of 40 mobile nodes in rectangular area of size 1000m x 1000m. Traffic pattern used up to 20 CBR/UDP connections, with sending rate of 4 packets per second between randomly chosen source-destination pairs. Connections begin at random times during the simulations. We use the identical traffic and mobility model pattern for different routing protocols. Data packets have fixed size of 512 bytes and the interface network queue size for routing is set to 50 packets for all scenarios.

Scheduling data packets: In AOMDV sender uses all available paths to a destination simultaneously. Data packets are sent over each individual path with equal probabilities. When one path breaks, the source stops using that path but does not initiate a new route discovery process. Only when all available paths are broken then a new route discovery process is initiated. In case of AODV, if existing route fails, it initiates the route discovery process.

Simulation results of AOMDV with AODV are compared to emphasize the advantages and disadvantage of multipath versus single path routing in ad hoc networks.

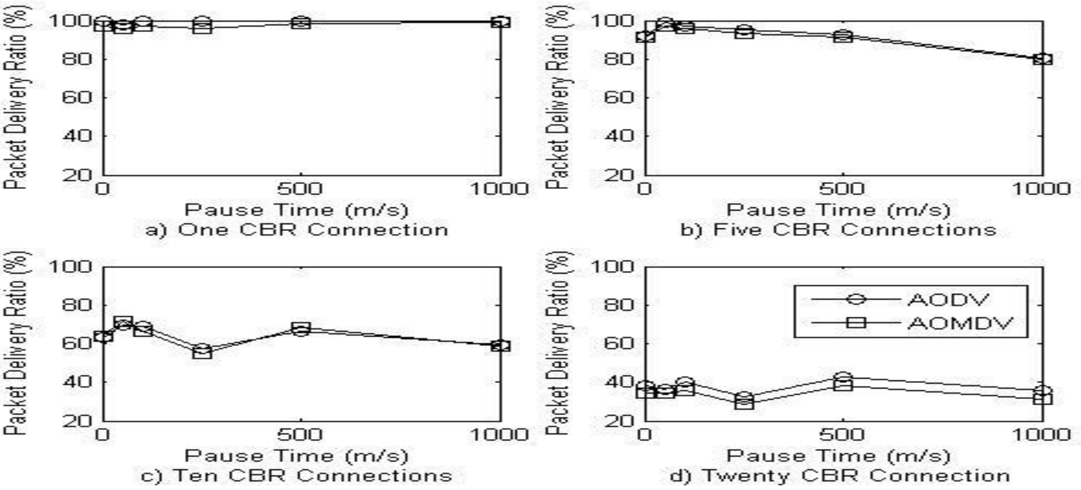
Packet Delivery Ratio:

Figure 2, 3 and 4 shows the packet delivery ratio of the originated application data packets each protocol was able to deliver, as a function of node mobility, pause time and network load. For both routing protocol, PDR is very high for low offered traffic load, and low PDR for high offered load. Routing protocols does well in low traffic 1 or 5 sources, delivery ratio is between 80% to 100%. In case of high traffic load 20 sources, PDR is less than 40%. Traffic load has lot of impact on routing protocols. Nodes speeds also affect the PDR metric, as node speed increase performance degrades. The reason is that the AOMDV source waits until all existing path break before sending a new route request and the probability that two paths break is lower if they are node disjoint.

Routing Overhead:

Figure 5, 6, and 7 shows the number of routing protocol packets sent by each protocol in obtaining the packet delivery ratio as shown in Figure 2, 3 and 4 respectively. In AOMDV protocol control overhead is almost independent of pause time. As the number of sources increase, routing overhead also increase. For low traffic AODV routing load is low, but in case of high load AOMDV uses less routing packets. Overhead is almost constant in low traffic. In the following we discuss in detail the routing overhead for each scenario.

Low mobility: The routing overhead in networks with low traffic is shown in Figure 5a. We clearly observe the higher (at least 50 times) over head of AOMDV compared to the AODV routing protocol. When comparing AODV and AOMDV, the high traffic (20 CBR) multipath routing protocols requires less control messages, approximately it saves at least 40000 routing packets compared to AODV shown in Figure 5d.



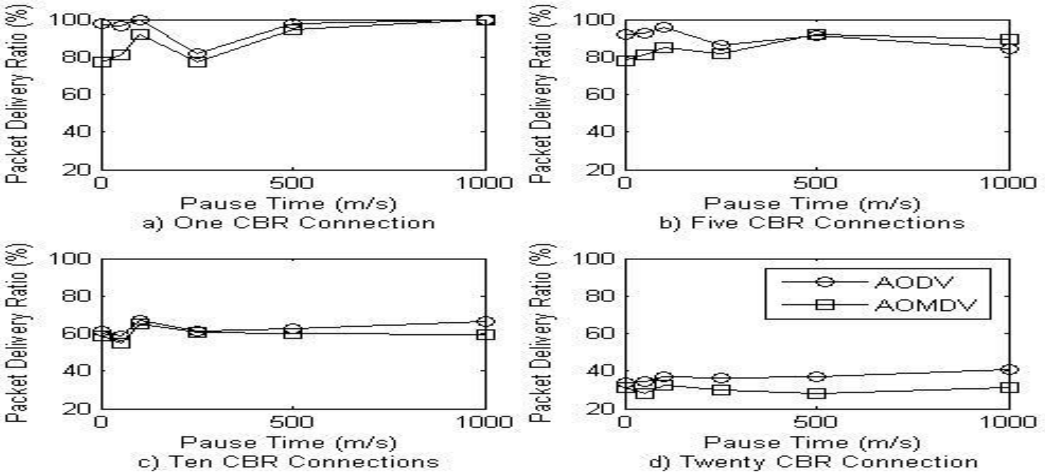
a) One CBR Connection

b) Five CBR Connections

c) Ten CBR Connections

d) Twenty CBR Connection

Figure 2 Comparison of the packet delivery ratio of application data packets successfully delivered as a function of pause time. Speed 1m/s



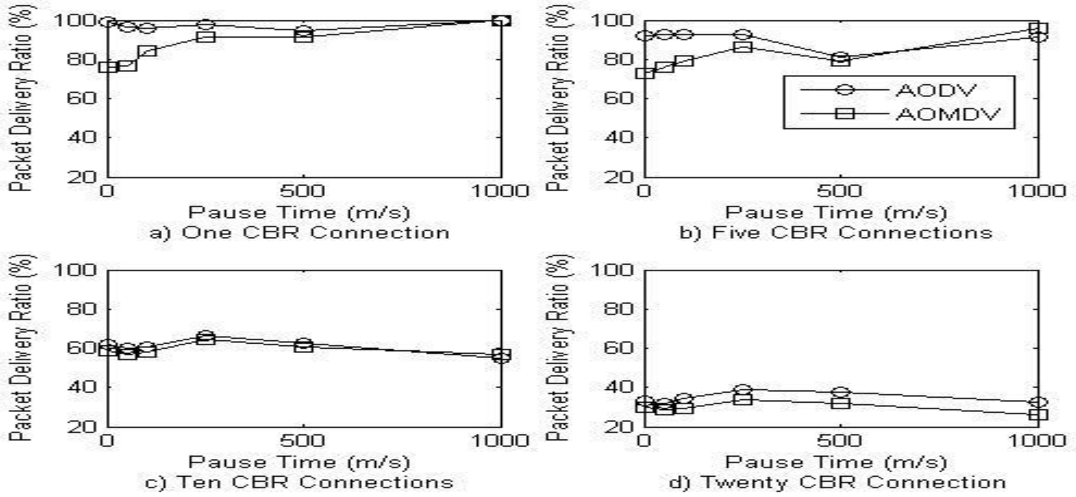
a) One CBR Connection

b) Five CBR Connections

c) Ten CBR Connections

d) Twenty CBR Connection

Figure 3 Comparison of the packet delivery ratio of application data packets successfully delivered as a function of pause time. Speed 10m/s



a) One CBR Connection

b) Five CBR Connections

c) Ten CBR Connections

d) Twenty CBR Connection

Figure 4 Comparison of the packet delivery ratio of application data packets successfully delivered as a function of pause time. Speed 20m/s

Moderate Mobility: In Figure 6 show the routing overhead for moderate mobility, the control over head is similar to low mobility shown in Figure 5. In case of low traffic control over head is almost independent of pause time. AODV overhead is low in low traffic few hundred routing packets are required but in case of high traffic it needs approximately 150000 routing packets.

High Mobility: In Figure 7 AODV routing over head is almost constant in low traffic as a function of pause time, but in case of high traffic routing overhead decline as a function of pause time increase. AOMDV routing is not affected by pause time. When number of connection changed from 1 to 5 only it contributed approximately 10000 routing packets. Whereas, number of connection varied from 10 to 20, control over head also double.

Overall in low traffic AODV performs better than AOMDV and in high traffic AOMDV is preferred compared to AODV and nodes mobility and pause time has little impact.

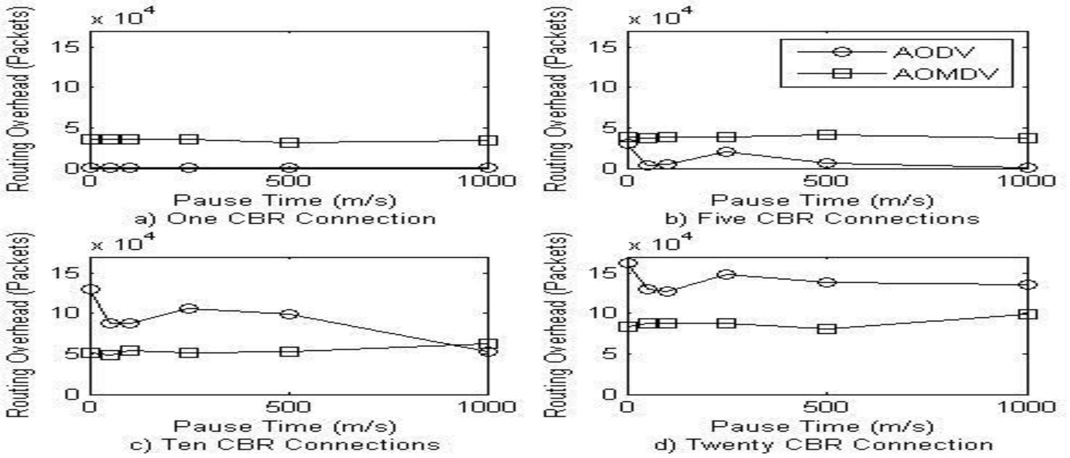


Figure 5 Comparison of the number of routing packets sent as a function of pause time. Speed 1m/s

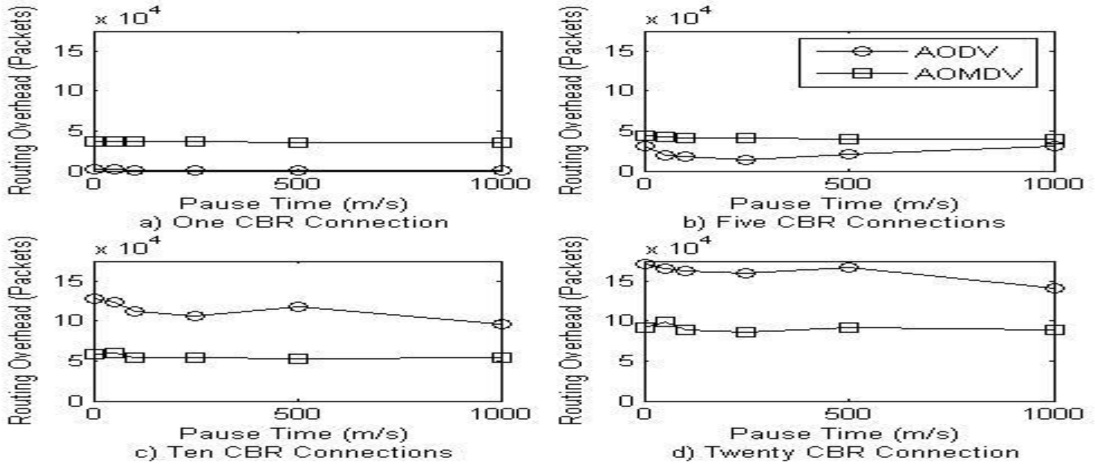


Figure 6 Comparison of the number of routing packets sent as a function of pause time. Speed 10m/s

4. CONCLUSION

Table 1 has given performance comparison both single path AODV and multipath AOMDV routing protocols, "high" denotes the best performance, "low" the worst. We summaries the performance of multipath and single path routing protocols for the different metrics. We find that: Multipath routing achieves in general better performance than single path routing in high traffic loads and vice versa.

Table 1 Single vs Multipath results

Traffic Sources	PDR		Routing Overhead	
	AODV	AOMDV	AODV	AOMDV
1 and 5	High	High	Low	high
10	low	Low	High	low
20	low	low	High	low

5. References

[1]. Zygmunt J. Haas, Marc R. Pearlman, The performance of query control schemes for the zone routing protocol, in: Proceedings of SIGCOMM'98, 1998.

[2]. D. B. Johnson and D.A. Maltz, Dynamic source routing in ad-hoc wireless networks, Mobile Computing, Kluwer Academic Publishers (1996), pp. 152–181.

[3]. Vincent D. Park, M. Scott Corson, A highly adaptive distributed routing algorithm for mobile wireless networks, in: Proceedings of INFOCOM'97, 1997.

[4]. Rendong Bai and M. Singhal, Doa: Dsr over aodv routing for mobile ad-hoc networks, IEEE Transactions on Mobile Computing 5 (10) (2006), pp. 1403–1416.

[5]. M. Tarique, K.E. Tepe, S. Adibi and S. Erfani, Survey of multipath routing protocols for mobile ad hoc networks, Journal of Network and Computer Applications 32 (2009), pp. 1125–1143.

[6]. C. Perkins, E. Royer, S. Das, Ad hoc on-demand distance vector routing, in: Proc. of the Workshop on Mobile Computing Systems and Applications, IEEE, February 1999, pp. 90–100.

[7]. A. P. Subramanian, A.J. Anto, J. Vasudevan and P. Narayanasamy, Multipath power sensitive routing protocol for mobile ad hoc networks, Proc. of the 1st IFIP/TC6 Working Conference on Wireless On-

Demand Network Systems (WONS'2004), LNCS vol. 28, Springer-Verlag (2004), pp. 171–183.

[8]. A. Nasipuri, S.R. Das, On-demand multipath routing for mobile ad hoc networks, in: Proc. of INFOCOM'99, IEEE, 1999, pp. 64–70.

[9]. S. -J. Lee, M. Gerla, Split multipath routing with maximally disjoint paths in ad hoc networks, in: Proceedings of IEEE International Conference on Communications 2001, vol. 3, June 2001, pp. 867–871.

[10]. Jie Wu, An extended dynamic source routing scheme in ad hoc wireless networks, Telecommun. Syst. 1 (2003) (4), pp. 61–75.

[11]. M. Marina and S. Das, Ad hoc on-demand multipath distance vector routing, ACM Mobile Comput. Commun. Rev. 6 (2002) (3), pp. 92–93.

[12]. M. K. Marina and S.R. Das, On-demand multi path distance vector routing in ad hoc networks, Proceedings of the Ninth International Conference on Network Protocols, IEEE Computer Society, Washington, DC, USA (2001), pp. 14–23.

[13]. K. Wu, J. Harms, Performance study of a multipath routing method for wireless mobile ad hoc networks, in: Proceedings of Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, August 2001, pp. 99–107.

[14]. A. Tsirigos, Z.J. Haas, Multipath routing in mobile ad hoc networks or how to route in the presence of frequent topology changes, in: Proceeding of IEEE Military Communications Conference, vol. 2, October 2001, pp. 878–883.

[15]. D. Saha, S. Toy, S. Bandyopadhyay, T. Ueda, S. Tanaka, An adaptive framework for multipath routing via maximally zone-disjoint shortest paths in ad hoc wireless networks with directional antenna, in: Proceedings of IEEE Global Telecommunications Conference, vol. 1, December 2003, pp. 226–230.

[16]. J. Shi, Z. Ling, S. Dong, Z. Jie, A stability-based multipath routing algorithm for ad hoc networks, in: Proceedings of 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, vol. 1, September 2003, pp. 516–520.

[17]. N. Wisitpongphan, O.K. Tonguz, Disjoint multipath source routing in ad hoc networks: transport capacity, in: Proceedings of IEEE 58th Vehicular Technology Conference, vol. 4, October 2003, pp. 2207–2211.

[18]. X. Li, L. Cuthbert, Stable node-disjoint multipath routing with low overhead in mobile ad hoc networks, in: Proceedings of The IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, October 2004, pp. 184–191.

[19]. A. M. Abbas, B.N. Jain, Path diminution in disjoint multipath routing for mobile ad hoc networks, in: Proceedings of 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 1, September 2004, pp. 130–134.

[20]. P. Papadimitratos, Z.J. Haas, E.G. Sirer, Path set selection in mobile ad hoc networks, in: MobiHoc '02: Proceedings of the 3rd ACM International Symposium on Mobile Ad hoc Networking & Computing, 2004, pp. 23–29.

[21]. S. -J. Lee, M. Gerla, AODV-BR: Backup routing in ad hoc networks, in: IEEE WCNC'2000, vol. 3, September 2000, pp. 1311–1316.

[22]. A. Valera, W.K.G. Seah and S.V. Rao, Cooperative packet caching and shortest multipath routing in mobile ad hoc networks, IEEE INFOCOM 2003 1 (2003) (April), pp. 260–269.

[23]. L. Wang, L. Zhang, Y. Shu, M. Dong, Multipath source routing in wireless ad hoc networks, in: Proceedings of Canadian Conference on Electrical and Computer Engineering 2000, vol. 1, March 2000, pp. 479–483.

[24]. A. Nasipuri, R. Castaneda and S.R. Das, Performance of multipath routing for on-demand protocols in mobile ad hoc networks, Mobile Networks and Applications (MONET) 6 (2001) (4), pp. 339–349.

[25]. L. Wang, Y. Shu, M. Dong, L. Zhang, O.W.W. Yang, Adaptive multipath source routing in ad hoc networks, in: Proceedings of IEEE International Conference on Communications, vol. 3, June 2001, pp. 867–871.

[26]. N. F. Maxemchuk, Diversity routing in store and forward networks, Ph.D. thesis, University of Pennsylvania, May 1975.

[27]. J. Moy, OSPF version 2, RFC 1247, 1991.

[28]. S. Vutukury, J.J. Garcia-Luna-Aceves, MDVA: A distance-vector multipath routing protocol, in: Proceedings of the IEEE INFOCOM, 2001, pp. 99–107.

[29]. A.E. Perkins, E.M. Royer, Ad-hoc on-demand distance vector routing, in: Proceedings of the IEEE WMCSA'99, New Orleans, LA, February 1999, pp. 90–100.

[30]. F. Ducatelle, G.D. Caro, L.M. Gambardella, Ant agents for hybrid multipath routing in mobile ad hoc networks, in: Proceedings of the Second Annual Conference on Wireless on demand Network Systems and Services (WONS), January 2005.

[31]. Y. Liang, S.F. Midkiff, Multipath fresnel zone routing for wireless ad hoc networks, in: Proceedings of 2005 IEEE Wireless Communications and Networking Conference, vol. 4, March 2005, pp. 1958–1963.

[32]. S. Roy, S. Bandyopadhyay, T. Ueda, K. Hasuike, Multipath routing in ad hoc wireless networks with omni directional and directional antenna: a comparative study, in: IWDC '02: Proceedings of the 4th International Workshop on Distributed Computing, Mobile and Wireless Computing, 2002, pp. 184–191.

[33]. Yuan Xue, K. Nahrstedt, Fault tolerant routing in mobile ad hoc networks, in: 2003 IEEE Wireless Communications and Networking, vol. 2, March 2003, pp. 1174–1179.

[34]. L. -P. Chou, C.-C. Hsu, F. Wu, A reliable multipath routing protocol for ad-hoc network, in: Proceedings of 10th IEEE International Conference on Networks, August 2002, pp. 305–310.

[35]. K.-L. Lee, X.-H. Lin, Y.-K. Kwok, A multipath ad hoc routing approach to combat wireless link insecurity, in: Proceedings of IEEE International Conference on Communications, vol. 1, May 2003, pp. 448–452.

BANDWIDTH ESTIMATION FOR RADIO BASED TRAIN CONTROL AND COMMUNICATION SYSTEM

Byungsik Yoon[1], Min-woo Jung[1], Sook-jin Lee[1], Keun-hong Min[2], and Young-Kyu Kim[2]
[1]Wireless Convergence Research Team, ETRI, Daejeon, Republic of KOREA
[2] Train Control & Communication Research Department, KRRI, Uiwang, Republic of KOREA

Abstract - *The candidates of frequency bands and mobile communication systems for radio based train control and railway communication are introduced in this paper. The strength and weakness of frequency bands and mobile communication systems are also addressed for developing optimum frequency utilization and cost efficient railway infrastructure. A new calculation method of frequency spectrum bandwidth is proposed for radio based train control and railway communication system. Proposed method would be used to request new frequency band for dedicated railway communication system in Korea.*

Keywords: Railway communication, Train control, CBTC, GSM-R.

1 Introduction

During the last two decades, overall railroad infrastructures have been rapidly developed with increased demand for railway services. Especially, railway signaling and communication are key factors to maintain reliable, safe and secure operation. In the past, the wired communication systems were used for train signaling and communications in railway industry. Nowadays, wireless communication systems have emerged with a variety of advantages. The radio based train control and communication system is consider as cost efficient digital replacement for existing wired and analogue old railway system. North America and European countries have developed the radio based train control system to improve interoperability, safety, and cost efficiency. Unlike, a U.S. communications-based train control (CBTC) uses unlicensed frequency band, European GSM-R (GSM-Railway) uses the licensed frequency band with 8 MHz bandwidth (876 ~ 880 MHz for uplink and 921 ~ 925 MHz for downlink) [1][2]. Although, there are a lot of difficulty to be given dedicated licensed frequency band from the government, licensed frequency has many advantages over train control system. Licensed frequencies has no interference with other frequency band and it allows higher transmit power to increase radio coverage. Therefore, the reliability for secure communication can be increased, whereas infrastructure efforts along the railway track will be reduced. Especially, mobile communication based train control system has cost-effective deployment based on open and standard based technology. Modern mobile communication systems provide very high data rate and very high mobility with secure IP equipment interoperability. These key features are required for modernized railway communication architecture.

In this paper, we introduce the analysis result of frequency bands and mobile communication systems for Korean Radio based Train Control System (KRTCS) [3]. The candidates of licensed frequency for KRTCS are addressed in section 2. In section 3, we describe various candidates of mobile communication system for railway infrastructure. The estimation results of spectrum bandwidth for each system are shown in section 4, and conclusions are drawn in section 5.

2 Frequency candidates for KRTCS

Since current Korean railway communication infrastructure is based on analog based radio system, railway communication services are restricted by poor quality of services, limited applications, large number of staff using the same channel and limited local connectivity. Because current radio system for railway infrastructure doesn’t meet the requirement of future railway communication, Korean government prepares new railway communication system. KRTCS is next generation train control and communication system which will replace current railway infrastructure in 2015. KRTCS use the dedicated radio spectrum for train control and railway communication. However, there are lots of challenges to develop and commercialize KRTCS. The biggest obstacle of implementing KRTCS is the acquisition of radio frequency. Unfortunately, only limited license frequency bands remain for public use.

700MHz spectrum bands (698 ~ 806 MHz) which are supposed to be freed up with analog to digital TV transition can be used for KRTCS [4]. Frequencies below 1 GHz are more attractive for radio based train control than those above 1 GHz. Lower path loss for greater radio coverage and lower Doppler shift for high speed train operation are the key features in 700MHz frequency. However, most country already has a plan to use this frequency band for next generation mobile communications, public safety, and ITS. New frequency allocation for railway infrastructures seems to be very competitive.

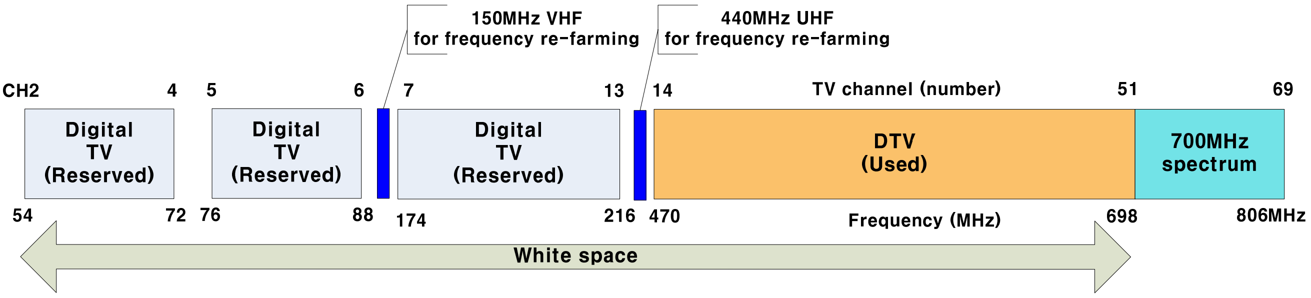


Fig. 1. Frequency candidates for KRTCS

White spaces refer to frequencies (54 ~ 698 MHz) allocated to a broadcasting service but not used locally [5]. A train control system which has very sophisticate location information can use these frequency bands. Even if white spaces have superior frequency properties and less competitive than 700 MHz, there are unlicensed spectrum allocation tendency for public usage. Because licensed spectrum provides very high reliability with minimum interferences, locally licensed white space frequency should be given for train control and communication system. Some technical difficulty also should be solved for implementing seamless frequency alternation by location information.

Exiting railroad infrastructures have their own radio frequencies for railway operation. 150 MHz, 440 MHz analog VHF and UHF bands are used for train driver and control center communication in Korea. Re-farming of the existing frequency has no need to allocate new frequency band and it can use superior frequency properties as well. However, new technologies for utilizing limited spectrum bandwidth should be needed and they increase the complexity of system and implementation cost. Frequency re-farming is heavily bounded by a very small amount of frequency bands which are currently used in Korea railway communication.

In conclusion, although there are very competitive to get frequency license, some frequency area of 700MHz spectrum bands is the most suitable candidate for KRTCS. This frequency area not only has good frequency spectral properties but also could be allocated up to 10MHz bandwidth for KRTCS. After careful estimation and analysis of required bandwidth for KRTCS, Korean railway companies and organizations have schedule to apply 700MHz frequency distribution process in 2011

3 Mobile communication candidates for KRTCS

The KRTCS technologies shall be based on international standardization frame work to enhance the economic feasibility. It should support for high data rate and high speed train operation. High data rate with low latency are needed to support high capacity network utilization and real time operation in full mobility. Many different mobile services are currently provided by various mobile telecommunication service companies in Korea. 2G mobile systems are ruled out due to the fact that it is not suitable for future railway communication requirement. Wideband Code Division Multiple Access (WCDMA), Wireless Broadband (WiBro, WiMax) and 3GPP Long Term Evolution (LTE) is 3G or beyond 3G mobile system. These systems currently provide high quality mobile services and these are strong candidate for KRTCS.

WCDMA is 3G mobile communication network which densely deploys across the nation. Existing WCDMA network can be used as fallback system when dedicated WCDMA based train control fail to operation. Meanwhile, high cost equipments (Serving GPRS Support Node (SGSN), Gateway GPRS Support Node (GGSN)) for packet transmission increase the implementation cost. Moreover, WCDMA is relatively old school telecommunication technology compare to WiBro and 3GPP LTE.

WiBro (WiMAX) is wireless broadband internet technology developed by Korea [6]. Although it offers a high data throughput and wide area of coverage with relatively low cost network equipments, the major telecom companies tend to adopt their next generation mobile system as 3GPP LTE technology.

In 2008, awareness was created on a new technology, with a promising name 3GPP LTE. This would impact GSM-R life cycle. 3GPP LTE is latest mobile network technology which evolved from WCDMA system. It provides high speed data rate (100 Mbps for downlink) and high mobility (up to 500 Km/h) over the all IP-based network. Recently, International Union of Railways (UIC) prepares to develop the LTE-Railway for next generation of GSM-R [7]. The popularity and globalization is very important for the interoperation and longtime maintenance of railroad infrastructure. However, the preparation of commercialized LTE chip sets and test equipments for developing the radio based railway communication system needs more time.

Most prominent candidate mobile system is 3GPP LTE system. Although 3GPP LTE is not fully commercialized at this moment, it satisfies future railway communication requirement and will evolve to global railway standardization. The public hearing and discussion will be scheduled for railway organizations to decide mobile system for KRTCS in 2011.

4 Spectrum bandwidth calculation

Spectrum bandwidth estimation method varies from one area to another. Determining how many frequency bands are needed is vital to frequency band request to government. In this paper, we estimate the frequency spectrum bandwidth based on ITU-R M.1390 for train communication system. ITU-R M.1390 is the general methodology to estimate spectrum bandwidth for various mobile communication services using 12 parameters [8]. These 12 parameters are analyzed to apply radio based train control system. The estimation of total data rate which are used during the train operation is important to calculate bandwidth estimation. To estimate maximum transmission data rate, we assumed that Seoul station is the highest data traffic transmission place and diameter of cell coverage is 6 Km. As a result, the data rate for automatic train control (ATC) is induced about 530 kbps. To calculate the voice communication data rate, we should consider the point to point call, group call, broadcast call and emergence call from the railway many difference communication scenarios. The total voice data rate is calculated as 224 kbps with 30 second average conversation time. 1.1 Mbps data rate is required for advanced railway data service like on-line ticketing, arrival and transit information. The surveillance video service is needed to improve the security of railway facilities and passenger monitoring. Video data rate is calculated as 2 Mbps respectively. Overall data rate is shown in table 1.

Table 1. Estimation of data rate for KRTCS

Function	Data rate	
	Up link	Down link
ATC	530 kbps	530 kbps
Voice call	224 kbps	224 kbps
Data service		1,100 kbps
Video service	2,304 kbps	2, 048 kbps
Total	3,058 kbps	3,902 kbps

One of the critical parameter for estimating bandwidth is net system capability of mobile communication system. It is a measure of how much data are transmitted to given frequency bandwidth in a specific condition of mobile communication radio cell. The net system capabilities for train control environment are induced as 0.51, 1.30, and 1.70 [bps/Hz/Cell] for WCDMA, WiBro, and LTE system respectively. Based on these parameters, we can deliver required spectrum bandwidth for establishing the KRTCS. For the highest reliability of train operation, fully duplicated network structure with overlaid radio cells is usually used in railway network planning. The duplicated network structure is shown in fig 2. As a result, total bandwidth should be double compare to single cell based system.

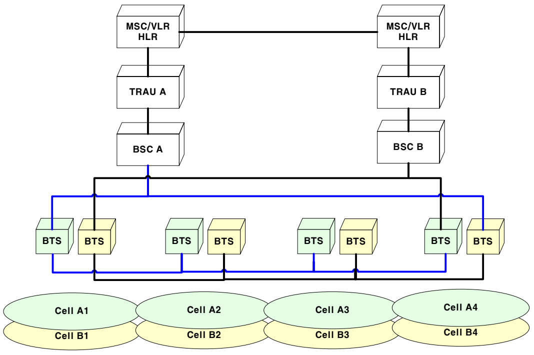


Fig. 2. Overlayed radio cells for railway communication

WCDMA is Frequency Division Duplexing (FDD) system and offers 5 MHz bandwidth for up and down link respectively. For constructing overlaid radio cell, 40 MHz bandwidth is needed for implementing KRTCS. WiBro uses Time Division Duplexing (TDD) system and has 5 MHz, 8.75 MHz, 10 MHz bandwidth profile in Korea. TDD can share the same bandwidth for up and down link connection. So, 17.5 MHz bandwidth can be used for overlaid radio cell. 3GPP LTE uses FDD and has 1.3 MHz, 3 MHz, 5 MHz, 10 MHz bandwidth profile in Korea. To implement overlaid radio cell, 12 MHz bandwidth are required for KRTCS. Consequently, the required spectrum bandwidth of WCDMA, WiBro, and LTE are shown in Table 2.

Table 2. Required spectrum bandwidth for KRTCS

Mobile System	Up link	Down link	Total (Overlayed Cell)
WCDMA	5.86 MHz	7.47 MHz	40 MHz (10 MHz x 4)
WiBro	2.30 MHz	2.93 MHz	17.5 MHz (8.75 MHz x 2)
3GPP LTE	1.70 MHz	2.24 MHz	12 MHz (3 MHz x 4)

5 Conclusions

In this paper, we describe candidates of frequency band and mobile communication system for Korean radio based train control system. Each frequency bands and mobile systems are analyzed to construct efficient train control system. Required frequency bandwidth is calculated using ITU-R recommendation under the railway environment.

The analysis results of each mobile communication system will be used to select the dedicated mobile communication system for the KRTCS. The strong candidate of frequency band and mobile system for KRTCS is 700MHz spectrum band and 3GPP LTE mobile system respectively. As a result of bandwidth estimation, 3GPP LTE mobile system requires 12 MHz bandwidth for dedicated radio based train control and railway communication. The calculation result of required bandwidth for KRTCS will be used to apply frequency bandwidth request process of 700MHz frequency distribution in 2011.

Acknowledgement

This research was supported by a grant (10PURT-B056851-01) from “Future Urban Railway Development Program” funded by Minster of Land, Transport and Maritime Affairs of Korea government.

6 References

[1] R. D. Pascoe and T. N. Eichorn, “What is communication-based train control,” IEEE Veh. Tech. Magazine, vol. 4, no. 4, pp. 16-21, 2009.

[2] C. Briso-Rodriguez, C. Cortes, F. J. Arques, and I. Alonso, “Requirements of GSM technology for the control of high speed trains,” in Proc. Pers., Indoor, Mobile Radio Comm. Conf., pp. 792-793, Feb. 2002.

[3] http://www.mltm.go.kr/USR/NEWS/m_71/dtljsp?lcmsp a ge =1&id=155702357, Oct. 2010.

[4] C. Bazelon, “Licensed or Unlicensed: the economic considerations in incremental spectrum allocations,” IEEE Communications Magazine, vol. 47, issue 3, pp. 110-116, Mar. 2009.

[5] S. Shellhammer, A. Sadek, and W Zhang, “Technical challenges for cognitive radio in TV white space spectrum,” Information Theory and Applications Workshop, pp. 323-334, Feb. 2009.

[6] M. Auado, O. Onandi, P. Agustin, M. Higuero, and E. Jacob Taquet, “Wimax on rails,” Vehicular Technology Magazine, IEEE, vol. 3, no. 3, pp. 47-56, Sep. 2008.

[7] G. Tingting, and S. Bin, “A high-speed railway mobile communication system based on LTE,” ICEIE, vol. 1, pp. 414-417, Aug. 2010.

[8] Recommendation ITU-R M.1390. “Methodology for the calculation of IMT-2000 terrestrial spectrum requirements.”

A Routing Protocol using Out-of-band Control for MANETs

Woosuck Chang[1], Chen Ni[2], Chulho Jang[2], and Chonggun Kim[2*]
[1]Department of Information & Communication Engineering, Kumi College, Kumi, Korea
[2]Department of Computer Engineering, Yeungnam University, Gyeongsan, Korea

Abstract—*Conserving energy is an essential consideration in designing routing protocols for mobile ad hoc networks. The majority of the work reported in the literature focuses on designing energy aware metrics to take place of the existing ones, and the improvements were not significant. In this paper, we propose a novel energy conserving routing protocol which use dedicated data and control channels instead of a single channel for both purposes, and the control channel, which we refer to as out-of-band channel, is used for control traffic. The proposed protocol, we refer to as ECOBR (Energy Conserving Out-of-Band Routing), can easily be applied to other routing protocols for wireless ad hoc and sensor networks. We theoretically analyze the effectiveness and efficiency of the ECOBR and give a rough condition which guarantees the effectiveness of the ECOBR. Simulation results show that ECOBR have a significant improvement in conserving energy in comparison with the AODV routing protocol when working over IEEE 802.11 MAC.*

Keywords: Energy Conserving; Out-of-Band; Ad Hoc Networks

1 Introduction

A MANET (or Mobile Ad hoc NETwork) [1] consists of a group of mobile, wireless nodes which cooperatively and spontaneously form a network independent of any fixed infrastructure or centralized administration. Accompanied with convenience brought by flexibility and scalability, energy supports for nodes in MANETs are usually limited due to their infrastructure-less nature. The energy depletion of nodes can easily lead to communication failures and network partitioning [9]. Therefore, saving energy from unnecessary use becomes a vital issue in prolonging the network lifetime.

Many efforts have been made to design routing protocols to save energy and elongate network lifetime for MANETs, however, the enhancements are usually limited. In [9,13], D. Kim and his co-researchers proposed a new metric, the energy drain rate, to be used to predict the lifetime of nodes according to current traffic conditions. Combined with the value of the remaining battery capacity, the metric is used to establish whether or not a node can be part of an active route and in turn avoids situations in which a few nodes allow too much traffic to pass through them. Reference [9] also mentioned that the overhearing problem results in the similar behaviors of all energy-aware routing protocols and prohibits performance enhancement.

IEEE 802.11 [2] is one of the dominating MAC protocols for ad hoc networks and is used by many proposed routing protocols as their underlying MAC protocol. Ad hoc mode operation defined in IEEE 802.11 MAC does not use any infrastructure: nodes communicate directly with all othe nodes that are in wireless transmission range. Because there are no base stations to moderate communication, nodes must always be ready to receive traffic from their neighbors. A network interface operating in ad hoc mode does not sleep [8]. Therefore, besides the energy consumed by routing control overhead and data transmission, overhearing is another significant energy drain in MANETs. In ad hoc networks, a transmission from one node to another is potentially overheard by all the neighbors of the transmitting node – thus all of these nodes consume power even though the data transmission was not directed to them**.**

Recent trends of reducing hardware cost and technique development have made it feasible to equip nodes with multiple interfaces and reduce the energy consumption of switching the radio into off state (sleep state). Multiple channels and multiple interfaces have been used in some research in MANET area. In [4], S. Singh and his coworker proposed a power-aware multiple access protocol (PAMAS) using a separate control channel, in which a node turns off its radio interface for a specific duration of time, when it knows that it will not be able to send and/or receive packets during that time due to the possibility of multiple access interference. However, such fine grained control of on-state of the radio interface at MAC layer requires fast transition between on and off state (within a millisecond in 802.11b, e.g.), which is far from realistic in the state-of-art commercial radios for high speed networks. Multi-channel has also been used in [5] and [7] separately to increase throughput and network capacity.

Because IEEE 802.11 has already been standardized and implemented by many commercial radios，it is not likely to be taken place by other MAC protocols in ad hoc networks for general use in near future. In this paper, we propose an energy conserving routing protocol which can conserve a significant amount of energy by largely alleviating the energy drain on overhearing in IEEE 802.11 based ad hoc

* Correspondence Author

networks. The proposed protocol is referred to as ECOBR (Energy Conserving Out-of-Band Routing). The ECOBR uses dedicated data and control channels instead of a single channel, and the control channel is used for control traffic (control packets at network layer). By separating the control traffic from data communication channel, the data communication radios on nodes that are not involved in any transmission can be turned off thus can save energy that would otherwise be wasted in overhearing unintended traffic. Note that network layer control of the on-off state of radios can greatly reduce the number of state switches between on and off in comparison with the MAC layer control. Following this instruction, the proposed ECOBR routing protocol is described in detail in Section II. Section III presents the simulation results showing that ECOBR can conserve a significant amount of energy in ad hoc networks when working over IEEE 802.11. Effectiveness and efficiency analysis are presented in Section 4. Finally, we conclude this work in Section 5.

2 Energy Conserving Out-of-band Routing

2.1 Background

2.1.1 Ad-hoc On-demand Distance Vector (AODV) Routing Protocol

Ad-hoc On-demand Distance Vector (AODV) [3] routing is one the most representative routing protocols for ad hoc networks. In AODV, when a source wants to send data to a destination and no route is available in its routing table, it initiates a route discovery process by flooding a route request (RREQ) for destination throughout the network. A RREQ packet is uniquely identified by a sequence number so that duplicate RREQs can be recognized and discarded. Upon receiving a non-duplicated RREQ, an intermediate node records previous hop and check whether there is a valid route entry to the destination existing in route table locally. If it is the case, the node sends back route reply (RREP) to the source; otherwise, it rebroadcasts the RREQ. When the destination receives the route request, it will reply with a RREP to the source. As the RREP traverses though the established reverse path, each node along the path set up a forward pointer, update corresponding timeout information and records latest destination sequence number (for the freshness checking by intermediate node to determine whether sending RREP or not). A node updates routing information and propagates the RREQ upon receiving further RREPs only if RREP contains either a greater destination sequence number (fresher) or a shorter route found. After receiving the RREP at the source node, a path to the destination is available and data transmission can begin.

For the path maintenance part, route failure can be detected by link layer feedback. When a route failure is detected, route error (RERR) packets are sent back to all sources to erase route entries using the failed link. A route discovery procedure will be issued if the route is still needed.

Note that all the processes mentioned above are done in a single channel. With only one channel in use, nodes cannot turn off their radios because they should always be ready to receive potential traffic from their neighbors, and therefore, waste lots of energy in overhearing unintended traffic.

2.1.2 Out-of-band Channel

An out-of-band channel is typically used to separate control traffic from data traffic. In this paper, we use two network interface cards to allow the using of two channels. The newly added out-of-band channel is used for interchanging control packets in network layer.

2.2 Overview of ECOBR

2.2.1 Networking Model

The ECOBR protocol is a combination of the original AODV and the idea of using an out-of-band channel. The single channel networking model used by AODV is shown in figure 1 in which both control and data traffic are going through the same protocol stack and the only channel. In contract, the ECOBR uses the networking model given in figure 2. Newly added parts are represented by gray boxes, which allow the control packets exchange of routing protocol to take place over the out-of-band channel that is separated from the channel used for data packet transmissions. ECOBR can operate over two channel dedicated to data and control traffic independently. We assume the radio for control channel is always on while that for data communication can be turned off when no route is using it. So in what follows, when we mention about radio, it represents the radio used for data communication.

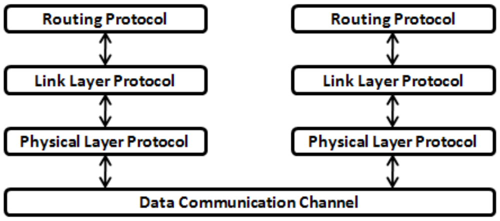


Figure 1. Single Channel Networking Model

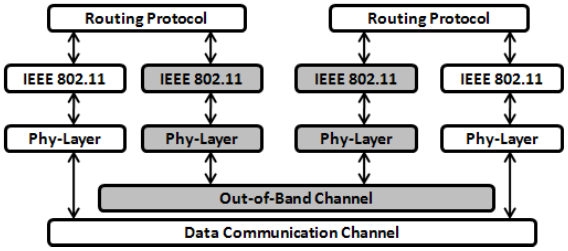


Figure 2. Double Channels Networking Model used by ECOBR

2.2.2 Route discovery

The route discovery process is done in the out-of-band channel follows the same procedures in the AODV. To enable coordinating the work in two channels, a few changes have to be made in original AODV route discovery.

We give a simple example to illustrate the working process and how it can control the on-off state of nodes based on network layer decision. As shown in figure 3a, 9 nodes are evenly distributed in a square area and their radios are turned off (represented by gray circles) at initial state. Here we assume that the transmission range of each node can only cover its one-hop neighbors; for instance, in figure 3, node 1 can only communicate directly with nodes 2, 4, and 5.

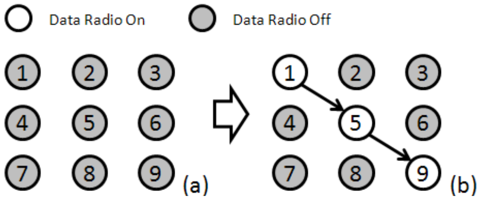


Figure 3. Radio state transition in route discovery

When node 1 wants to transmit data to node 9 and no route is available in its routing table, it broadcast the route request in the control channel and all the intermediate nodes follows the operation specified in AODV. When node 9 receives the route request, it turns on its radio and sends back a route reply. Upon receiving a route reply, a node turns its radio on and waits for the coming data transmission. As shown in figure 3b, after the route reply reaches node 1, a forward path with all the nodes along it turned on is available and data transmission can begin. And in this case, all the nodes that are not on the path are turned off thus will not expend energy on overhearing.

Please note that in contrast with the MAC layer operation of on-off state of nodes, in which the interval between a node being turned off and then turned on is counted by milliseconds, our scheme controls the on-off state of radios based on routing strategies, and the interval is generally in the order of the lifetime of a route. Not only can our scheme significantly reduce the number of times to switch the on-off state of nodes (which can save energy expended on switching state), it is much easier to achieve in real commercial products as well.

2.2.3 Route Maintenance

The route maintenance process is also done in the out-of-band channel. According to the specification of AODV routing protocol, route error packets will be sent once a path failure occurs. Differing from the single channel case, the proposed protocol requires a change that a node turns off its radio upon receiving a RRER packet if there is no other communication session (or has no available route in its routing table) using it. The downstream nodes of the failed link will turn their radio off, after being idle for a period of time (See Section 2.2.4 for explanation), The transition of radio state in aforementioned example when a link failure occurs is given in figure 4. When the link between nodes 5 and 9 fails (figure 4a), route error is send back to erase routes in upstream nodes and turn their radio off (figure 4b). And node 9 will turn itself off (figure 4c), after waiting for a period of time and no data are transmitted to it.

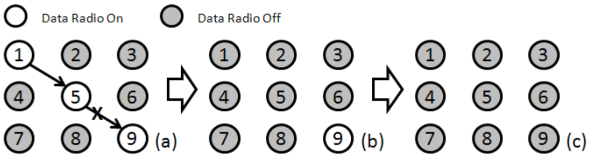


Figure 4. Radio state transition after a link failure

2.2.4 Radio State Transition in ECOBR

A key issue for ECOBR is the on-off state control of data communication radio. The state diagram outlining the behavior of the radio state in ECOBR is given in figure 5. The ACTIVE_ROUTE_TIMEOUT is a constant used by routing table to purge stale routes. Routing table will set a path to invalid after being idle for this period of time. According to [3], this value is set to 3000 milliseconds by default, but we use 10,000 milliseconds in our simulation to fit the network environment.

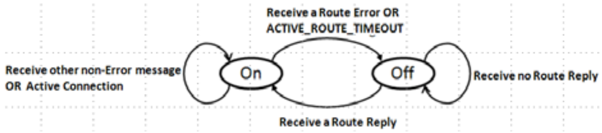


Figure 5. Radio on-off state diagram

During the route discovery process, a node switches its radio on when receives a route reply from control channel. And when it receives a route error message or does not receives any message using the current route within ACTIVE_ROUTE_TIMEOUT, the node checks if there is any other connection using it; if not, it will turn the radio off; or otherwise, keep the radio on since it still has to serve the other connection session. We use ACTIVE_ROUTE_TIMEOUT here because in the AODV if a node does not receive any message using the current path within this period of time, it will invalidate this path in its routing table; so even in original AODV, when next time a data transmission needs to use this node for the same route, it has to start a route request again, and in ECOBR case, its radio can be switched on again upon receiving the route reply.

3 Simulations

The goal of this section is to investigate the effect of the proposed ECOBR routing protocol on energy conserving. We implement the proposed routing protocol in Ns2 [12] and compare the performance of ECOBR with that of AODV when both are working over IEEE 802.11 MAC protocol. We assume both of the data communication channel and the control channel have the same characteristics. We run simulations over a variety of scenarios and see a significant improvement in conserving energy.

3.1 Simulation Settings

In this section, we list the key parameters used in our simulation, these parameters apply to all the following scenarios if without explicative declaration.

3.1.1 Energy and Radio Parameters

Important energy and radio parameters used in our experiment are listed in Table I. Referencing to [4, 11], the radio we use consumes 15 watts during while transmitting, 11watts while receiving and 50mW in idle mode.

TABLE I. KEY ENERGY AND EADIO PARAMETERS

Sending Power	15 watts
Receiving Power	11 watts
Idling Power	0.05 watt
Initial Energy	1000 Joules

3.1.2 Traffic Parameters

Traffic parameters used in this work are listed in Table II.

TABLE II. TRAFFIC PARAMETERS

Packet Size	512 bytes
Packet Type	UDP, cbr
Sending Rate	4 packets / second
Traffic pattern	Uniform

3.2 Simulation Results

3.2.1 Line Topology

A line topology of 5 nodes is used to evaluate the ECOBR and AODV is shown in figure 6. Node 0 serves as the source node and sends traffic to Node 4. We use line topology to investigate the scenario in which no other nodes will overhear the transmission excluding the nodes on the path. Hence, line topology is expected to make minimum difference between the proposed ECOBR and the original AODV. This simulation runs 300 seconds.

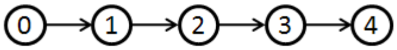


Figure 6. Line Topology

The residual energy of each node after running ECOBR and AODV over the given topology is shown in Table III. As we expected, the energy conserving performances of the two protocols are almost the same. The slightly higher residual energy in ECOBR is resulted by separating the traffic into two channels, and hence reduces the contention in data channel.

TABLE III. RESIDUAL ENERGY IN LINE TOPOLOGY

	Node 0	*Node 1*	*Node 2*	*Node 3*	*Node 4*
AODV	670.17	659.23	659.23	659.23	750.19
ECOBR	671.54	660.62	660.62	660.62	750.78

3.2.2 Lattice Topology

As shown in figure 7, the lattice topology used in our experiment consists of 49 static nodes evenly distributed in a square area size of 1000m x 1000m. Here, we use lattice topology to test the static regular scenario. First, connections 1, 2, and 3 are used to generate traffic as given in the figure. The simulation runs 150 seconds. Then we run another simulation by adding connection 4, 5, and 6 in the same topology.

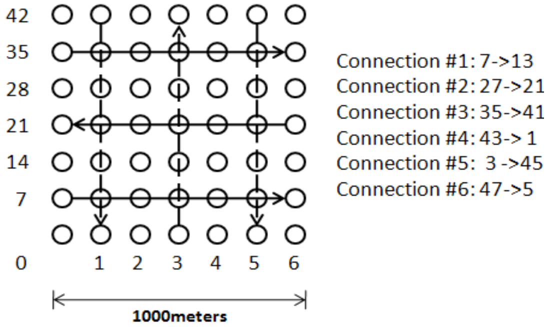


Figure 7. Static Lattice topology

Total consumed energy information of two scenarios after running simulation is given in Table IV. Note that few nodes are died in the tested scenario. We can see in the scenario with 3 connections, the ECOBR consumes 43% energy of that consumed by AODV; and in the 6 connections case, the ECOBR consumes 69% energy of that consumed by AODV. This difference is mainly resulted by the situation that more nodes are involved in data communication in the scenario with 6 connections hence fewer nodes can be prevented from expending energy in overhearing.

TABLE IV. TOTOAL CONSUMED ENERGY IN LATTICE TOPOLOGY

	3 connections	*6 connections*
AODV	19247	33888
ECOBR	8402	23464

3.2.3 Dynamic Random Topology

In the dynamic random topology, the initial positions of mobile nodes are randomly selected within an area of 1500m x 300m; the “random way-point” model is used to simulate nodes movement. These mobile nodes move with speeds uniformly selected between 0m/s and 10m/s to their randomly chosen destinations. Each node stops for 50 seconds after arriving at its destination before moving again. We vary the node density, traffic rate, and the number of connections in our experiment to see their impact on energy conserving performance of the proposed protocol. Each of the following simulations runs 300 seconds.

Figure 8 shows the results generated by varying node density—the number of nodes within the test area. The number of connection sessions is fixed as 3. In 20 nodes case, the ECOBR consumes only 25% energy of that consumed by the AODV. In the following cases, the energy consumed by the AODV increases drastically as the number of nodes increases while that of ECOBR stays low and increases with a very small gradient. This is largely contributed by the increase of nodes that are not involved in the data communication but expend energy in overhearing the unintended traffic.

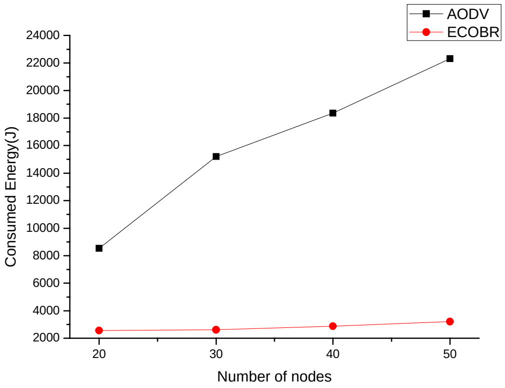


Figure 8. Consumed Energy vs. Number of Nodes

We fix the number of nodes as 30 and vary the traffic rate from 2 to 8 packets/second. The results of consumed energy by each case are given in figure 9. We can observe the phenomenon that in both AODV and ECOBR cases, the consumed energy values are proportional to the traffic rate. However, in AODV case, the consumed energy is obviously much higher than that in ECOBR.

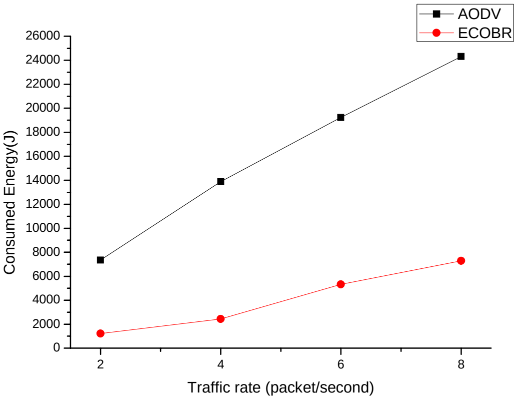


Figure 9. Consumed Energy vs. Traffic Rate

The number of nodes is also set to 30 for this scenario and we vary the number of connection sessions from 1 to 5. This time, we see a different phenomenon in figure 10 in comparison with previous ones—the energy consumed by ECOBR increases much faster than aforementioned cases. But if we take a look at the scenario given in lattice topology section, we may easily come up with the answer: fewer nodes can be saved from expending energy in overhearing as the number of connections increases.

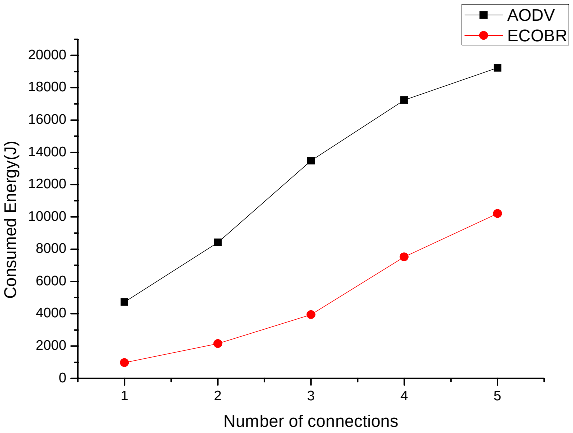


Figure 10. Consumed Energy vs. Number of connections

4 Effectiveness and Efficiency Analysis

In this section, we explain the energy consuming behaviors in wireless ad hoc networks and present an analysis on effectiveness and efficiency of ECOBR by taking major energy drains into consideration. The key issue that affects the effectiveness and efficiency of the proposed ECOBR is the energy cost introduced by the control channel. It is true that ECOBR will cost more energy than single-channel case when using network interfaces which consume significant amount of energy in their idling state. So, we try to derive a rough condition which takes network environment into account and guarantees a better energy conserving performance in our proposed networking model

4.1 Energy Consuming Behaviors

From the perspective of the network layer, the possible behaviors that consume energy include:

1) Transmitting, Energy consumption is P_{trans} *per second.*

2) Receiving and overhearing, Energy consumption is P_{recv} *per second.*

3) Idling, Energy consumption is P_{idle} *per second.*

4.2 Network Characteristics

We define the following variables to represent network characteristics:

1) Density of nodes: represented by d nodes/m^2*, this affects the number of nodes overhearing a transmission.*

2) Affect Range: represented by r meters, this variable is typically related to the transmitting power and *transmission range, within which all the nodes will expend energy in overhearing the transmission.*

3) time: represented by ti seconds, each i shows one connection.

4) Number of connections: represented by N.

4.3 Comparision between ECOBR and single-channel routing

Assuming the costs of control traffic and necessary transmitting and receiving of data traffic are the same in both cases. Then, effectiveness and efficiency of ECOBR are mainly affected by the comparison between the energy used in idle state in the out-of-band channel and the energy saved from being overhearing unintended traffic in data channel. These two energy value can roughly be estimated as:

$$E_{idle} = \sum_{i=1}^{n} P_{idle} \times ti \qquad (1)$$

$$E_{save} = \sum_{i=1}^{n} P_{recv} \times ti\,(N \times (d\pi r^2 - ci)) \qquad (2)$$

where t is the simulation time measured in second and c represents the number of nodes which are within the *Affect Range* and on active paths(thus will not be turned off and will still expend energy on overhearing) as well. And note that values d, r and N positively contribute to c. From the above discussions, the efficiency can be represented by E_{save}/E_{idle}, and if it is larger than 1, we say ECOBR is effective in conserving energy. Let $E_{save} > E_{idle}$, we have:

$$\frac{P_{resv}}{P_{idle}} > \frac{1}{\sum_{i=1}^{n} N \times (d\pi r^2 - ci)} \qquad (3)$$

In (3), the large values of N, d, and r will contribute positively to making the expression satisfied; hence these factors have positive impact on the efficiency of the proposed ECOBR. As we can see, even in the situation that the difference between P_{recv} and P_{idle} is not significant, our proposal is still effective in network environments that allow the satisfying of (3).

5 Conclusion

In this paper, we propose a novel energy conserving routing protocol for mobile ad hoc networks named ECOBR. The ECOBR is implemented by extending original AODV and differs from traditional routing protocol by operating the route control processes in an out-of-band channel thus allows the on-off state controlling of data communication radios. Properly turning off radios on nodes that are not involved in data communication can largely prevent wasting energy on overhearing unintended traffic. Theoretical analysis gives a rough condition which guarantees the effectiveness of the ECOBR. Simulation result show that, compared against single-channel routing protocol, ECOBR can conserve a significant amount of energy.

6 Acknowledgement

This research was supported by yeungnam university research grants in 2010.

7 References

[1] J. Macker and I. Chakeres, Mobile Ad-hoc Networks (manet), https://www.ietf.org/html.charters/manet-charter.html, IETF Working Group Charter, 2008.

[2] IEEE standards Department, Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE standard 802.11-1997, 1997.

[3] C. E. Perkins, E. M. Belding-Royer, and I. Chakeres. "Ad Hoc On Demand Distance Vector (AODV) Routing." IETF Internet draft, draft-perkins-manet-aodvbis-00.txt, Oct 2003 (Work in Progress).

[4] S. Singh, C.S. Raghavendra, Power Aware Multi-Access protocol with Signalling for Ad hoc Networks, ACM SIGCOMM Computer Communication Review, Volume 28 , Issue 3, pp. 5-26, July 1998.

[5] E. Shim, S. Baek, J. Kim, D. Kim, Multi-Channel Multi-Interface MAC Protocol in Wireless Ad Hoc Networks, Proceedings of ICC 2008, pp.2448-2453, May 2008.

[6] B.S. Manoj, C. Siva Ram Murthy, On the use of out-of-band signaling in ad hoc wireless networks, Computer Communication, Vol. 26, pp.1405-1414, 2003.

[7] P. Kyasanur, N.H. Vaidya, Routing and Link-layer Protocols for Multi-Channel Multi-interface Ad Hoc Wireless Networks, ACM SIGCOMM Mibile Computing and Communication Review, Vol. 10, issue 1, pp.31-43, Jan. 2006.

[8] L.M. Feeney, M. Nilsson, Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment, Proceedings of INFOCOM 2001, Vol. 3, pp.1548-1557, 2001.

[9] D. Kim, J.J. Garcia-Luna-Aceves, K. Obraczka, J.C. Cano, and P. Manzoni, Routing Mechanisms for Mobile Ad Hoc Networks Based on the Energy Drain Rate, IEEE Transactions on Mobile Computing, Vol.2, No.2, pp.161-173, April-June 2003.

[10] J. Deng, B. Liang, P. K. Varshney, Tunning the Carrier Sensing Range of IEEE 802.11 MAC, GLOBECOM 2004, pp.2987-2991, 2004.

[11] J.J. Garcia-Luna-Aceves, Chane L. Fullmer and Ewerton Madruga, “Wireless Mobile Internetworking”, Manuscript.

[12] The Network Simulator ns-2, http://www.isi.edu/nsnam, 2006.

[13] Mary Wu, C.kim, A cost matrix agent for shortest path routing in ad hoc networks, JNCA 33, pp.646-652, 2010

An Energy-efficient Coverage Pattern of WSNs for High Rate Data Transmissions

Manh Thuong Quan Dao[1], Ngoc Duy Nguyen[1], Vyacheslav Zalyubovskiy[2], and Hyunseung Choo[1]
[1]School of Information and Communication Engineering, Sungkyunkwan University, Korea
[2]Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Russia
Email: dmtquan@skku.edu, duyngoc30@skku.edu, slava@math.nsc.ru, choo@skku.edu

Abstract— *The coverage problem is a fundamental issue in wireless sensor networks (WSNs). Recent coverage patterns only utilize the adjustable ranges of sensors to minimize the sensing energy consumption. However, a big amount of energy consumption for communication is not strictly taken into account, especially for high rate data transmission applications. In this paper, we introduce an energy-efficient pattern that is used to minimize the communication energy consumption for high rate data transmission WSNs, while simultaneously, guarantee a high degree of sensing coverage. Theoretical calculations and extensive simulation are conducted to evaluate the efficiency of the new pattern compared to existing ones in terms of various performance metrics. Our proposed pattern can minimize the total energy consumption at most 17.4% compared to the existing patterns, which were known as the best patterns.*

Keywords: coverage problem, deployment patterns, high rate data transmission, wireless sensor networks.

1. Introduction

Modern technology has enabled a new generation of wireless sensor networks (WSNs) that are feasible for a wide range of commercial and military applications. A wireless sensor network is composed of a large number of autonomous sensors that are densely deployed into a target sensing field to monitor physical phenomena of interest [1]. Replenishing power resources is a difficult and impossible task in most cases, since each sensor has a limited power battery. Energy-saving optimization is an important criteria to evaluate the success of WSNs. Recent analysis [4] shows that each sensor spends a larger portion of power for a communication task than for a sensing task. The power usage for WINS Rockwell seismic sensor for completing a data transmission task and a sensing task are 0.7W and 0.02W, respectively. In low rate data transmission application, it is reasonable to concentrate on minimizing the sensing energy consumption since sensors perform sensing task throughout their lifetime, and rarely transmit the data. However, in high rate data transmission applications, the communication energy consumption is drastically higher than the sensing energy consumption since sensors may need to transmit the real-time tracking data continuously [8], [10]. Therefore, this paper considers the problem of how to minimize the communication energy consumption by adopting a strategic coverage pattern.

A fundamental issue in WSNs is the coverage problem, which concerns how well the target sensing field is monitored or tracked by sensors. In this paper, we consider the full coverage problem in the sense that every point in the target sensing area is covered by at least one sensor. On the other hand, there are two mechanisms in sensor deployment: deterministic deployment and random deployment. In deterministic deployment, a sensor can be placed exactly at a pre-defined position in the target sensing field. In contrast, the position of a sensor is not known a priori in random deployment. Our proposed strategic pattern is useful in deterministic deployment as well as in random deployment. Another important consideration in WSNs is the node scheduling problem, as it has a significant impact on extending the network lifetime. A node scheduling mechanism operates in such a way that a set of active sensor nodes is selected to work in a round. Another set is selected in the next round, as long as the coverage goal is met [12]. The results of this paper can be used as a guideline to select active sensors in each round that two conflicting goals are satisfied simultaneously: minimizing the energy consumption and keeping a high degree of coverage.

As shown in [4], power consumption can be divided into three domains: sensing, communication, and data processing. Recent coverage patterns only consider how to minimize the overlapped sensing areas of sensors, and thus optimize the sensing energy consumption of WSNs. In [3], [12], [14], the authors utilize the adjustable sensing range of sensors to achieve significant improvement in coverage efficiency. Recently, the authors in [14] proposed the optimal sensing energy patterns using two adjustable sensing ranges. These patterns outperform the existing ones with respect to various performance metrics. However, none of the previous work considered patterns that minimize the communication energy consumption. A huge amount of energy is spent on data transmissions in applications that require high rate data collection in real-time [8], [10]. This problem becomes critical since sensors cannot replenish their power, thus the network lifetime is decreased significantly. For this reason, communication energy consumption should be taken into account when designing energy-efficient coverage patterns.

In this paper, we propose a coverage pattern that concentrate on minimizing the communication energy consumption. Since data transmissions are the most important source for sensors' energy depleting in high rate data transmission WSNs. In summary, this paper makes the following key contributions:

- A novel coverage pattern is constructed so that it can be used in deterministic deployment as well as in random deployment. In deterministic deployment, the pattern becomes a strategic plan to design efficient WSNs. In cases where sensors are randomly deployed, our proposed pattern aids a node scheduling mechanism to select active nodes in each round.
- The proposed pattern has a structure that is easy to implement and design due to their simplicity. Moreover,all sensors in our proposed pattern only use a uniform sensing range.
- To the best of our knowledge, when considering an energy-efficient pattern for the coverage problem, we provided the best pattern in terms of communication energy consumption.

The remainder of this paper is organized as follows. In section 2, we discuss related work. Section 3 presents the system model and assumptions while Section 4 presents our proposed pattern. Section 5 shows the performance evaluation result. Finally, we conclude our work in section 6.

2. Related Work

A survey on the energy-efficient coverage problem is researched by Cardei and Wui [4]. The paper summarizes various problems on coverage area as well as their corresponding solutions. One of the mechanisms to reduce the redundant energy is using a node scheduling strategy. In this strategy, the network is scheduled to operate in turn, in the sense that one set of sensors is selected to monitor fully the entire target sensing field, and another set will be selected at another time, after the current set of sensors goes into a dormant state. Coverage ratio is one of the measurements of the system's quality of service. It plays an important role in evaluating whether or not a WSN topology is good. There is always a lower bound of coverage ratio. If the coverage ratio falls below this threshold, the network may not operate correctly. Therefore, the major challenge for the success of WSNs is designing an energy-efficient pattern that provides a high degree of coverage.

Another approach to extend the network lifetime is using sensors with adjustable ranges. In [12], two coverage patterns were proposed to reduce the sensing energy consumption of WSNs. The authors constructed patterns based on regular polygon-tiles that cover the entire target sensing field without overlap. They proposed an efficient pattern which was based on the regular triangle tile as shown in Fig. 1. However, the authors did not consider the adjustable sensing range of sensors. In [14], the authors introduced the concept of coverage density that was used as a standard metric to evaluate the efficiency of a pattern in terms of coverage efficiency. Then, they proposed the optimal sensing energy consumption patterns, using two adjustable sensing ranges for the triangle tile, called pattern A, and the square tile, called pattern B as shown in Fig. 2. The ratio between the large disk's sensing range and small disk's sensing range of pattern A and pattern B are $\sqrt{31}$ and $\sqrt{5}$, respectively. These patterns have been shown to outperform prior ones with respect to various performance metrics. Similar to [3], [12], and [14], we consider the problem of energy-efficient area coverage patterns. However, this paper supplements the important limitation of previous studies by introducing a pattern that is considered to be the best among the existing ones in terms of communication energy consumption.

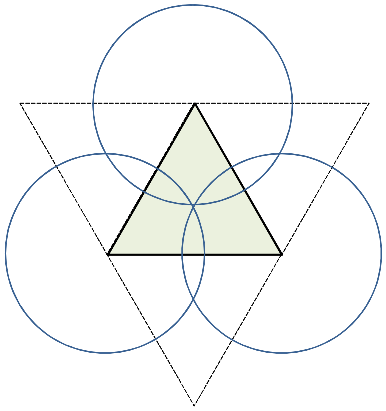
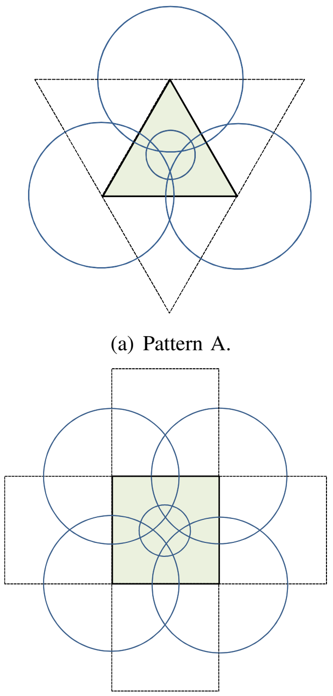


Fig. 1: Uniform Sensing Range Pattern based on Regular Triangle Tile.

The analysis of the power usage for the WINS Rockwell seismic sensor shows that the power usage for communication is between 0.74 W and 1.06 W, for the idle state it is 0.34 W, for the sleep state 0.03 W, and for the sensing task 0.02 W [9]. Obviously, the communication energy consumption is much higher than the sensing energy consumption. Recently, more applications such as monitoring industrial processes, geophysical environments, and civil structures (buildings, bridges, *etc*.), require high-data rate signals [8], [10]. A key challenge in those applications is how to collect efficiently those fidelity data subject to limited radio bandwidth and the battery of sensors. Therefore, finding an energy-efficient pattern in terms of communication energy consumption is crucial in such applications.

In [2], Bai *et al.* propose deployment patterns to achieve full coverage with three-connectivity and full coverage with five-connectivity, under different ratios of sensor communication range over the sensing range for WSNs. The authors in [15] and [13] consider the *k-coverage* problem with



(a) Pattern A.

(b) Pattern B.

Fig. 2: Optimal coverage patterns with two adjustable sensing range.

any arbitrary sensing shape and find the weak sub-regions degrading the overall coverage performance. Various related coverage problems have been discovered recently. Examples include the coverage problem in three dimensional space [6], coverage for estimating localization error [11], and barrier coverage problem [13].

3. System Model and Problem Formulation

3.1 System Model

In this paper, we assume that sensor nodes are randomly deployed in a two-dimensional target sensing field, where each node uses a Global Positioning System or a localization scheme to know its position. The sensing area of each sensor is a disk of a given sensing range. The sensors are in charge of monitoring a target sensing field which is assumed to be very large compared to the sensing area of a sensor, and thus we can ignore the boundary effect of the target sensing field.

The sensing range and the transmission range of a sensor in a coverage pattern is illustrated in Fig. 3. The transmission range (R_C) is usually larger than the sensing range (R_S) to guarantee the connectivity between sensor nodes. Two nodes can communicate only when they are in the transmission range of each other. The total area inside the sensing range and transmission range are called the sensing disk and communication disk, respectively.

We construct a minimal spanning tree among active nodes when calculating the communication energy consumption. Each node adjusts its communication range to the farthest node on the tree to guarantee network connectivity. We also assume that the energy consumed by communication for a sensor is proportional to the square of the distance from itself to its farthest node in the minimum spanning tree when calculating the theoretical energy consumption.

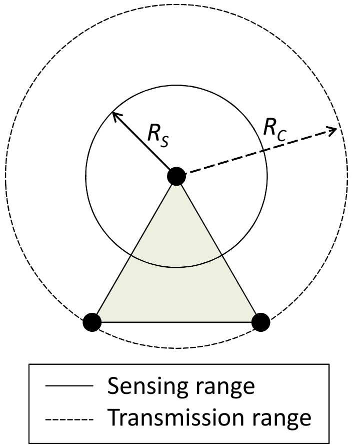


Fig. 3: Sensing Range and Transmission Range.

Finally, since our sensor deployment is random deployment, we may not find a sensor that has the exact location in the pattern. In this case, we select a sensor that is closest to the ideal position in the corresponding pattern. Similar to previous studies, we construct the patterns based on a regular polygon-tile that cover the whole target sensing field without overlap. We also suppose that all tiles are covered in the same manner. Sensors are placed at the vertices of the polygons and the circles represent the sensing areas or the communication areas of sensors.

3.2 Important Definitions

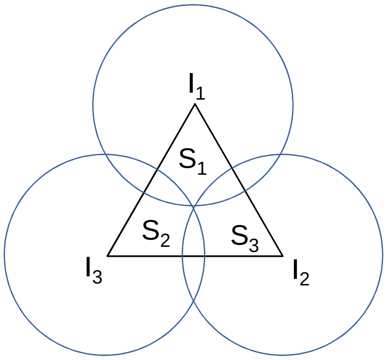


Fig. 4: Coverage Density.

Before going into the details of our proposed pattern, we introduce three important metrics that are used in [4] and [14] to compare the efficiency among patterns in terms of coverage density and communication energy consumption.

Definition 1. Coverage density (D) is the ratio of the total area of the parts of sensing disks inside the tile divided by the area of the tile. Given a coverage pattern as in Fig. 4, the coverage density D is calculated as $D = \frac{S_1+S_2+S_3}{S_{I_1 I_2 I_3}}$, where S_1, S_2, and S_3 denote the areas of parts of sensing disks of sensors I_1, I_2, and I_3 inside the tile, respectively; $S_{I_1 I_2 I_3}$ denotes the area of the triangle tile.

Definition 2. Sensing energy consumption per area (SECPA) is the part of the sensors' sensing energy used by the nodes inside a tile divided by the tile's area. We suppose that the sensing energy consumption is proportional to the area of sensing disks by a factor of μ_1, or the power consumption per unit. Then, SECPA is $SE = D.\mu_1$

Definition 3. Similar to SECPA, communication energy consumption per area (CECPA) is the part of the sensors' communication energy used by the nodes inside a tile divided by the tile's area. When calculating CECPA, we replace the sensors' sensing disks with the sensors' communication disks and calculate in the same manner with SECPA.

3.3 Problem Formulation

We now give a formal description of the energy-efficient coverage pattern formulation. We assume that there has been n nodes deployed in the working area. Similar to [16], the total energy consumption of each sensor is calculated as follows:

$$E = kS(R)^x + (1-k)T(R)^y + C,$$

where $S(R)$ and $T(R)$ denote the sensing range and communication range of a sensor, respectively; x and y are constants between 2 and 4; and k is a constant, such that $0 \leq k \leq 1$. The energy consumed by the idling radio and processor of each sensor is a constant, C.

Given the ratio k between the sensing tasks and the communication tasks, find the coverage pattern or the set of nodes to work in one active period that minimizes the total energy consumption E of all sensors in this set. The coverage pattern should satisfy the sensing coverage constraint which follows that every point in the working area is covered by at least one sensor. It also means that every point in the working area must be inside the sensing range of at least one sensor.

4. Proposed Pattern

As presented in [15], the uniform sensing range pattern based on the triangle tile is the optimal topology, in the sense that it provides the minimum number of sensors used to cover fully the entire target sensing field. However, the communication energy consumption of the uniform sensing range pattern based on the triangle tile is considerably high. Therefore, to retain the advantages of this pattern, and simultaneously improve the communication energy consumption, we construct another pattern, called pattern C, based on a hexagon tile, as shown in Fig. 5.

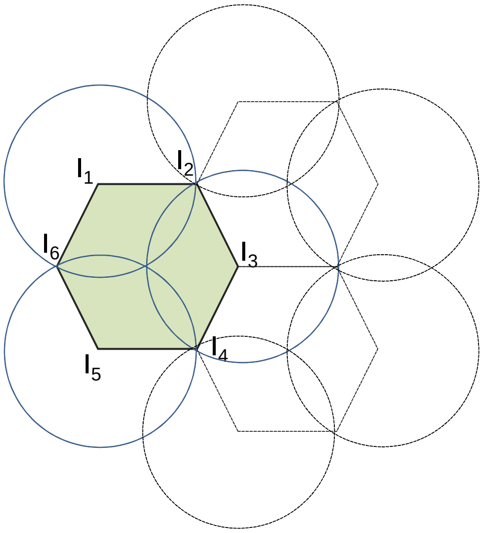


Fig. 5: Pattern C.

As opposed to the previous patterns, we assign different tasks to sensors. In Fig. 5, only three sensors placed at I_1, I_3, and I_5 retain both roles: sensing the monitored field and continuously communicating with other sensors. On the other hand, sensors placed at I_2, I_4, and I_6 only need to turn on their communication ability. The main reason for this strategy is that the topology of all sensors retaining both roles is the same with the topology of the uniform sensing range pattern based on the triangle tile. Therefore, the coverage density and the sensing energy consumption of pattern C are equal to the coverage density and the sensing energy consumption of the uniform sensing range pattern based on the triangle tile, respectively. Due to the symmetry, sensors in pattern C can take turns to switch on/off their sensing ability. For example, in the first round, sensors at I_2, I_4, and I_6 turn on their sensing ability, while sensors at I_1, I_3, and I_5 turn off their sensing ability. In the next round, sensors placed at I_2, I_4, and I_6 take on the sensing responsibility, while sensors at I_1, I_3, and I_5 can safely turn the function off. Therefore, the energy consumption is more well-balanced between sensor nodes.

We apply the same method as in [14] to calculate the coverage density D and sensing energy consumption per area. For pattern C, the coverage density and sensing energy consumption are:

$$D_C = 2\Pi/3\sqrt{3} \approx 1.2091$$

$$SE_C \approx 1.2091\mu_1,$$

where μ_1 is the sensing power consumption per unit.

We assume that all sensors are involved in communication. To calculate the communication energy consumption of this pattern, we construct a minimal spanning tree that spans all the sensors. We assume that the energy consumed by communication for a sensor is proportional to the square of the distance from itself to its farthest neighbor on the tree by a factor of μ_2, where μ_2 is the communication power consumption per unit. We ignore the edge effect and calculate CECPA for the case of an infinite grid, as for the estimation of coverage density. The communication energy consumed by each node in each rectangle is $\frac{1}{3}I_1{I_2}^2\mu_2$, since each node contributes to three hexagons. Finally, the CECPA of pattern C is calculated as follows:

$$CE_C = \frac{\frac{6}{3}I_1{I_2}^2\mu_2}{S_{I_1I_2I_3I_4I_5I_6}} \approx 0.7698\mu_2,$$

where $S_{I_1I_2I_3I_4I_5I_6}$ denotes the area of the hexagon tile.

Since patterns A and B outperform the prior ones in terms of sensing energy and communication energy consumption. In this paper, we only need to compare the energy consumption of our proposed pattern with these patterns. Table 1 summarizes the SECPA and CECPA of three patterns. We can obviously see that patterns A and B have the lowest value of sensing energy consumption, whereas pattern C is the best pattern in terms of communication energy consumption. In the next section, we will compare the energy efficiency of all patterns with extensive simulation.

Table 1: Energy consumption per area for different patterns

Type	*SECPA*	*CECPA*
Pattern A	$1.10\mu_1$	$1.15\mu_2$
Pattern B	$1.17\mu_1$	μ_2
Pattern C	$1.20\mu_1$	$0.7698\mu_2$

5. Performance Evaluation

5.1 Simulation Environment

To evaluate the efficiency of our new pattern (pattern C), we use the same simulation environment as in [12], [14]. We randomly deployed 1000 sensors in a $50m \times 50m$ area. The sensing range of the large disk, R, varies from $4m$ to $12m$. We first construct a minimal spanning tree among the working nodes to estimate the communication energy consumption. We assume that the energy consumed by communication for a working sensor is proportional to n's power of the distance to its farthest neighbor in the tree ($n = 2, 4$). As mentioned earlier, we suppose that we can find a sensor at any desirable position. Since this assumption may not hold in practical applications, we select sensors that are closest to the pre-defined positions in our ideal patterns. We use the following metrics to compare the performance of all patterns:

1) Sensing energy consumption per area (SECPA) in one round.
2) Communication energy consumption per area (CECPA) in one round.
3) Total energy consumption per area (TECPA) in one round.

We use the energy cost model as in [16], to estimate the TECPA of the entire network. In this model, the total energy consumption of each working node is calculated by the following formula:

$$E = kS(R)^x + (1-k)T(R)^y + C,$$

where $S(R)$ and $T(R)$ denote the sensing range and communication range of a sensor, respectively; x and y are constants between 2 and 4; k is a constant, such that $0 \le k \le 1$. The energy consumed by the idling radio and processor of each sensor is a constant, C. Similar to [16], we select $x = y = 4$ and $C = 2000$ for our simulation.

5.2 Simulation results

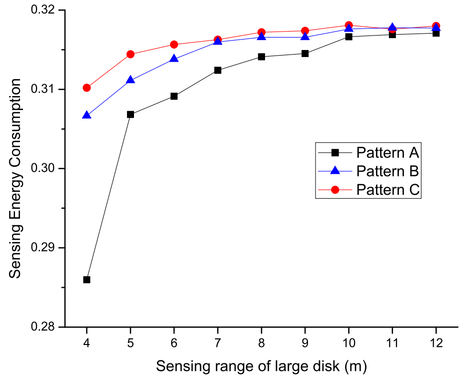


Fig. 6: Sensing energy consumption.

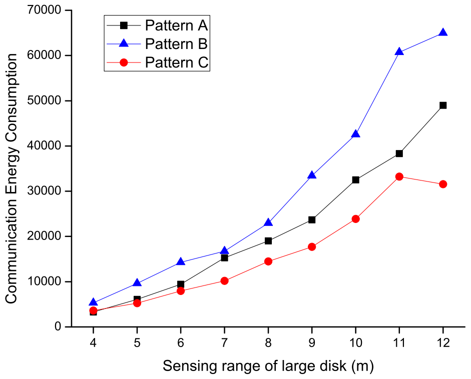


Fig. 7: Communication energy consumption ($n = 4$).

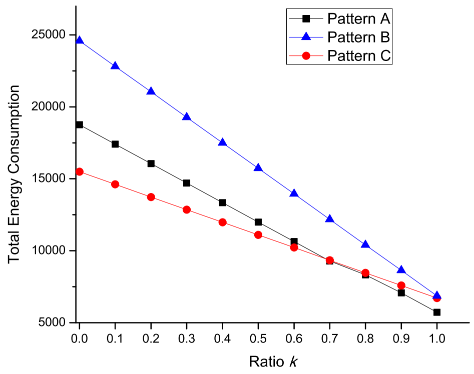


Fig. 8: Total energy consumption.

As shown in Fig. 6, the simulation results for the sensing energy consumption are correlated with the theoretical analysis in section 4. Patterns A and B are better than pattern C in terms of the sensing energy consumption. Pattern C has the highest sensing energy consumption since its active sensors in one round create the same pattern with the uniform sensing pattern based on the regular triangle tile.

Fig. 7 shows the communication energy consumption of all patterns when the path loss exponent is $n = 4$. Pattern C is the best pattern with respect to the communication energy consumption, since the average distance between neighbor nodes in pattern C is smaller than other patterns. Thus, the communication energy consumption is minimized.

Fig. 8 shows the total energy consumption of three patterns. It is obvious that our proposed pattern C outperforms the other patterns in terms of the total energy consumption in most cases. When the ratio $k < 0.7$, the total energy consumption of pattern C is always lower than other patterns. In high rate data transmission WSNs, the ratio k between the sensing and transmission is considerably low. Therefore, pattern C is the most preferred pattern in a WSN that has frequent traffic. On the other hand, patterns A and B are suitable for WSNs that have low traffic. Moreover, in a WSN which the data transmission rate is dynamic, we can use both pattern A and C to have the most benefit of those two patterns. When the transmission rate is low, we can use pattern A. Pattern C can be switched on when the transmission rate increases.

6. Conclusion

In this paper, we constructed an energy-efficient coverage pattern under the condition of a high ratio of sensing coverage. Our mathematical and simulation results show that our new proposed pattern significantly improves energy consumption of patterns A and B that were previously known as the best by up to 17.4%. Our future research will consider scheduling algorithms employed on the proposed pattern and improve the sensing energy consumption as well as minimizing the number of deployed sensors.

Acknowledgment

"This research was supported in part by MKE and MEST, Korean government, under ITRC NIPA-2011-(C1090-1121-0008), WCU(No. R31-2008-000-10062-0)and PRCP(2011-0018397) through NRF, respectively."

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," IEEE Communications Magazine, vol. 40, pp. 102-114, 2002.

[2] X. Bai, D. Xuan, Z. Yun, T. H. Lai, and W. Jia, "Complete optimal deployment patterns for full-coverage and k-connectivity ($k \leq 6$) wireless sensor networks," In Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 401-410, 2008.

[3] M. Cardei, J. Wu, and M. Lu, "Improving network lifetime using sensors with adjustable sensing ranges," International Journal of Sensor Networks, vol. 1, pp. 41-49, 2006.

[4] M. Cardei, J. Wu, "Energy-efficient Coverage Problems in Wireless Ad-hoc Sensor Networks," International Journal of Computer Communications, pp. 413-420, 2006.

[5] G. J. Fan, S. Jin, S. Cheng, "An Approach to Finding Weak Regions for Wireless Sensor Networks with Arbitrary Sensing Areas," In Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 445-446, 2008.

[6] C. F. Huang, Y. C. Tseng, L.C. Lo, "Coverage and Connectivity in Three-Dimensional Underwater Sensor Networks," Wireless Communication and Mobile Computing, vol. 8, no. 8, pp. 995-1009, 2008.

[7] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B.Srivastava, "Coverage problems in wireless ad-hoc sensor networks," In Proceedings of the 20th IEEE INFOCOM, pp. 1380-1387, 2001.

[8] L. Porta, T. IIlangasekare, P. Loden, Q. Han, and A. Jayasumana, "Continuous Plume Monitoring Using Wireless Sensors: Proof of Concept in Intermediate Scale Tank," Journal of Environment Engineering, vol. 135, no. 9, pp. 831-838, 2009.

[9] V. Raghunathan, C. Schurgers, S. Park, and M. B. Srivastava, "Energy-Aware Wireless Microsensor Networks," IEEE Signal Processing Magazine, vol. 19, pp. 40-50, 2002.

[10] W. Song,R. Huang, B. Shirazi, and R. LaHusen, "TreeMAC: localized TDMA MAC protocol for real-time high-date-rate sensor networks," In Proceeding of the IEEE Internation Conference on Pervasive Computing and Communications, pp. 750-765, 2009.

[11] W. Wang, V. Srinivasan, B.Wang, and K. C. Chua, "Coverage for target localization in wireless sensor networks," IEEE Transactions on Wireless Communications vol. 7, no. 3, pp. 667-676, 2008.

[12] J. Wu, and S.Yang, "Energy-efficient node scheduling models in sensor networks with adjustable ranges," International Journal of Foundations of Computer Science, vol. 16, pp. 3-17, 2005.

[13] G. Yang, and D. Qiao, "Multi-Round Sensor Deployment for Guaranteed Barrier Coverage," In Proceedings of the 30th IEEE INFOCOM, pp. 2462-2470, 2010.

[14] V. Zalyubovskiy, A. Erzin, S. Astrakov, and H. Choo, "Energy-efficient Area Coverage by Sensors with Adjustable Ranges," Sensors Journal, vol. 9, pp. 2446-2460, 2009.

[15] H. Zhang, and J. C. Hou, "Maintaining Sensing Coverage and Connectivity in Large Sensor Network," International Journal of Ad Hoc and Sensor Wireless Networks, pp. 89-124, 2005.

[16] Z. Zhou, S. R. Das, and H. Gupta, "Variable radii connected sensor cover in sensor networks," ACM Transaction on Sensor Networks, vol. 5, no. 1, 2009.

Random Point Process Modeling of Location and Time Dependent Interference in Uplink of Mobile Communication Systems

K.S.Subramanian Iyer[1], **Vijayalakshmi Chetlapalli**[2]
[1]School of Management, International Institute of Information Technology, Pune, India
[2]School of Information Technology, International Institute of Information Technology, Pune, India

Abstract— *This paper proposes a random point process model for the total interference at a base station in a mobile communication system. The expected cumulative interference from users generating voice and/or data traffic, located at different distances from the base station is analytically evaluated . Analytical solution for the total interference is presented for the cases when users are static and mobile. The model is generic and will aid in interference studies of several types of wireless communication systems. It is proposed to verify the analytical solution by system level simulations.*

Keywords: Interference Modeling, Mobile communications, Random Point Processes

1. Introduction

The uplink interference in mobile cellular systems is characterized by several random factors viz., varying user load, no. of users, types of traffic, distance of the the user from the base station (BS) and time dependent arrival of users. Whatever be the method of multiple access (TDMA/FDMA/OFDMA/CDMA), an accurate estimation of uplink interference is critical for optimal use of radio resources and higher spectral efficiency. The received signal at the BS at any given time is a composite of the desired signal (from the mobile that the BS is communicating with) plus the interference signals received from all the other active mobiles within the coverage area of the BS. Interference is also present in the downlink. However, the number of interferers in the downlink is much less as compared to the number in the uplink. So, we focus our analysis on evaluating interference in the uplink.

Random point processes form a special class of stochastic process [1] and are currently used in modeling a variety of problems in wireless communications. Interference in mobile communication systems is closely dependent on the continuously varying distance separation of the users from the BS and the random arrivals/departures of users from the system. The aim of this paper is to present a probabilistic model for interference at the BS based on random point processes. The model reflects network variability in time and space. It has been reported by many authors that shot noise is a good model for representing interference in wireless networks. In the classic shot noise model, shots arrive according to Poisson process. The shots are independent and identically distributed random variables. The overall effective is additive. The exact formulation of this requires point process. In the case of a mobile communication system, time and distance variation of the users can be modeled as point processes to evaluate cumulative interference.

This paper is organized as follows. In Section 2, we present the random point process model for cumulative location and time dependent interference. Section 3 presents the solution for the probability frequency function of the interference using partial differentiation and Fourier transforms. The solution is presented for the cases of static and mobile users. Some numerial results and plots of the expected value of the cumulative interference are presented in Section 4. In Section 5, we elucidate the utility of the interference model and discuss the scope for future work.

2. The Interference Model

Consider the case of a single BS serving a cell, located at a position x (see figure 1). Users are assumed to be located randomly at spatial locations distributed over the area of the cell. The users can initiate voice or data calls. We propose to use point process for modeling the interference as experienced by the BS . The model is based on following postulates:

1) Call arrivals from users in the system are Poisson
2) Time spent by each user within the system (service time) follows negative exponential distribution
3) Position of each user w.r.t BS can be static or random (i.e., users are either stationary or mobile)
4) Users may generate power as per the traffic type

We associate with each interfering user a random interference power $a^j(x_i, t_i)$ where x_i is the distance of the user i from the BS at time t_i and j represents the type of traffic (voice/data) generated by the user. $a^j(x_i, t_i)$ constitute a set of statistically independent random variables.

Let $h(x - x_i, t - t_i)$ be the unit response due to the user i at distance the base station at instant t where $t > t_i$. h is assumed to be deterministic, representing the pathloss. A typical realization of total interference at the BS at time t due to user type j may be written as

$$Y^j(x,t) = \sum_i a^j(x_i, t_i) h(x - x_i, t - t_i) \qquad (1)$$

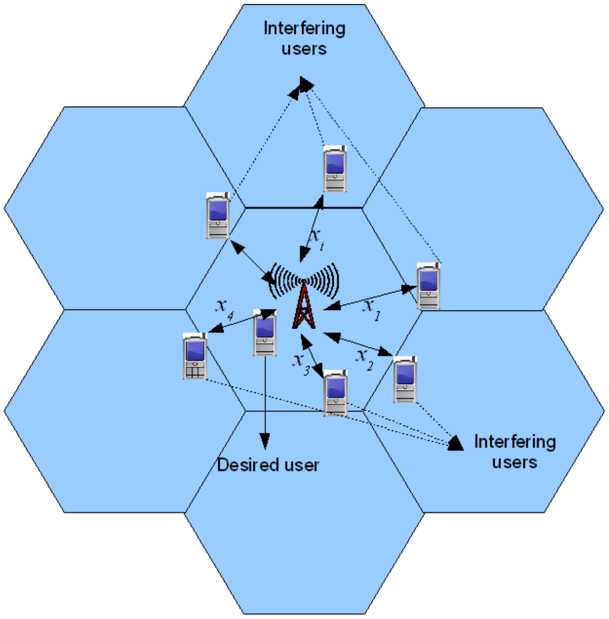


Figure 1: Interference Scenario in the uplink

$a^j(x_i, t_i)$ being a random variable, Y^j is also a random variable, since h is deterministic. It may be noted that Y^j is non-Markovian and we can express the moments of Y^j by using probability frequency function $\pi(y^j, t)$.

For a time-independent system, we can express the moments of Y^j as

$$E[Y^j(x) = \sum_i E[a(x_i)]h(x, x_i) \tag{2}$$

For a time-dependent and location-independent system (i.e, static users), we can express the moments of Y^j as

$$E[Y^j(t) = \sum_i E[a(t_i)]h(t, t_i) \tag{3}$$

Voice and data call arrivals can both be modeled as Poisson processes with exponential service times [2], [3], [4]. The inter-arrival times and service times for each of these traffic types will be different,though. For the rest of the discussion, we drop the index j, as the solution for Y will be identical for both data and voice traffic, except for the values of the Poisson parameter λ and the mean service rate μ.

3. The Solution

The cumulative interference Y is completely characterized by its probability frequency function $\pi(y, t)$. All the moments of Y can be derived from $\pi(y, t)$. In this section, we investigate the solution for $\pi(y, t)$. The solution is presented for two cases—the case of static users and the case of mobile users.

$\pi(y, t)dy$ denotes the probability that Y takes a value lying between Y and $Y + dy$ at time t, Y being zero at $t = 0$. $p(a_i)da_i$ is the probability of interference power lying between a_i and $a_i da_i$. Let the poisson arrivals of calls be characterized by the parameter λ.

Assuming the system to be time invariant, we can analyze the events that occur in $(0, \Delta t)$. There are two possible events: (I)

1) A user arrives in $(0, \Delta t)$ with probability $\lambda \Delta t$, in which case a contribution $a(x_i)h(x_i)$ arises and the contribution to $\pi(y, t)$is given by $\lambda \Delta t \int \pi(y - ah, t)p(a)e^{(-\mu t)}da$ where μ is the servicing rate.
2) No user arrives in $(0, \Delta t)$ with probability $1 - \lambda \Delta t$, in which case the origin is shifted to Δt with the result $\pi(y, t - \Delta t)dy$ will give the probability that y lies between y and $y+dy$ at time t. The corresponding contribution to $\pi(y, t)$ is $(1 - \lambda \Delta t)\pi(y, t - \Delta t)$.

Combining (1) and (2) we get

$$\begin{aligned}\pi(y, t) =& (1 - \lambda \Delta t)\pi(y, t - \Delta t) + \\ & \lambda \Delta t \int \pi(y - ah, t)p(a)e^{(-\mu t)}da\end{aligned} \tag{4}$$

This leads to the partial differential equation

$$\frac{\partial \pi}{\partial t} = -\lambda \pi + \lambda \int \pi(y - ah, t)p(a)e^{(-\mu t)}da \tag{5}$$

Defining $\bar{\pi}(s, t) = \frac{1}{2\pi} \int \pi(y, t)e^{isy}dy$, we get

$$\frac{\partial \bar{\pi}(s, t)}{\partial t} = -\lambda \bar{\pi}(s, t) + \lambda \int e^{iash} \bar{\pi} e^{-\mu t} p(a)da \tag{6}$$

Defining $\Phi(u) = \int p(a)e^{iua}da$, the characteristic function of $a(x, t)$, and substituting in above,

$$.\frac{\partial \bar{\pi}(s, t)}{\partial t} = -\lambda \bar{\pi}(s, t) + \lambda \bar{\pi} e^{-\mu t} \Phi(sh) \tag{7}$$

The solution of $\bar{\pi}(s, t)$ can be written as

$$\bar{\pi}(s, t) = e^{-\lambda t + \lambda \int_0^t e^{\mu \tau} \Phi(sh) d\tau} \tag{8}$$

We note that $\bar{\pi}(s, t)$ is the characteristic function of the cumulative interference power Y. It yields all the moments of Y. The first moment N is given by

$$\frac{\partial N}{\partial t} = -\lambda N + \lambda N e^{-\mu t} \Phi(0) + h\Phi^{'}(0)\lambda e^{-\mu t} \tag{9}$$

Solving for $N(t)$ we get

$$N(t) = h\Phi^{'}(0)\lambda e^{-\Psi(t)} \int_0^t e^{\Psi(\tau) - \mu \tau} d\tau \tag{10}$$

where
$\Psi(t) = \lambda t - \frac{\lambda \Phi(0)}{\mu}(1 - e^{-\mu t})$
$\Phi(0) = ka_m$
$\Phi^{'}(0) = k\frac{a_m}{2}^2$
a_m= maximum possible interference power generated by a single user
k= probability with which a user generates a_m

$N(t)$ represents the expected value of the cumulative interference power at the BS. It is proposed to simulate the system based on (1) and (10).

What is crucial to the solution of $\pi(y,t)$ is the determination of $p(a)$. $p(a)$ is the probability distribution function of the interference contribution from individual users. The determination of $p(a)$ has to be dealt with differently for the cases of static and mobile users. These two cases are discussed in detail in the following subsections.

3.1 Evaluation of $p(a)$ for Static Users

When all users are stationary, a is only time-dependent, and is characterized by the distribution of the traffic type. In this case, $p(a)$ will be the distribution function of the traffic (poisson for voice, etc). For the traffic types that cannot be represented using known distributions, $p(a)$, λ and μ have to be evaluated from simulations or traffic experiments and used for arriving at $\pi(y,t)$.

3.2 Evaluation of $p(a)$ for Mobile Users

For this section of the discussion, we will denote the distance separation of the user from the BS as x instead of $x - x_i$, to simplify the notation. When the users have mobility, a depends on both distance and time. In this case, distance of the user from the BS varies with time. We define $p(x \mid x_0, t \mid t_0)$ as the probability that the user is at a distance x from BS at time t , given that the position of user at $t = t_0$ is x_0 . Essentially, $p(a)$ is now replaced by $p(x \mid x_0, t \mid t_0)$. Since x varies with time t, $p(x \mid x_0, t \mid t_0)$ has to be evaluated using Kolmogorov forward equations. The maximum allowed separation distance between the BS and the user beyond which handover takes place is denoted by x_c. We make use of the Markovian nature of $x(t)$ to represent the following:

$$p(x \mid x_0), t + \Delta t) = \int p(x^{'} \mid x_0, t) p(x \mid x^{'}, \Delta t) dx^{'} \quad (11)$$

where

$$\begin{aligned} p(x \mid x^{'}, t, \Delta t) =& R(x \mid x^{'})\Delta t + \\ & \delta(x - x^{'}) \left[1 - \Delta \int R(x \mid x^{'}) dx \right] \end{aligned} \quad (12)$$

$R(x \mid x^{'})\Delta t$ represents the probability that the user moves from $x^{'}$ to x in Δt and is proportional to Δt . This depends on the velocity of the user, or more generally, the mobility model used. The second term in the equation represents the probability that the user continues to remain at the same distance x from the BS.

Substituting in the Kolmogorov forward equation, we get

$$\begin{aligned} \frac{\partial p(x \mid x_0, t)}{\partial t} =& - p(x \mid x_0, t) \int R(x^{'} \mid x) dx^{'} + \\ & \int p(x^{'} \mid x_0, t) R(x \mid x^{'}) dx^{'} \end{aligned} \quad (13)$$

Assuming $R(x^{'} \mid x)dx^{'} = R(x^{'})dx^{'}$ and $\int_{x_0}^{x_c} R(x)dx = b$, the above equation reduces to

$$\frac{\partial p}{\partial t} = -bp + R(x) \quad (14)$$

Using initial condition $p(x \mid x_0, 0) = \delta(x - x_0)$, we can solve the above equation thus

$$p(x \mid x_0, t) = \frac{R(x)}{b}(1 - e^{-bt}) + \delta(x - x_0)e^{-bt} \quad (15)$$

4. Numerical Results

To gain some insight into the validity of the interference model, the values of$N(t)$ have been plotted for various values of λ and μ, for the case of static users. The traffic is assumed to be homogeneous for each of the plots, i.e., no combination of dissimilar traffic types has been considered. The probability k is assumed to be 0.1. Typically, this value depends on the percentage of users at the cell edge, as these will be transmitting at maximum power. As we have considered only static case, the response factor h only models pathloss, without any fading effects. The study for the case of mobile users is under progress.

4.1 Variation of $N(t)$ with Arrival Rate

Figure 2 shows the plot of $N(t)$ for two values of arrival rates, 1/7 and 1/20. In mobile communications, the typical average interarrival time for voice calls is 7 seconds [5]. The arrival rate of data calls is modeled to be $1/3^{rd}$ of that of the voice calls. The figure shows variation of N(t) for interarrival times of 7 and 20 seconds. It can be seen that the expected interference increases with arrival rate, for a given service rate.

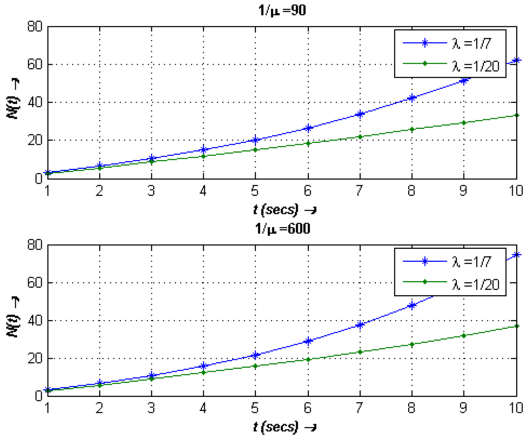


Figure 2: Variation of $N(t)$with λ

4.2 Variation of $N(t)$ with Service Rate

The variation of $N(t)$ with service rate μ is depicted in Figure 3. The typical call holding times ($1/\mu$) for voice calls is taken to be 90s and that of data calls to be 600s. The expected value of interference is higher when the call holding time is longer, for a given arrival rate. This is because the average number of interfering users in the system will be more for longer call holding time.

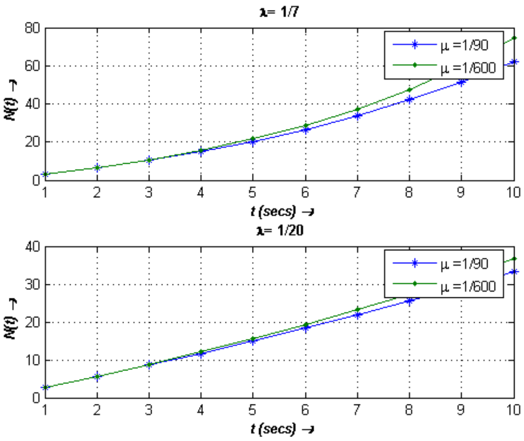


Figure 3: Variation of $N(t)$with μ

5. Conclusion

In this work, an attempt has been made to represent all possible aspects in modeling the cumulative interference power in the uplink of a mobile communication network $viz.$, varying user load, number of users, types of traffic, distance of the users from the BS and time dependent arrival of users. One of the significant uses of this model is that it accounts for heterogeneous traffic. The cumulative interference from different traffic types can be estimated by including respective values of $\lambda$and μ for each of the traffic types. Though the model presented here is for the case of a single cell, it can be extended for the study of cellular communication systems comprising of multiple cells, with appropriate changes. Further, it is proposed to validate the analytical model by simulations. Simulations are under progress and the results will be produced for publication in the near future.

6. Acknowledgement

The authors would like to thank International Institute of Information Technology, Pune, India for their approval and support in carrying out this research.

References

[1] Srinivasan S.K and Iyer, K.S.S., *Random Processes Associated With Random Points on a Line*, Zastosowania Matematicae, Applications Mathematicae VIII (1966), p 229.

[2] *E*TSI, TR101 112v3.2.0, Selection Procedures for the choice of radio transmission technologies of the UMTS, UMTS 30.03 v3.2.0

[3] Seyed Mohammadreza Hashemiannejad, Aziz Mahmoodi and Jahangir Dadkhah Chimeh, *Delay Calculation of Internet Traffic in Mobile Systems* International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009.

[4] David Soldani, *QoS Management in UMTS Terrestrial Radio Access FDD NetworksŤ Helsinki University of Technology* Ph.D. Thesis, October 2005.

[5] A. Pattavina, A. Parini *Modelling voice call interarrival and holding time distributions in mobile networks.* Proc. of 19th International Teletraffic Congress 19, 2005.

Authentication/Association Flooding Dos Attacks And Their Impact On Wireless Mesh Networks(ICWN' 11)

Rupinder Cheema[1], **Ms. Divya Bansal**[2], **and Prof. Sanjeev Sofat**[2]
[1]Computer Science, PEC University of Technology, Chandigarh,Chandigarh,India.
[2]Computer Science, PEC University of Technology, Chandigarh,Chandigarh,India.

Abstract - *Wireless mesh network is the key technology for present generation in wireless networking for providing fast and hassle free services to users. The conformity of Wireless Mesh Networks to offer Internet connectivity is becoming a popular choice for Wireless Internet Service Providers, as it allows a flexible, easy and inexpensive network deployment. However, these networks are prone to critical attacks because of the security fissures in the draft standard. Moreover security in Wireless Mesh Networks is still in its infancy as very little attention has been devoted so far to this topic by the research community. The security mechanisms proposed so far can work well for securing only data frames.Management frames are unencrypted and have been sent transparently, thus facilitating the launch of dos attacks. So for pervasive network connectivity, it is mandatory to thwart these attacks for the uninterrupted and stable network availabilty. In this paper we have analyzed the impact of Authentication and Association flooding dos attacks over the network performance.*

Keywords: Authentication flooding, Association flooding, DOS attacks, Vulnerabilities, Wireless Mesh Networks.

1 Introduction

The disposition of Wireless networks have been proliferated at a tremendous rate, because of the functionalities offered by these networks over their wired counterparts. WiFi has made it possible to relax the wired constraints by offering wireless access[1]. Thus it has been made feasible for anybody to connect from anywhere to the Internet or to the local network without any physical connection. This world of wireless saw the birth of new technologies like wireless ad-hoc networks and wireless mesh networks which enable great flexibility of deployment[2]. The shortcomings of Wireless networks have been overcome by the draft standard 802.11s referred to as Wireless Mesh Networks. [3] If the network access has to be provided to the distant users , who are not physically located at the same place, then in that case Wi-Fi has not been able to serve the purpose. In that case, Wireless Mesh Networks have to be deployed[ ] . Due to its contributions to eliminate the complexity of installation, configuration and maintenance of wireless network, to ensure a better quality of services and to provide compatibility with external and heterogeneous networks, Wireless Mesh networks have become a universal and topical issue and captured the interest of university research and industry [1] [2]. Wireless Mesh Networks (WMNs) represent a good solution to Wireless providing wireless Internet connectivity in a sizable geographic area. This new and promising paradigm allows for network deployment at a much lower cost than with classic WiFi networks [16]. The architecture of a WMN involves different components which ensure the execution of the network operations. A Wireless mesh network is a communication network made up of radio nodes organized in a mesh topology. Wirereles mesh networks often consists of Mesh clients, Mesh routers, gateways and Portal. Mesh stations can collocate with 802.11 access points and provide access to the mesh network to 802.11 stations [3]. Mesh portal is stationary and located inside the building and at the backend connected via wires to the gateway for providing the network access [14] The Mesh points can be deployed at different places depending upon the requirements.. Fig. 1 shows the architecture of the wireless mesh networks.The mesh nodes can be clients (or stations) and relay routers at the same time, and so, they can be integrated in the traffic routing. This makes it possible to guarantee the multiplicity of the paths to reach any destination in the mesh network. Moreover , both of these actions can be performed simultaneously[16] . As it has been assumed that all, the STAs are communicating with their MAPs using IEEE 802.11 b standard while MPs/MAPs are communicating with each other using IEEE 802.11a standard [13]. So, the communication between MAPs and STAs has not been interfering with forwarding actions between MPs and MAPs. Because of its architecture, IEEE 802.11s standard draft has been providing end users with better experiences, more achievable bandwidth , fewer cost , and more fairness than 802.11 standards do [16].

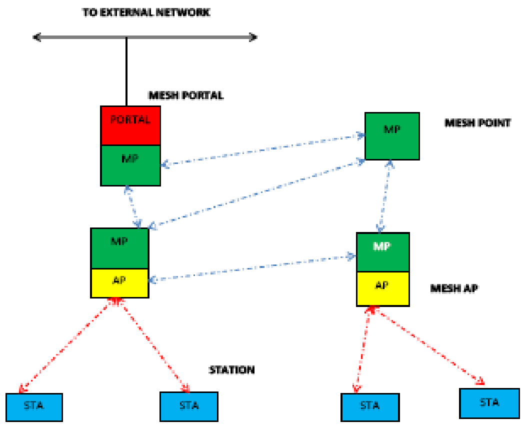


Fig. 1 Architecture of wireless mesh networks

2. Security Mechanisms

The problem of security is a great concern in all types of wireless networks. While networks continue to be developed, many efforts are concurrently ongoing to make sure that network access has been granted to the authorized users only [6]. Moreover, the network complexity usually grows with the increase in the number of applications, the nodes mobility and the degree of medium opening towards the outside. Wireless mesh networks have introduced several new risks along with the facilitation of communication and the various advantages they offer [3]. As Wireless mesh networks have become inevitable for gaining network access, so there is an indispensable need for the ubiquitous network availability. And for ensuring that, there has been the imperative need of devising security mechanisms for these networks. So that, the threats and vulnerabities to which these networks are prone to can be detected. The Internet is a set of interconnected heterogeneous computer networks [17]. In order for these networks to communicate deterministically, communication protocols are standardized in standards bodies like the [IETF].These protocols are a set of published rules that specify the structure of the information on the wire and their associated semantics. Different protocols have been proposed by the IEEE for ensuring security. The choice among these depends upon the specific type of the application for which it has to be deployed [10]. Firstly comes the WEP, the most simpler one based on the RC4 algorithm and simply relies on concatenating the fixed part with the 24 bit variable IV for generating the key. It has got several vulnerabilities such as limited size of the initialization vector , which leads to key reuse. As with this length of IV only 2^{24} unique keys can be generated. And it can be compromised within few minutes, and is prone to many attacks such as the FMS attack. So the use of WEP has not been considered as the good choice for ensuring security. [11] Then comes the WPA standard that had been proposed by the RSN (robust security network). This standard has got hardware compatibility with the already existing version but it has been upgraded in software. Moreover the key generation has been complicated by the introduction of the key mixing function and extending it to four levels [17]. So this has been complicated the key generation process, thus making it more secure. But still it have got some vulnerabilities and can be compromised. Then Task group 802.11i had proposed the most secure protocol 802.11i. In this standard, the integrity has been provided by the MICHAEL. The use of nonces have been introduced for generation of keys. As these nonces can be used only, so the chances for the compromise of the keys have been reduced. Moreover, this standard is based on the most secure AES algorithm, thus has added up to the security [13]. In this case, the process of authentication is not limited to clients and APs but there has been the introduction of one more functionary Radius Server, which is the main authority for granting access [10]. This mechanism has been relying on the use of secure protocols such as EAP, PEAP. So it cannot be easily cracked. Hence, Wireless mesh networks as more susceptible to attacks have been relying upon this protocol for security [16].

3. Security Breaches

Although all the networks are vulnerable to security breaches, but in wired networks, physical barriers reduce risk by limiting media access. Ethernet switches can be locked away in closets and offices, and unused jacks can be disabled [1]. But in wireless networks, the medium is the air. Walls, doors, and floors reduce signal strength, but do not stop attacks launched from stairwells, lobbies, parking lots, or nearby buildings [2]. These new vectors can be used to exploit vulnerabilities that are inherent to 802.11 and 802.1X [10]. The radio communication in case of wireless networks take place by the exchange of different frames [12]. These frames are of different categories such as data frames, control frames and management frames. Each frame incorporates the information like the sequence number, time stamp, duration, bssid, source address, destination address [2]. Wireless mesh networks have been relied upon the use of standard protocols for ensuring security. The security mechanisms existing so far are viable for data frames only[13]. Management and control frames are unencrypted and sent in clear. So this communication has heard by the attacker too. After that, the attacker masquerades the frame and launches the attack over the network [3]. A key problem in detecting denial of service attacks is that the source address of the packets has been spoofed. This ensures that the compromised machines will remain undetected and thereby can be used for other attacks [15]. So there has been the indispensable need to devise the security mechanism for defending against dos attacks that can secure these frames and hence the network state has been preserved [12].

4. MAC Layer Vulnerabilities

As the MAC layer is very much susceptible to the dos attacks because of the breaches like the unencrypted management frames that lure hackers. The attackers can easily gather the credentials of the legitimate clients by simply using certain freely available hacking tools [6,7]. These tools can be easily operated. So anyone with little knowledge can use these for launching the attacks. Moreover, the victims are not even aware of this, as the attackers have been obfuscated their presence over the network[4].

These vulnerabilities are categorized into two groups :-

4.1 Identity Vulnerabilities

These security breaches have been aroused because of spoofing. This 48 bit unique address imprinted in hardware over the adapter has been providing uniqueness to the different stations and the Access points [4]. In case of wireless communication the trust lies in this MAC address only [2]. It can be changed in software thus leaving room for the launch of attacks [6,7].

4.2 Media Access Vulnerabilities

As in case of radio communication too the access to the shared medium is required for gaining access. This access has been intervened by the attacker, thus long delays have been introduced for the legitimate users before gaining access [2]. Thus the launch of DOS attacks have been facilitated [13].

5. Measures for Thwarting Attacks

There has been an inescapable need for the security as the MAC layer has got a number of security breaches which can be exploited easily. Moreover, there has not been much work explored in securing the wireless networks[10]. In case of IEEE 802.11 based networks, although some protocols have been deployed for encryption and authentication. Only data frames have been benefitted by the use of such protocols [10]. There has been no security mechanism devised so far for protecting the managerment and control frames. So the launch of the attacks over these cannot be vetoed [13]. Although there are some mechanisms for detecting the presence of attacks but they have not considered the situation in which network is congested because of the load placed due to the legitimate clients[17]. Consequently, the large number of false positives has been generated. Moreover these security mechanisms have not been able to serve WMNs because of the distributive, mobile and complex nature of these networks [5]. However, the architecture and configuration of this type of network do not ensure protection against unauthorized use of the network. This is because the basic used security measures do not include the notion of mobility, which characterizes these networks[17]. So there has been an unavoidable need for the mechanisms for safeguarding security in Wireless mesh networks [13].

6. DOS Attacks

Dos attacks aim at overwhelming the user with the reservoir of data so that, the network’s capability to entertain legitimate client’s requests has been vanished [12]. As a consequence the network has been mobbed and the performance starts deteriorating. Finally the connection has been disrupted and thus leading to the nullification of the network’s avaialabilit[1,2]. These attacks have been focused at consuming the network’s resources such as the bandwidth. They can be launched at all the layers of the OSI reference model such as jamming attack in case of the physical layer.[1] In this paper we are mainly focusing on the MAC layer attacks at the data link layer. The mode of operaion at this layer in case of WMNs is identical to IEEE802.11 networks.

7. Related Work

Wireless networks are vulnerable to DOS attacks and the results can be anything from degradation of the Wireless network to a complete loss of availability of the wireless network within the organization [12]. The contribution of various authors in this area is illustrated here. The following literature reviews investigated these attacks, their cause, implementation and detection of these attacks for providing security.

In [2] the authors focused that DOS attacks were possible by circumventing the normal operation of firmware in commodity 802.11 devices. Moreover two important classes of DOS attacks were implemented and the range of their practical effectiveness was investigated. In [4] it was focused that various tools may be implemented in firmware for injecting raw 802.11 frames into the channel. They examined the MAC layer and identified a number of vulnerabilities that could be exploited. The reasons behind identity vulnerabilities were the non-verifiability of the source and destination address contained in the MAC management frames which could have been easily spoofed by the attacker [4]. Then attacker could have intervened the communication and acted as client. So that the AP responded to the requests sent by the attacker as if they were sent by the client itself, thus DOS attacks could have been facilitated [18]. In [2] it was focused that the main reason behind the launch of these DOS attacks was the one way authentication that had been established in 802.11 networks. There was no provision for the client to authenticate AP, only AP was having the flexibility to validate the identity of the client. Consequently the launch of Rouge AP attack was facilitated. The part of the message exchanged between the client and the AP was not authenticated by any of the keyed mechanisms. So that had been spoofed and redirected by the attacker. As a consequence whole communication would have been stopped that resulted in the unavailability of the network resources to the legitimate users [13]. .Mina Malekzadeh, Abdul azim,Jatil desa,Shamala” An experimental evaluation of dos attacks and its impact over throughput in wireless networks “ demonstrated the impact of these attacks over the network performance [15].

8. Communication Services

The communication between the Mesh AP and the Mesh client takes place by the exchange of several frames. As initially the client has been neither associated nor authenticated so is in state 1[1]. Then the client probes for the available networks by using probing either active or passive. After the selection for the network has been made the client sends the authentication request to the appropriate network. In response, the AP after verifying the credentials if wishes to authenticate that network to itself will send the appropriate reply, thus validity the entity of the station to it[2]. Then after the exchange of these set of frames the station enters state 2, that is authenticated and unassociated. So the station has been sending the association request to the AP for the establishment of the connection for attaining network access. And the AP responds with the association response if has not yet attained the figure of maximum allowable stations which it can allow [6]. It may be feasible that, a station has been authenticated to more than one networks but associated to only one network at an instant.Fig. 2 shows the whole transition process and the states through which the entities would have to pass for gaining network access securely [12].

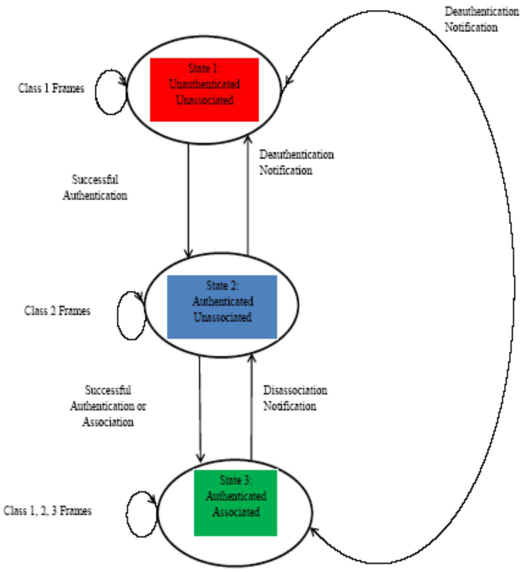

Deauthentication Notification
Class 1 Frames
State 1: Unauthenticated Unassociated
Successful Authentication
Deauthentication Notification
Class 2 Frames
State 2: Authenticated Unassociated
Successful Authentication or Association
Disassociation Notification
Class 1, 2, 3 Frames
State 3: Authenticated Associated

Fig.2 state diagram of management frames

9. Description of the Attacks

The attacker has been launching the attacks by using fake MAC address, thus obfuscating his presence over the network[9]. The AP in turn has been responding to this same fake address and as it actually doesnot exists, so leads to the wastage of networks resources. Thus the network's performance has been deteriorated and vanished [15]. In this paper we are discussing the impact of two such attacks:-

9.1 Authentication Flooding Attack

As in the state diagram it has been discussed that the first step for the station or client in gaining network access via the AP is to validate its identity with that AP [1]. Firstly, the attacker gets access to the network credentials via sniffing and spoofing as he wished to obfuscate his presence over the network, so later on his identity has not been revealed [2]. The attacker sends the authentication request to the AP with spoofed address and the AP responds with the authentication response. But as the request has been sent by the fake address, so the response has not been received by the same. As it actually does not exists and has been created to betray the legitimate APs.[2] So the AP continues sending the response and as this response has not been received. So this flood of response frames has been overwhelming the network causing congestion and thus prohibiting the AP to serve the legitimate clients [6].

9.2 Association Flooding Attack

Association is the second step towards gaining network access. After being authenticated, the client has been sending the association request to the AP for connection establishment [12]. This time too the client has been using the same fake address so the AP can match it with the list of authenticated clients before leveraging connectivity [4]. And as this address have not been present in that list, so the AP responds by sending the deauthentication frames [6] . The whole process has been initiated by the fake addres and as no such client exists so the ACK have been received by the AP. This continual sending of deauth frames by the AP has been devastating the network thus leading to hike in network traffic. Consequently network performance has been disrupted [15].

10. Network Scenario Used

We have launched the attack over the victim client connected to the mesh AP, and the attacker machine running different soft wares thus facilitating monitoring network in promiscuous mode . The network performance seems to be normal before the attack but after the attack it has been degrading considerably and drops down to zero. Network performance can be measured by the parameters viz throughput, bandwidth, jitter, bandwidth. We have considered two parameters throughput and bandwidth for analyzing the performance of our Wireless mesh test bed before and after the attack.

10.1 Authentication Flooding Attack Launch

Fig 4 and Fig. 5 shows the impact of the triggering of theattack over the network performance. We can analyze the impact of the attack over the network performance by varying several metrics say buffer length, size of packets and number of packets . In this paper we have used size of packets to analyze the sway of the attack over the network performance. And it has been noticed that the network performance seems to be normal before the attack and after the launch of the attack the throughput as well as bandwidth starts deteriorating and reaches zero. Thus halting the communication.

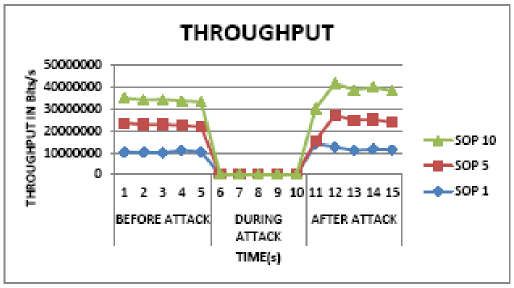


Fig. 3

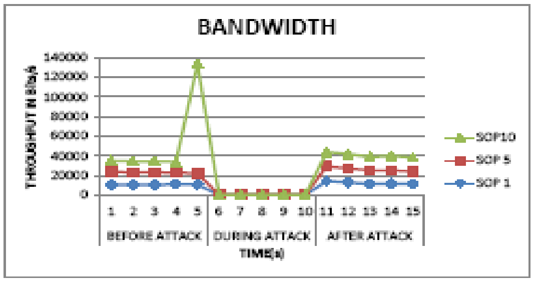


Fig. 4

10.2 Association Flooding Attack Launch

Fig.5 and Fig.6 shows the impact of the association flooding attack over the network performance. In this case too as the attack has been launched till the tenure of the attack the throughput as well as bandwidth starts putrefying and then vanishes. The reason behind this is the congestion in the network caused by these association response floods . Moreover, the resources available with the network are limited so not able to entertain this reservoir of traffic. It requires time for the network to regain its strength and stabilize after the attack.

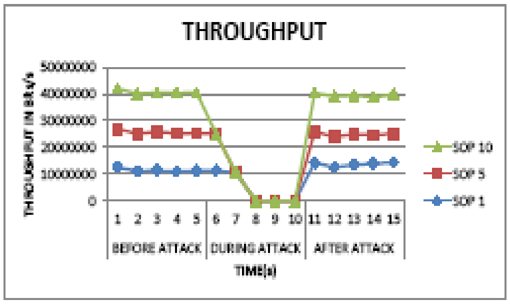


Fig. 5

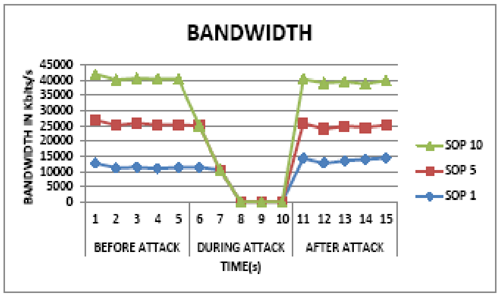


Fig.6

11. Conclusion And Future Work

Wireless mesh network has offered improved performance and ubiquitous connectivity and mobility features. Security is still very important issue in these networks because of the open , mobile and distributive nature of these networks. Wireless mesh networks derive their security from the WPA2 standard. Howerever, these networks are still prone to a wide variety of DOS attacks because of the lack of encryption mechanisms in management frames. Denial of Service Attacks pose a serious threat to networks, and the distributed and ever-changing nature of these attacks makes them difficult to detect and suppress...In this paper we have implemented two common dos attacks against the wireless mesh networks. We have demonstrated the impact of the authentication flooding and association flooding attack on throughput and bandwidth of the mesh network real testbed. From the outcomes of the experiments it has been concluded that dos attacks are serious security problem over the network performance. The results show that these attacks consume the wireless network resources so that the remainder of the throughput almost zero is not enough to continue the normal operation of the network.The mechanisms proposed so far can only detect the occurence of these attacks. So there is the imperative need to devise the solutions for the prevention of these attacks so that the disruption caused to network performance by these attacks can be mitigated.

12. References

[1] Stuart Compton, Charles Hornat. 802.11 Denial Of Service Attacks and Mitigation. SANS Institute InfoSec Reading Room, 14-18, May 17^{th} 2007.

[2] John Bellardo and Stefan Savage. Denial of Service Attacks: Real Vulnerabilities and Practical Solutions In the Proceedings of the 12^{TH} USENIX Security Symposium Washington, D.C., U.S.A. Aug 4 -8, 2003.

[3] N. Bisnik and A. Abouzeid, Delay and Throughput in Random Access Wireless Mesh Networks In the Proceedings Of IEEE International Conference on Communications (ICC 2006), vol.1, pp. 403 -408, Jun 2006.

[4] Joshua Wright Detecting wireless LAN MAC Address spoofing, GCIH, CCNA, pp 1-5, Jan 21, 2003.

. [5] Baber Aslam, M Hasan Islam and Shoab A.Khan. 802.11 Disassociation Dos Attack and its Solutions: A Survey In the proceedings of the First International Conference on Mobile Computing and Wireless Communication, pp 221-222, Amman, Sep 17-20, 2006.

.[6] AirJack. http://sourceforge.net/projects/airjack/

[7] Airsnarf. http://airsnarf.shmoo.com/

[8] D. Dasgupta, F. Gonzalez, K. Yallapu and M. Kaniganti. Multilevel Monitoring and Detection Systems (MMDS). In the proceedings of the 15th Annual Computer Security Incident Handling Conference (FIRST), Ottawa, Canada June 22-27, 2003.

[9] E. D Cardenas. MAC Spoofing {An Introduction. http://www.giac.org/practical/GSEC/Edgar Cardenas GSEC.pdf

[10] J. Hall, M. Barbeau and E. Kranakis. Using Transceiver Prints for Anomaly Based Intrusion Detection. In Proceedings of 3rd IASTED, CIIT 2004, St. Thomas, US Virgin Islands, USA, November 22-24, 2004.

[11] F. Robinson. 802.11i and WPA up Close. Network Computing, 2004.

[12] Peter Egli, Product Manager Wireless & Networking Technologies," Susceptibility of wireless devices to denial of service attacks"/http://www.netmodule.com

[13] A. Gerkis and J. Purcell. A Survey of Wireless Mesh Networking Security Technology and Threats. Sans Infosec Reading Room, Sep 2006.

[14] Wireless Mesh Networks /http://www.wikkipedia

[15]M. Malekzadeh, A azim, J. Desa, S. subramaniam " An experimental evaluation of dos attacks and its impact on throughput of 802.11 netwroks" IJCSNS, Vol.8 , Aug 2008

[16]White Paper "Wireless Mesh Technology: Connecting the new millennium". An IJIS institute Briefing Paper.

[17]Ian F. Akyildiz, Xudong Wang, ''Security in Wireless Mesh Networks''.December 19, 2006

[18] Michael Lowry Lough. A Taxonomy of Computer Attacks with Applications to Wireless. PhD thesis, Virginia Polytechnic Institute, April 2001.

Wi-Fi Technology in the Smart Grid Backhaul

Carlos Henrique Rodrigues de Oliveira and Elaine Franca Fonseca
Telecommunications Research and Development Center (CPqD)[1] / Eletrobras
CEP 13086-902, Campinas, SP / CEP 20071-003, Rio de Janeiro, RJ (Brazil)
carloshe@cpqd.com.br / elaine.fonseca@eletrobras.com

***Abstract*—Smart Grid solutions are being driven by the desire for more efficient energy usage worldwide. Nowadays Smart Grid communications network is a heterogeneous network based on many different standards. This paper describes the suitability of Wi-Fi technology for use in the Smart Grid backhaul.**

***Index Terms*— Smart Grid, Wi-Fi, IEEE P2030.**

I. Introduction

Important Smart Grid design principles are secure, reliable, scalable, manageable, modular, future proof and "open" standards-based interoperable.

The IEEE P2030 is a standard guide for Smart Grid interoperability that addresses the basic Smart Grid definitions, frameworks, challenges and three different architectural perspectives (Power & Energy, Communications and IT).

IEEE P2030 provides the basis for ongoing standards development discussions but it is still a work in progress and currently Smart Grid communications network is a heterogeneous network based on many different standards.

II. Background [1]

The need for Smart Grid solutions is being driven by the emergence of distributed power generation and management/monitoring of consumption, and the desire for more efficient energy usage worldwide. Smart Grid advancements will apply digital technologies to the grid, enabling two-way communications and real-time coordination of information from generating plants, distribution resources and demand-side end points.

Standards are critical to enabling interoperable systems and components. Mature international standards are the foundation of markets for the millions of components that will have a role in the future Smart Grid.

According to [1], Smart Grid Standards must:

- ✓ Provide two-way communication among grid users, e.g. regional market operators, utilities, service providers and consumers;
- ✓ Allow power system operators to monitor their own systems as well as neighboring systems that affect them;
- ✓ Coordinate the integration into the power system of emerging technologies such as renewable resources, demand response resources, electricity storage facilities and electric transportation systems;
- ✓ Ensure the cyber security of the grid.

III. The Smart Grid Framework

The Smart Grid Framework is presented in Figure 1. The Smart Grid is a large system of 17 subsystems.

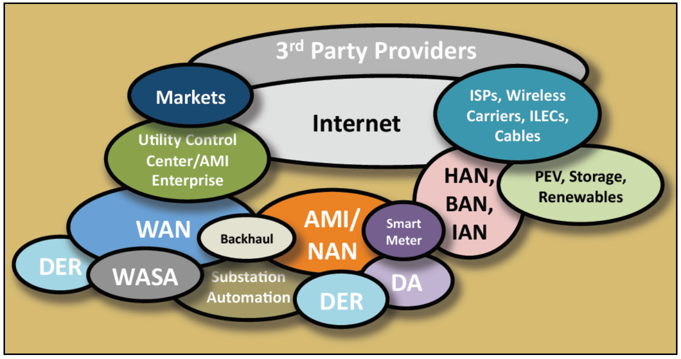


Figure 1. Smart Grid Framework [2]

The Smart Grid will be created by IT, Telecom, embedded platforms, advanced control, cyber security, power electronics and energy engineers and specialists.

IV. Smart Grid Architecture

The backhaul is present in the end-to-end Smart Grid Communications Architecture as showed in Figure 2 and in the IEEE P2030 Smart Grid Communications Reference Architecture (SG-CRA) as showed in Figure 3.

The focus of this work is in the Telecommunication area, the wireless communication between DA/AMI/NAN and WAN, the backhaul.

[1] This work was supported by the Smart Grid Parintins Project of Eletrobras - Brazil.

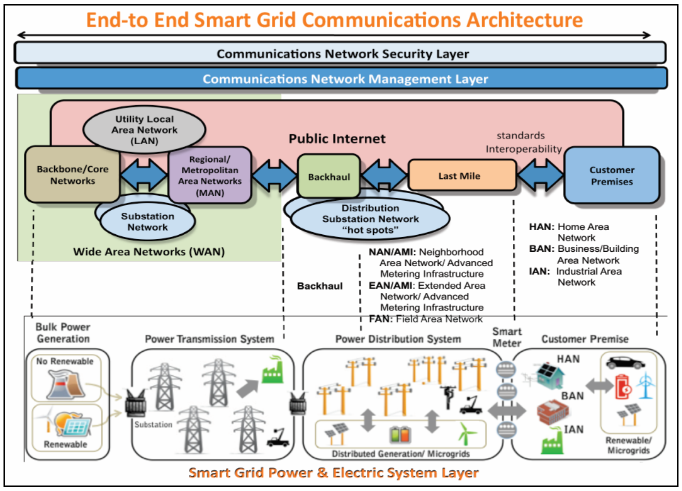

End-to End Smart Grid Communications Architecture
Communications Network Security Layer
Communications Network Management Layer
Utility Local Area Network (LAN)
Public Internet
standards Interoperability
Backbone/Core Networks
Regional/ Metropolitan Area Networks (MAN)
Backhaul
Last Mile
Customer Premises
Substation Network
Distribution Substation Network "hot spots"
Wide Area Networks (WAN)
HAN: Home Area Network
BAN: Business/Building Area Network
IAN: Industrial Area Network
NAN/AMI: Neighborhood Area Network/ Advanced Metering Infrastructure
EAN/AMI: Extended Area Network/ Advanced Metering Infrastructure
FAN: Field Area Network
Backhaul
Bulk Power Generation
Power Transmission System
Power Distribution System
Smart Meter
Customer Premise
No Renewable
Renewable
Substation
HAN
BAN
IAN
Renewable/ Microgrids
Distributed Generation/ Microgrids
Smart Grid Power & Electric System Layer

Figure 2. End-to-end Smart Grid Communications Architecture [3]

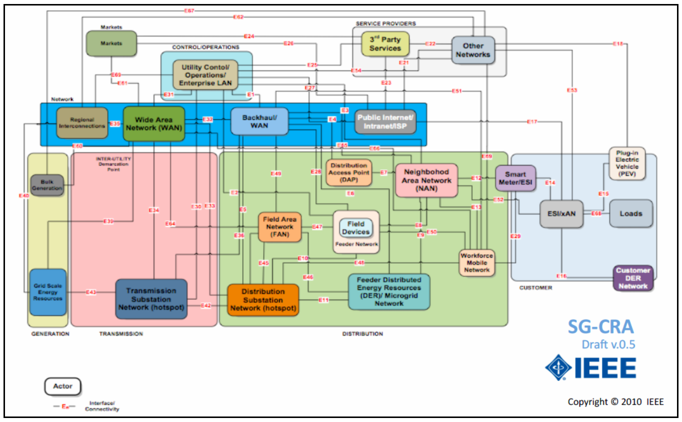
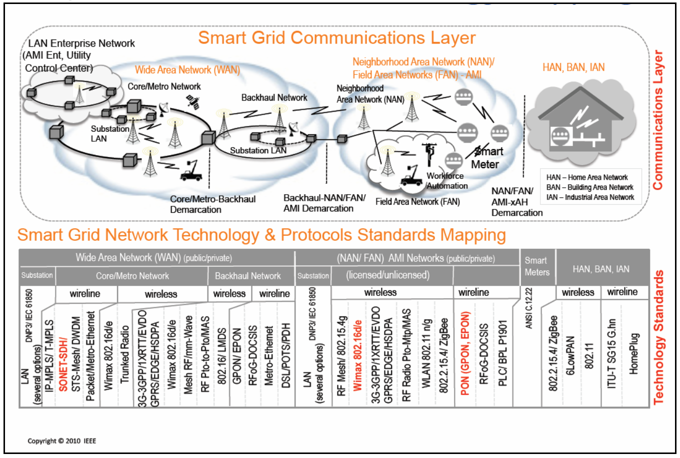

Markets
Markets
SERVICE PROVIDERS
3rd Party Services
Other Networks
CONTROL/OPERATIONS
Utility Contol/ Operations/ Enterprise LAN
Network
Regional Interconnections
Wide Area Network (WAN)
Backhaul/ WAN
Public Internet/ Intranet/ISP
INTER-UTILITY Demarcation Point
Bulk Generation
Distribution Access Point (DAP)
Neighbohod Area Network (NAN)
Smart Meter/ESI
Plug-in Electric Vehicle (PEV)
ESI/xAN
Loads
Field Area Network (FAN)
Field Devices
Feeder Network
Workforce Mobile Network
Grid Scale Energy Resources
Transmission Substation Network (hotspot)
Distribution Substation Network (hotspot)
Feeder Distributed Energy Resources (DER)/ Microgrid Network
Customer DER Network
CUSTOMER
GENERATION
TRANSMISSION
DISTRIBUTION
SG-CRA
Draft v.0.5
IEEE
Actor
Interface/ Connectivity
Copyright © 2010 IEEE

Figure 3. IEEE P2030 Smart Grid Communications Reference Architecture (SG-CRA) [3]

V. Smart Grid Network Technology Mapping

The Smart Grid Network Technology Mapping presented in Figure 4 suggests RF mesh to wireless backhaul network among others.


Smart Grid Communications Layer
LAN Enterprise Network (AMI Ent, Utility Control Center)
Wide Area Network (WAN)
Core/Metro Network
Backhaul Network
Neighborhood Area Network (NAN)/ Field Area Networks (FAN) - AMI
Neighborhood Area Network (NAN)
HAN, BAN, IAN
Substation LAN
Substation LAN
Smart Meter
Workforce Automation
Core/Metro-Backhaul Demarcation
Backhaul-NAN/FAN/ AMI Demarcation
Field Area Network (FAN)
NAN/FAN/ AMI-xAH Demarcation
HAN – Home Area Network
BAN – Building Area Network
IAN – Industrial Area Network
Communications Layer
Smart Grid Network Technology & Protocols Standards Mapping
Technology Standards
Copyright © 2010 IEEE

Wide Area Network (WAN) (public/private)					(NAN/ FAN) AMI Networks (public/private)			Smart Meters	HAN, BAN, IAN	
Substation	Core/Metro Network		Backhaul Network		Substation	(licensed/unlicensed)				
	wireline	wireless	wireless	wireline		wireless	wireline		wireless	wireline
DNP3/ IEC 61850 LAN (several options)	IP-MPLS/ T-MPLS SONET-SDH/ STS-Mesh/ DWDM Packet/Metro-Ethernet	Wimax 802.16d/e Trunked Radio 3G-3GPP/1XRTT/EVDO GPRS/EDGE/HSDPA Wimax 802.16d/e Mesh RF/mm-Wave RF Pto-to-Pto/MAS	802.16/ LMDS	GPON/ EPON RFoG-DOCSIS Metro-Ethernet DSL/POTS/PDH	DNP3/ IEC 61850 LAN (several options)	RF Mesh/ 802.15.4g Wimax 802.16d/e 3G-3GPP/1XRTT/EVDO GPRS/EDGE/HSDPA RF Radio Pto-Mtp/MAS WLAN 802.11 n/g 802.2.15.4/ ZigBee	PON (GPON, EPON) RFoG-DOCSIS PLC/ BPL P1901	ANSI C.12.22	802.2.15.4/ ZigBee 6LowPAN 802.11	ITU-T SG15 G.hn HomePlug

Figure 4. Smart Grid Network Technology Mapping [2]

VI. Wi-Fi [1]

Wi-Fi is cost effective, scales to cover large geographies, mature and proven technology that implements many of the Smart Grid application scenarios today.

Wi-Fi networks can be deployed to meet the Smart Grid requirements for robustness, manageability, performance and security. Moreover, Wi-Fi technology has an ongoing roadmap of innovation and established mechanisms for collaboration (via the Wi-Fi Alliance and IEEE) to meet the evolving needs of Smart Grid applications well into the future.

A. *Wi-Fi CERTIFIED PROGRAM*

The Wi-Fi Certified Program tests devices based on the 802.11 family of standards for interoperability and quality. The Wi-Fi Certified Program provides a widely-recognized designation of interoperability and quality and has contributed to the success of Wi-Fi technology. Key attributes of Wi-Fi include:

- ✓ Mature technology with more than a billion nodes already deployed [4];
- ✓ Mechanisms to deliver robust performance in shared-spectrum and noisy RF environments including listen-before-talk protocol, RF noise awareness and reporting, and received signal strength;
- ✓ Transports all IPv4 and IPv6-based protocols, thereby supporting all IP-based applications;
- ✓ Extensive radio performance and network management mechanisms to provide radio link quality, history reports and channel selection optimization;
- ✓ Low costs due to economies of scale: Wi-Fi chipset shipments now exceed one million units per day and will grow past one billion units per year by 2011 [4];
- ✓ One standard that allows implementation of several interoperable performance/power dissipation profiles;
- ✓ Rates ranging from 1 Mbps (802.11b) to 600 Mbps (802.11n);
- ✓ Networks can scale from a single pair of devices to thousands of access points and clients;
- ✓ Security protections: Link-, network-, and application-level security based on international standards which meet FIPS 140-2 certification [5];
- ✓ Rogue device and intrusion detection tools.

VII. Smart Grid Communications Networks and Wi-Fi [1]

The Smart Grid communications network is typically partitioned into three segments: Home Area Network (HAN), Neighborhood Area Network (NAN), and Wide Area Network (WAN/Backhaul). Wi-Fi technology addresses all three segments.

A. Wi-Fi in the Metropolitan Area Network

The MAN for Smart Grid will aggregate data from multiple Neighborhood Area Networks and convey it to the utility private network. Such "backhaul" can be implemented via point-to-point and/or point-to-multipoint and/or multipoint-to-multipoint wireless links. Smart Grid WANs may cover a very large area and could aggregate ten thousand supported devices. Multi-megabit capacity will be required, and the links involved may range from sub-kilometer to multi-kilometer distances.

Existing city-wide deployments of Wi-Fi networks demonstrate the clear applicability of Wi-Fi as a Smart Grid MAN technology.

Today such metropolitan area MAN incorporating standard 802.11 Wi-Fi in point-to-point or point-to-multipoint or multipoint-to-multipoint links embody a variety of proprietary network management approaches, demonstrating that Wi-Fi technology could be similarly incorporated into the future standardized Smart Grid management framework for MAN communication.

A key advantage of Wi-Fi for the Metropolitan Area Network Smart Grid is its use of free, unlicensed spectrum.

This makes it practical for a city or utility to own and operate a large private wireless network for Smart Grid. WiMAX and LTE for cellular data networks can provide the required service, but are usually owned and operated by large carriers who pay for the frequency licenses that are not designated to utilities services, according to the Brazilian Telecommunications Regulatory Agency (Anatel).

B. Wi-Fi Network Measurement and Management

All wireless networks require network measurement and management to help optimize RF performance.

Wi-Fi has extensive network management capabilities that integrate with existing enterprise network management systems and that make Wi-Fi suitable for very large scale deployments in the Smart Grid.

Ensuring that these networks securely and reliably serve the hundreds of thousands of users that rely upon them requires both autonomous and centrally administered problem diagnosis and performance optimization.

Wi-Fi Network Management systems in place today provide:

- ✓ Visibility into device performance and usage;
- ✓ Historical trend reporting;
- ✓ Threshold-based alerts;
- ✓ Scheduled events and reports;
- ✓ Device configuration and reconfiguration, including multi-vendor management when networks are comprised of wireless devices from more than one manufacturer;
- ✓ Centralized software updates.

One major addition to the IEEE 802.11 standard, 802.11k, provides network measurement protocols. A second addition 802.11v, adding a new suite of network management capabilities.

These two standards extend existing commercial device and management system product capabilities, allowing Wi-Fi devices to measure and report radio link and traffic characteristics. Availability of this information enables optimization in performance and reliability through both local responses (e.g. transmit power control and radio channel change of a wireless Access Point or client device) and centralized management of these extended Wi-Fi networks.

C. Enterprise Mobility for Utility Companies

Utilities will need to have high-speed, reliable communications throughout their service territory for Smart Grid communications. The utility's private Wi-Fi NAN used for AMI could also be used to carry both voice and data to support mobile applications for service technicians and field personnel, complementing existing cellular data and voice networks. Wi-Fi is an obvious choice for this wireless network since low cost interoperable Wi-Fi clients are available and already integrated into mobile phones, laptops, and tablet PCs.

D. Wi-Fi for Smart Grid: Gap Analysis

As discussed in the Report to NIST (National Institute of Standards and Technology) on the Smart Grid Interoperability Standards Roadmap, gap analysis is a critical part of the overall Smart Grid protocol requirements determination [6]. Of particular importance here with respect to the IEEE 802.11/Wi-Fi standard is the identification of any potential gaps with regard to Smart Grid application support.

1) IP Protocol Support

All Wi-Fi devices support the Internet Protocol (IP), both IPv4 and IPv6. No new work is needed for full support of IP.

2) Smart Energy Profile

The Smart Energy Profile helps build a framework for Smart Grid Applications. Version 2.0 of the Smart Energy Profile is PHY independent and thus transportable by Wi-Fi. The only work required for an implementer would be to port SEP 2.0 software to Wi-Fi-based devices.

VIII. Wi-Fi backhaul

In the current status of IEEE 802.11n technology, there is no available 4x4 MIMO with 4 streams per transmission antenna. In these conditions the maximum data rate is 600 Mbps either in 2.4 GHz or 5.8 GHz in unlicensed spectrum.

An alternative is dual-radio node operating in 2.4 GHz and 5.8 GHz each in "bonded" mode. In this mode, both radios are combined to operate as a single unit that provides double the bandwidth of a single radio equivalent. The performance is maximized to the same maximum data rate of 600 Mbps reached using 3x3 MIMO with 2 streams in each radio (2.4 GHz and 5.8 GHz).

This solution is available in the market with both radios in mesh mode. Wi-Fi mesh is a very interesting backhaul solution because offers natural redundancy first due to the C(n, k) link possibilities in a random wireless channel and second due to have two active links (2.4 and 5.8 GHz) operating simultaneously to guarantee link continuity in the case one of them turns off.

IEEE 802.11n technology operates in unlicensed spectrum and so is subject to interference. Although, it is designed with 52 OFDM data subcarriers to be relatively resilient against interference in uncontrolled spectrum, the recommendation is to have 100% first Fresnel zone clear raising the tower height.

Tower is an expensive element of infrastructure but nowadays is available in the market fiber-glass pole with 50-80 lifetime to be mounted in up to 3 sessions of 12 m each one resulting in a pole of 36 m and cheaper than a tower of same height.

IX. Eletrobras Smart Grid Pilot Project of Parintins

Parintins is a city in the far east of the Amazonas State in Brazil. The population for the entire municipality is 102,066 and its area is 7,069 km² [7]. The city is located on Tupinambarana island in the Amazon River. Parintins is known for a popular folklore festival held there each June called Boi-Bumbá.

Eletrobras is a Brazilian Federal Government Company focused on electric power generation and transmission areas that is initiating a pilot project in Smart Grid area in Parintins island aiming DA (Distribution Automation), AMI (Advanced Metering Infrastructure) and DR (Demand Response) supported by a backhaul presented in the item VIII.

X. Conclusions

In a Smart Grid project a mesh backhaul in unlicensed spectrum represents a good strategy considering: a) Wi-Fi mesh is a very interesting backhaul solution because offers natural redundancy, high capacity and is much cheaper than carrier-grade point-to-point digital radios to licensed or unlicensed spectrum; b) The highest capacity in the mesh solution available in the world market since such conditions are reached only working with the newest technologies of the IMT-Advanced (International Mobile Telecommunications Advanced) as WiMAX or LTE both projected to the licensed spectrum.

References

[1] http://www.wi-fi.org/knowledge_center_overview.php?docid=4686

[2] http://ewh.ieee.org/r6/scv/comsoc/Workshop_092510_EnablingSmarterGrid.pdf

[3] https://mentor.ieee.org/2030/dcn/10/2030-10-0053-00-0011-p2030-smart-grid-comms-reference-architecture-status-update.pdf

[4] ABI Research, Wi-Fi IC Market Data, 3Q 2009.

[5] Sheila Frankel et al., “Establishing Wireless Robust Security Networks: A Guide to IEEE 802.11i,” February 2007, NIST Special Publication 800-97, http://csrc.nist.gov/publications/nistpubs/800-97/SP800-97.pdf (8 September 2009).

[6] Electric Power Research Institute, 10 August 2009, “Smart Grid Interoperability Standards Roadmap,” http://www.nist.gov/smartgrid/Report%20to%20NISTlAugust10%20(2).pdf (8 September 2009).

[7] http://www.parintins.com/docs/parintins/?p=historiadeparintins

Improving Performance of Transmission Control Protocol for Mobile Networks

Dulal Kar, Swetha Pandala, and Ajay Katangur
Department of Computing Sciences, Texas A&M University-Corpus Christi, Corpus Christi, Texas, USA

Abstract - *Advances in wireless communications have enabled access to the Internet for mobile users from anywhere, anytime using wireless networks. However, the underlying traditional Transmission Control Protocol (TCP), which is tuned for wired networks, involves mechanisms that often lead to many problems for wireless networks including frequent disconnections of mobile hosts and low throughput due to high bit errors. Recent research to improve the performance of TCP in wireless networks has led the development of Mobile Transmission Control Protocol (M-TCP) as an extension of TCP for mobile networks, which can support multimedia services over high bandwidth links. Similarly, the Snoop protocol has been designed to improve the performance of TCP over networks that have both wired and wireless links. M-TCP and Snoop protocols are some of the improvements made earlier to the TCP to improve its performance over wireless networks. In this work, we propose an enhanced Mobile Transmission Control protocol that exhibits improved performance of TCP when a mobile node uses Internet services. Our experimental results suggest improved performance of the approach over existing M-TCP.*

Keywords: Transmission control protocol, wireless networks, mobile TCP.

1 Introduction

In recent years, the rapid increase of mobile computing devices such as personal digital assistants (PDA), personal computers, and laptops has driven a revolutionary change in the computing world. Internet services have grown rapidly and millions of people are using these services in their day to day life. The proliferation of mobile computing devices with improved processing capabilities allows mobile users to connect to the global Internet. The impact of this phenomenal growth has changed the modality of communication and increased its challenges. The most common issues are management of the wireless communication and poor performance of the existing protocols over wireless networks. These problems often act as the major obstacles for the large-scale deployment of wireless technologies.

The Internet uses the Transmission Control Protocol (TCP) as the transport layer protocol for communications over wired, fixed node networks. TCP is a connection-oriented protocol that provides a reliable service over the Internet. Many applications and services that make use of application layer protocols such as HTTP, FTP, SMTP, and telnet use TCP for communication over the Internet. Besides reliability, TCP provides network congestion control service that uses some control mechanisms to avoid network congestion. These mechanisms control the flow of traffic and keep the data flow below a certain rate that would cause the congestion [1]. However, the TCP congestion control protocol can cause performance degradation in mobile networks.

In a mobile computing environment, there is a combination of wired networks and wireless networks as shown in Figure 1.

Wireless networks are prone to frequent disconnections because of high bit error rates and the frequent hand-offs. Since the traditional TCP protocol is designed for wired hosts, any delay in ACK (acknowledgement) for a transmitted segment of data is normally viewed as congestion. To recover from such congestion, TCP defensively slows down the rate of transmission of segments over the network by invoking a congestion control algorithm [1]. A congestion control algorithm reduces the sender's window (meaning the sender TCP reduces the rate at which segments are sent) which in turn reduces the load to the network to relieve congestion. However, in mobile networks the delay in ACK or no ACK from the receiver may not be due to congestion, but due to the momentarily unreachable or out-of-range location of the mobile receiving/sending node in the wireless network or due to frequent transmission errors suffered by the wireless link. This can result in a significant reduction in the throughput in an active connection [5].

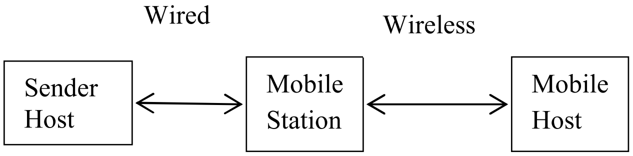


Figure 1. Mobile Networks

The connection-oriented service of TCP leads to several shortcomings. The major shortcomings with the TCP in wireless networks are waste of available capacity during slow-start process and corruption of segments due to high bit errors. Disconnections are common in mobile networks which result in loss of segments. Another problem is serial timeouts which result from continuous retransmission of segments at the time of connection loss [2]-[8],[13][14].

Various modifications have been proposed to overcome the difficulties of using TCP for mobile networks. In this work, we propose a new algorithm called Enhanced Mobile Transmission Protocol (EM-TCP) which is a significant enhancement of the Mobile TCP proposed earlier by Brown [5]. This protocol retains all the salient and necessary features of TCP related to flow control. It uses a split connection approach to implement the protocol. The EM-TCP improves the performance of the TCP connections when a mobile node uses the services across the Internet.

In the following, in Section 2, we review existing solutions for improvement of TCP over wireless networks. In Section 3, we qualitatively analyze the Snoop and M-TCP protocols, identify their drawbacks, and provide possible ways to improve TCP's performance further. In Section 4, we discuss the network model used for simulation study of our proposed EM-TCP protocol. Finally, in Section 5 we provide and discuss our simulation results. Section 6 concludes the paper with some future research directions.

2 Related Works

Several mechanisms have been proposed to improve the performance of TCP over the wireless network. Different properties between the wired network and the wireless network are considered by these mechanisms. For example, the Snoop Protocol is a TCP protocol designed to improve the performance of TCP over networks that have both wired and wireless links. This protocol deals with the problem of segment loss due to network congestion [3]. The Snoop protocol works by deploying a Snoop agent at the base station, performing retransmissions of lost segments based on duplicate ACKs (as duplicate ACKs are strong indicators of lost segments), and locally estimating last-hop round-trip times [4].

The end-to-end semantics of the transport layer connection is maintained in the Snoop protocol [2] [3]. The packets passed across the wired-wireless link are buffered at the base station. The buffered packets are used to retransmit unacknowledged packets and reduce the number of timeouts by suppressing the duplicate ACKs. When an ACK is received from the mobile host, Snoop distinguishes it as genuine, spurious, or duplicate and performs the appropriate action. Snoop avoids timeouts and maintains a larger value of TCP's congestion window, thus resulting in better throughputs [14]. It can improve TCP performance quite well in wireless links but has a problem; when there are no duplicate ACKs, the Snoop protocol cannot notice the segment loss until the local retransmission timer is expired [6] as it cannot identify the disconnection due to handoff.

M-TCP (Mobile-TCP), an enhanced protocol over Snoop, provides better performance in cases of frequent disconnections, changing bandwidths, and low bit wireless links [8]. The mechanism for M-TCP essentially splits TCP into two protocol blocks: Mobile TCP (M-TCP) and Supervisory Host TCP (SH-TCP). Figure 2 shows how TCP is split. At the sender's side, TCP protocol remains unchanged and at the supervisory host (SH), it uses a modified TCP called SH-TCP to communicate with the sender. The M-TCP is used for communication between a mobile host (MH) and the supervisory host.

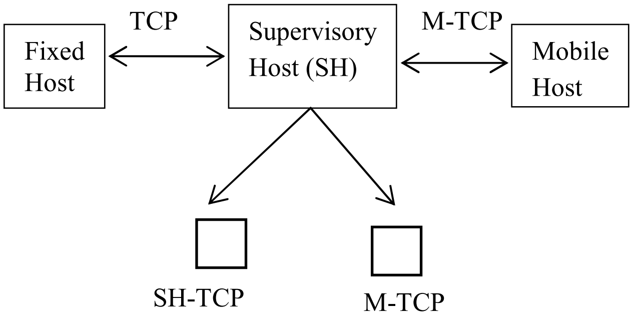


Figure 1. Splitting a TCP connection [5].

The end-to-end semantics are maintained in the M-TCP approach. Upon receiving a segment from the fixed host, the SH-TCP does not immediately send any ACK of the segment to the fixed host (wired network) unless the segment is transmitted to the mobile host (wireless network) and the segment is acknowledged by the mobile host [5]. When M-TCP does not receive an ACK from the mobile host for a transmitted segment, it causes the sender (fixed host) to be in the persistent state where the sender sets its window size to zero. In this state the sender does not suffer from time-out and does not slowly reduce its window size. When the mobile host reconnects, the supervisory host sends a greeting segment to the fixed host with an ACK for the previously sent segment of data. Once the greeting segment is received by the fixed host, it sets its congestion window size to its previous window size. There are certain drawbacks with the M-TCP protocol. When a segment is lost, it is retransmitted by the fixed host in the wired network, which is retransmitted back by the supervisory host to the wireless network. Although this process of retransmission of lost segments is effective when the mobile host moves temporarily out of the range but this model does not help in improving throughput when loss of segments occurs due to bit errors.

3 Enhancing Mobile TCP

In the previous section, we indicate some shortcomings of the Snoop and M-TCP protocols in the context of wireless networks. In the following, we show how further improvement of throughput over M-TCP and Snoop is possible. We assume the same split-connection network model as used in the Snoop and M-TCP protocols. Let us consider the following notations:

RT_{fs} : roundtrip time from the fixed host to the supervisory (or intermediate) host,

RT_{sm} : roundtrip time from the supervisory host to the mobile host,

RT_{fm} : roundtrip time from the fixed host to the mobile host, and

RTO : TCP retransmission timeout at the fixed host.

Typically, a mobile host is in close proximity of the supervisory host and therefore, the propagation delay is very small. On the contrary, the propagation delay over a long-hauled internet path is very large. Also, a wireless link bandwidth is much greater than the effective bandwidth of a long-hauled internet path. As a result, for most situations, RT_{sm} will be significantly smaller than RT_{fs}. This fact justifies the reason of having an intermediate, supervisory, or proxy TCP host as the endpoint of the wired part of the path that can buffer and retransmit the segments on behalf of the fixed host to improve throughput.

The Berkeley Snoop module running on the supervisory host inspects the header of all TCP data and ACK segments as well as buffers copies of all data segments. It also forwards data and acknowledgement segments in both directions. When Snoop detects a duplicate acknowledgement (originating from the mobile host) which is an indication of lost segments, it checks its buffer to retrieve the lost segments, if any, and then retransmits them over the wireless link to the mobile host. By doing local retransmissions, it saves at least RT_{fs} amount of time for each case of successful detection. However, to detect each case of lost segments, the Snoop module needs to wait at least ($RT_{fm}/2$ + $RT_{sm}/2$) amount of time. In order to detect a duplicate ACK, it has to allow the fixed host to timeout and to retransmit data at least once, which takes $RT_{fm}/2$ time (assuming symmetric path bandwidth) to reach the mobile host. It takes another $RT_{sm}/2$ to receive the duplicate ACK from mobile host. In addition, it also causes of shrinking of the congestion window since the fixed host is made aware of loss of segments, which in turn causes the fixed host to slow down its data transmission. Evidently we observe some opportunity to improve the throughput by devising a protocol that can reduce such wait time of ($RT_{fm}/2$ + $RT_{sm}/2$). Accordingly, we take this opportunity to incorporate techniques in our proposed EM-TCP protocol to improve TCP's performance.

Snoop maintains a roundtrip timer and retransmits unacknowledged segments accordingly to the mobile host. In order to prevent the fixed host invoking congestion control, Snoop also maintains a persist-timer of 200 ms whose expiration triggers retransmission of some data from the supervisory host itself. It is reported that Snoop performs well in high BER environments as well as for bursty loss of 2-6 packets. However, Snoop performs poorly when the wireless link experiences frequent and lengthy disconnections. One of our goals in the proposed EM-TCP protocol is to address such issues using similar ideas as found in M-TCP.

Like Snoop, M-TCP uses the split TCP network model. However, the design of M-TCP does not emphasize how to improve TCP's throughput under high bit errors in wireless environment. Instead, it is designed to improve TCP's performance in a situation a mobile host experiences long or frequent disconnections. Frequent disconnections can cause serial timeouts at the TCP sender, thus in turn triggering its exponential back-off action on its retransmission timer for the retransmission. M-TCP chokes the TCP sender when the mobile host is disconnected and allows the sender transmits at full speed by manipulating the TCP sender's window. The manipulation of the TCP sender's window by M-TCP is done by controlling acknowledgements transmitted to the sender in a specific way such as sending ACKs with the receiver's window size to 0. To handle disconnections, M-TCP uses a modified version of TCP on the mobile host that sends a reconnection ACK when the mobile host regains its connection.

As reported in [5], M-TCP performs very well under environments of disconnections and low bandwidths. However, the performance of M-TCP in environments of high bit errors is not as good as Snoop. Another drawback with the M-TCP protocol is that, when a packet is lost, it has to be retransmitted by the TCP sender, thus incurring extra time since the lost segment has to be transmitted by the TCP sender, but not by the supervisory host, which incurs an RTT (round-trip time) including the time to retransmit and to receive ACK by the fixed host. In our proposed protocol, we handle this situation more efficiently. Also, the retransmit and persist timers can be controlled more effectively to increase the throughput, which we incorporate in our EM-TCP.

We propose an Enhanced M-TCP protocol that provides solutions that addresses the drawbacks found in Snoop and M-TCP protocols. The drawback with the M-TCP protocol is that, when a packet is lost, it is retransmitted by wired networks. In case of bit errors and data loss in wireless network we do not have to force the wired network to retransmit just because there is no ACK from the mobile host. We could just as well try to retransmit the lost segment multiple times from the supervisory host so that even if one copy of the segment is in error another might reach the mobile host. The M-TCP model does not consider much about the data lost during the transmission in a wireless network. If there is no ACK, it makes the sender to retransmit the packet again.

Our enhanced mobile transmission protocol improves the performance of TCP throughput over the wireless Internet based on the end-to-end approach. The main goal in developing EM-TCP protocol is to lower the effect of high error rates on throughput by means of dynamic connection migration while keeping the benefits of M-TCP to handle disconnections efficiently. To some extent, EM-TCP attempts to achieve the benefits of both Snoop and M-TCP by merging the techniques of both protocols. It characteristically improves TCP performance over Snoop and M-TCP in mobile networks while maintaining end-to-end TCP semantics.

In EM-TCP, no congestion control is performed over the wireless link. Instead, EM-TCP aggressively retransmits the same segment multiple times with progressively smaller and smaller time intervals in between two successive transmissions. This alleviates the problem of high bit errors and disconnections invoking the wired network to retransmit the packet again thereby increasing the throughput. The wired network does not have to retransmit the data lost in the wireless network due to high bit error rates and frequent disconnections unless the mobile host becomes disconnected for a long time, i.e., more than the round-trip-time. In that case, the TCP sender is set into the persist-state by sending an ACK packet with the receiving window size set to 0, much like the way M-TCP does.

In the following we discuss the network model and a simulation study for evaluating the performance of the EM-TCP protocol.

4 Network Model for EM-TCP

EM-TCP is based on the same network model as used in M-TCP or Snoop protocol. It allows modifying TCP on the mobile network to increase the throughput in disconnections and varying bandwidth as well as to handle high bit error rates [3][5]. Figure 3 shows the network path and elements for our simulation study of EM-TCP to evaluate its performance.

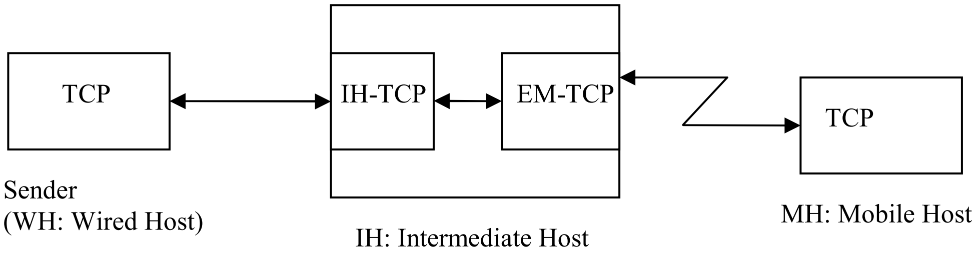


Figure 3. EM-TCP network model.

In this study of the TCP protocol, a closed loop network is considered. The wired host (WH) is the sender who transmits data and uses the ACKs received as feedback from the mobile host (MH) to increase or decrease its congestion window size (*cwnd*). The ACKs received from the network are used in TCP congestion control and flow control mechanisms. Hence, in case of disconnections a signal should be sent as indication. Then TCP could react appropriately by preventing unnecessary deflation of the *cwnd*. In this way, the available bandwidth of the network can be preserved for other TCP communications. During the data transfer the WH connects to the intermediate host (IH) through a WLAN link, the IH is connected to the wired network.

Figure 4 shows the state diagram of the EM-TCP module running at the intermediate host. When a segment is received from the wired host, the intermediate host buffers the segment and then transmits it to the mobile host. The intermediate host when receives some acknowledgement from the mobile host, it checks for any duplicate ACK and accordingly only forwards non-duplicate ACKs to the wired host. It also transmits over the wireless link the same segment multiple times, progressively with shorter and shorter time intervals between two successive transmissions. These time intervals are far less than the retransmission timer at the wired sender. In the process of resending the data packets, the duplicate ACKs are rejected when received. When the retransmission timer expires then the EM-TCP is moved to a persistent state. At this state, the wired TCP sender is set to persist. Then EM-TCP waits for ACK from the mobile host. Since EM-TCP at IH only enters the persistent state when the data is not acknowledged, the mobile receiver needs to send an ACK to remove it from the persistent state. For that EM-TCP keeps sending data to the mobile host until it receives ACK.

We simulate TCP, M-TCP, and EM-TCP over UDP (user datagram protocol) by opening UDP sockets for two end hosts (WH and MH) and one intermediate host (IH) [9]-[12]. The default TCP behavior is simulated using UDP to run for WH. However, IH-TCP, EM-TCP, M-TCP, and TCP on the mobile host are all modified versions of TCP (Figure 3). The simulation of EM-TCP to measure its performance includes the following features:

Connection Establishment

For a connection originating on a fixed network, a TCP connection between the fixed host and the intermediate host is completed. Then it initiates and completes the connection between the IH and MH. The connection is initialized by calling the connect method of the Socket. Then the remote address is set by the connect method, and sends the SYN packet. The connection attempt goes through three-way handshaking process.

From the connection point of view, IH is made transparent to the TCP sender in the fixed network. On connection setup, IH creates sockets with local address bonded to both TCP sender and MH's addresses.

round-trip propagation delay estimated between itself and the MH.

Slow-Start Mechanism

The simulation uses the slow-start mechanism. The congestion window size is not reduced on timeouts because they do not occur due to congestion. Due to this, the slow-start behavior is limited to the beginning of the connection.

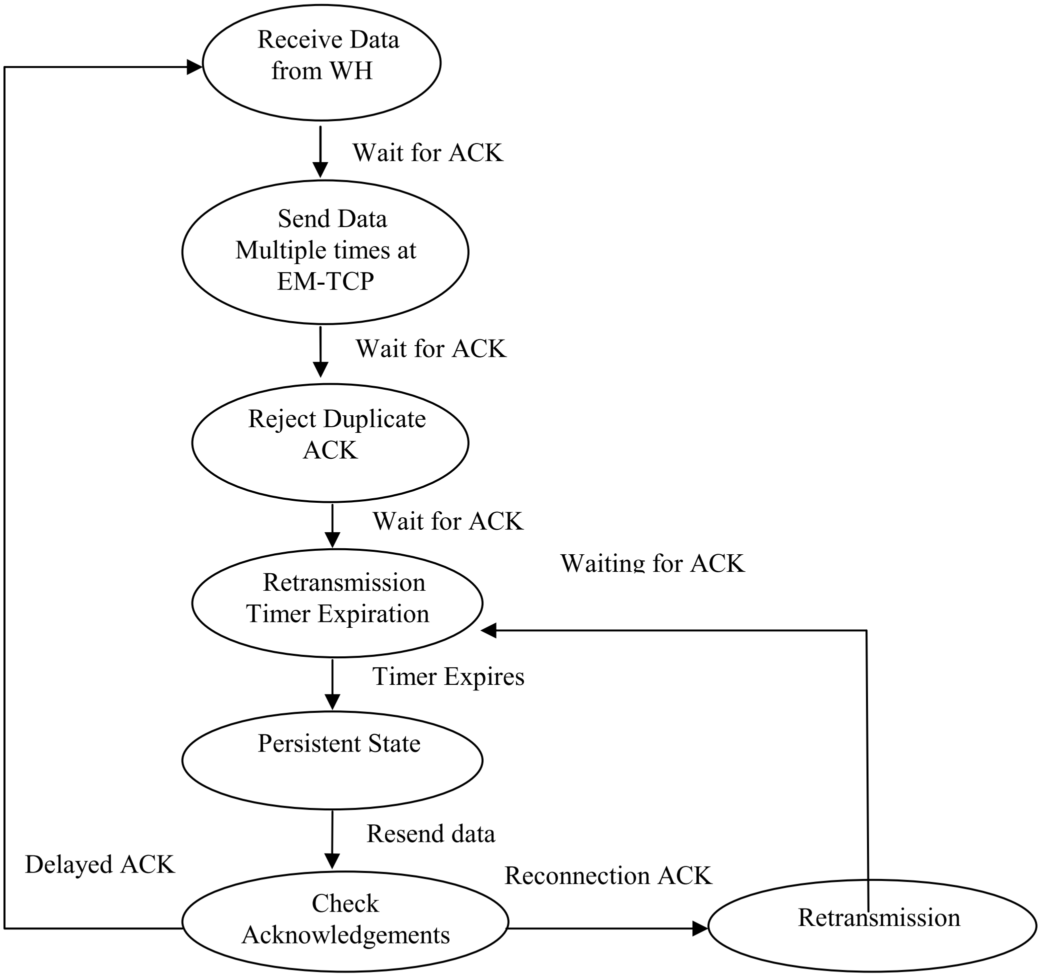


Figure 4. State Transmission Diagram for EM-TCP.

Congestion Window Settings

A timer function determines when to send a window size reduction update to the TCP sender (WH). Before the sender invokes the congestion control and timer expires, IH must generate a packet and send it to the TCP sender to stop invocation of congestion control. Since IH is situated in the middle, IH can estimate RTT between WH and IH and RTT between IH and MH and hence can estimate retransmission timeout (RTO) of the TCP sender retransmission timer. It is to be noted that the wired host's RTO is based on the sender's

5 Performance Testing

The performance of EM-TCP is tested against that of M-TCP protocol. Several data transfer operations at various bit error rates (BER) are performed in the simulation environment and accordingly, data transfer times are recorded for the purpose of comparison of the proposed EM-TCP protocol with the other protocol. Each simulation for a protocol, whether EM-TCP or M-TCP, is run under the same path conditions in terms of bandwidths, propagation delays, BERs and so on. The total time to transfer a data from the fixed host TCP sender (WH) to mobile host (MH) is measured. Particularly, we consider a test case with the following parameters:

Data size = 3 MB,

Maximum segment size = 1024 bytes,

Timeout for EM-TCP or M-TCP to receive ACK from MH = 0.15 ms,

Retransmission timeout (Freeze timeout) at IH or SH = 1 ms,

Persist Timeout at IH = 5 ms,

Maximum interval between two consecutive segments transmitted = 0.05 ms,

Number of hops = 3, and

Wireless link: IEEE 802.11g

The results are shown in Figure 5. For this particular test case, one can see that EM-TCP performs better than M-TCP when the bit errors are limited within 300 segments. However, when the bit errors exceed over 300 segments, both protocols achieve the same or similar performance.

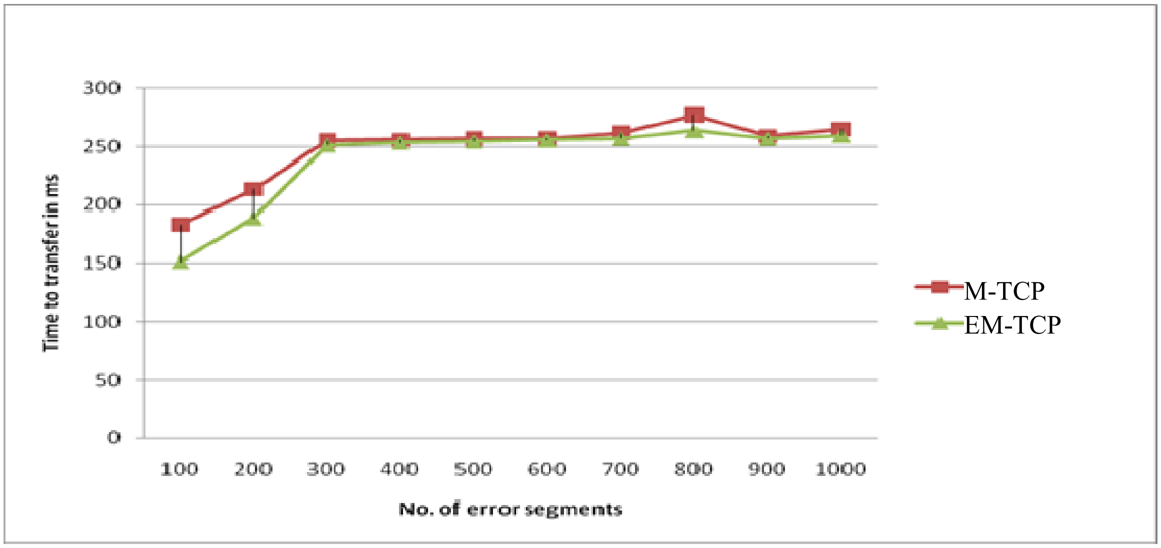


Figure 5. M-TCP vs. EM-TCP under various error conditions.

6 Conclusions

In this work, we propose a scheme to enhance TCP's performance for mobile applications, in which network services are provided over wireless networks tethered to a wired network infrastructure. Based on earlier works on Mobile TCP (M-TCP) and Snoop, we develop a enhanced M-TCP protocol to handle communications from the wired end-point to a mobile host. As our test results demonstrate, the EM-TCP protocol improves TCP's performance in terms of throughput and bandwidth usage in a network and adapt to dynamically changing bandwidth, frequent disconnections, and handoffs over the wireless link. The EM-TCP algorithm reduces the data transfer time, increases the throughput, and reduces retransmission attempts compared to TCP and M-TCP.

The basic assumption of our simulation study of the EM-TCP protocol is that bit errors occur in a continuous stream of segments. That is, several consecutive segments are in error in a stream of segments. This may not be the case in some communication scenario where the packets may be corrupted anywhere in the data. One of the future tasks would be to test the simulation by sending the error segments randomly, and accordingly analyze and explore the new possibilities that arise from this scenario. With the help of this simulation with random distribution of error packets, it is possible to devise a protocol that can offer better performance in all situations.

7 Acknowledgement

This work was supported by a grant from National Science Foundation, under grant number: CNS 0708573.

8 References

[1] M. Allman and V. Paxson. "TCP Congestion Control." http://www.ietf.org/rfc/rfc2581.txt (last visited May 23, 2011).

[2] E. Amir, H. Balakrishnan, S. Seshan, and R. Katz. "Efficient TCP over Networks with Wireless Links." In Proceedings of 5th. Workshop on Hot Topics in Operating Systems, pp. 35–41, May 1995.

[3] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. Katz. "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links." ACM/IEEE Transactions on Networking, vol. 5, pp. 756–769, December 1997.

[4] H. Balakrishnan. "The Berkeley Snoop Protocol." Available from http://nms.lcs.mit.edu/~hari/papers/snoop.html (last visited May 23, 2011).

[5] K. Brown and S. Singh. "M-TCP: TCP for mobile cellular networks." ACM SIG-COMM Computer Communication Review. vol. 27, pp. 19-42, October 1997.

[6] Y.-B. Cho, G.-S. Won, and S.-J. Cho. "Improvement of TCP throughput with the snoop+α protocol." IEEE Computer. pp. 849- 853, 2005.

[7] C. S. Hong, Y. Niu, and J.-J. Lee. "An Adaptive TCP Protocol for Lossy Mobile Environment." Lecture Notes in Computer Science. Springer Berlin / Heidelberg. pp. 466-478, January 2002.

[8] J. R. Ferreira, M. A. Mara, and N. S. Luis. "Mobility over Transport Control Protocol/Internet Protocol (TCP/IP)." The Electrical Engineering Handbook Series, pp. 329-343, 2003.

[9] Available from http://www.eventhelix.com/EventStudio/tcp_ip_sequence_diagrams/Client%20Socket%20Unit%20Test%20Procedures%20-%20TCP%20Project.htm (last visited May 23, 2011).

[10] C. M. Kozierok. TCP/IP Guide. No Starch Press, 2005.

[11] B. Kurniawan. "Using .NET Sockets." Available from http://www.ondotnet.com/pub/a/dotnet/2002/10/21/sockets.htm (last visited May 23, 2011).

[12] B. S. Mitchell. "Application Development with TCP/IP." Available from https://www.cs.drexel.edu/~bmitchel/course/mcs721/tcpipad.pdf (last visited May 23, 2011).

[13] S. A. Mondal and B. F. Lugman. "Improving TCP performance over wired–wireless networks." Computer Networks: The International Journal of Computer and Telecommunications Networking. vol. 51, pp. 3799 - 3811, 2007.

[14] S. Vangala and M. A. Labrador. "The TCP SACK-Aware Snoop Protocol for TCP over Wireless Networks." IEEE Computer. vol. 4, pp. 2624- 2628, October 2003.

A Novel Open Platform for Interoperability of Heterogeneous Personal Health Devices in u-Health care Environment

Minwoo Jung and Jeonghun Cho
School of Electronic Engineering, Kyungpook National University,
Daegu, Republic of Korea
jungminwoo80@gmail.com, jcho@ee.knu.ac.kr

Abstract – *Advances in Information & Communication technology (ICT) are bringing new opportunities in the field of interoperable and standard-based systems oriented to ubiquitous environments and wearable devices used for digital homecare patient telemonitoring. This paper introduces a novel open platform for u-Health care. A novel open middleware platform assures interoperability between standard and proprietary Medical Devices (MDs). We are used a smart device as gateway. A smart device is decreasing complexity of hardware and software. Then, proposed middleware is migrated to mobile operating system.*

Keywords: u-Health care, middleware, smart device, open platform

1. Introduction

As the number of the aged is growing increasingly, seniors who suffer from chronic disease grow rapidly [1]. Then, u-Health care that patient can manage consistently one's health in their daily life is being studied. u-Health care is a medical system that it measure consistently a vital sign with Personal Health Device(PHD) in their daily life and transmit to medical center. u-Health care facilitate early diagnostic and decreasing medical expenses through efficient health care, it will be solved to lack medical professional.

Communications and interfaces among components of patient monitoring system and between Medical Devices(MDs), become now very important in exploiting all the possibilities offered by the information gathered [2].

A freedom in elaborating high quality sensors combined with user's centered features raises the number of MDs introduced by each manufacturer in the market. Such proprietary protocol raise the interoperability problem among MDs. standardization of PHD is necessary in order to assure interoperability among MDs. The ISO/IEEE 11073 consolidates previous IEEE 1073 Medical Information Bus and CEN standards, to cover different levels of the ISO Model, with models for access to the data and with services and communication protocols for interoperability between medical devices [3]. Lack of interoperability among MDs and third-party hospital information system solutions introduces communication overhead, imposes the installation and provision of complex network designs, and limits the monitoring capabilities while moving inpatients for diagnostic examinations [4].

In this context, we propose open platform that it assures interoperability faced when interacting with medical devices that utilize standard or proprietary communication protocol. We utilize smart device for gateway which collect vital sign from MDs and display state of patient's health. Proposed middleware is migrated to mobile operating system.

2. Designed platform

Middleware as a means to simplify application development by providing abstraction of complex low-level concepts has been a topic of interest for many years. Recently, the importance of middleware is considerably growing especially with the emergence of wireless sensor networks, whereby the variety, complexity and dynamicity [5]. Proposed open platform assures interoperability among MDs that utilize standard and proprietary communication protocol.

Wireless communication is essential for implementation of u-Health care. Recently, the release of the Bluetooth Health Device Profile (HDP) has given a strong impulse to achieving interoperability among wireless devices [6]. HDP is a standardized specification for Bluetooth communication between medical devices, mainly targeted to support a variety of in-hospital and in-home healthcare applications. Then, we adopt Bluetooth for communication between MDs and gateway. We are used smart device that function as gateway. Smart devices support Bluetooth without additional module. Proposed open platform consist of MDs and a smart device that function as gateway and a server that store patient's vital sign. MDs transmit smart device measuring vital sign with various sensors. Smart device divide standard transport layer and proprietary transport layer. A data of standard transport layer transmit to application layer through ISO/IEEE 11073 stack. A data of proprietary transport layer convert to ISO/IEEE 11073 format with translate mechanism. Converted data transmit to application layer through ISO/IEEE 11073 stack. Transmitted vital sign display with smart device, and transfer to server through WiFi or CDMA in order to store vital signs.

3. Conclusion

In this context, we suggest middleware framework in order to implementation of open platform of u-Health care. Proposed middleware framework assures interoperability for interacting with medical devices that utilize standard or proprietary communication protocol. From the user, open platform frees about platform and device. From the manufacturer, it facilitate reduce of cost and development time due to can reuse platform. Utilization of smart device in proposed system is fined complexity of hardware and software. Future research aims to deploy combination of u-Health system and medical information system such as HL7, DICOM.

4. Acknowledge

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

5. Reference

[1] Joon-HO Lim, Chanyong Park, Soo-Jun Park, “Home Healthcare Settop-box for Senior Chronic Care using ISO/IEEE 11073 PHD Standard “, 32nd Annual International Conference of the IEEE EMBS, 2010.

[2] S.Pedersen, “Interoperability for information systems among the health service providers based on medical standards”, Informatik-Forschung Und Entwicklung, Vol. 18, pp. 174-188, 2004.

[3] I. Martinez, J. Escayola, M. Martinez-Espronceda, “Standard-based Middleware Platform for Medical Sensor Networks and u-Health”, 2008 proceedings of 17th International Conference on Computer Communications and Networks, 2008.

[4] Nikolas Stylianides, Marios D. Dikaiakos, Harald Gjermundrod, George Panayi, Theodoros Kyprianous, “Intensive Care Window: Real-Time Monitoring and Analysis in the intensive Care Environment”, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, Vol. 15, No. 1, 2011.

[5] Agustinus Borgy Waluyo, Isaac Pek, Xiang Chen, Wee-Soon Yeoh, “design and evaluation of lightweight middleware for personal wireless body area network”, Pers Ubiquit Comput, Vol. 13, pp. 509-525, 2009.

[6] Fioravanti A, Fico G, Arredondo MT, Salvi D, Villalar JL, “Integration of Heterogeneous Biomedical Sensors into an ISO/IEEE 11073 Compliant Application”, 32nd Annual International Conference of the IEEE EMBS, 2010.

Frequential Reconfiguration by Changing the Substrate EBG / PBG

Catarine Medeiros Resende, Antonio J. F. Vieira and Humberto Cesar Chaves Fernandes
Electrical Engeneering Department, Federal University of Rio Grande do Norte, 59078-970, Natal-RN,Brasil
resende.caty@gmail.com , tom_rpm@hotmail.com , humbeccf@ct.ufrn.br

Abstract - *This article presents a technique for reconfiguring the resonant frequency for a microstrip antenna using different substrate materials including EBG / PBG. Initially the substrate was used in 2.2 and later replaced by others of relative permittivity of 8.703 (p polarization) and 10.233 (s polarization). The results were obtained by Ansoft-Design for return loss and radiation pattern in 2D and 3D. Comparisons are made with results from improved for frequencies of 2.36, 6.8 and 7.35 GHz.*

Keywords: Microstrip Antennas, Return Loss, Pattern Fields, PBG, EBG.

1 Introduction

The microstrip antennas, as shown in Figure 1, are widely employed in wireless communication systems, and have been the target of several research studies, because of its wide use in modern technology for its efficiency, low manufacturing costs, easy integration with other devices or other types of antennas. However, it has some disadvantages, problems with bandwidth, sensitivity to environmental factors, low efficiency due to dielectric and conductor losses, and has no reconfiguration for the resonant frequency of the systems that today has fundamental importance [1]-[7].

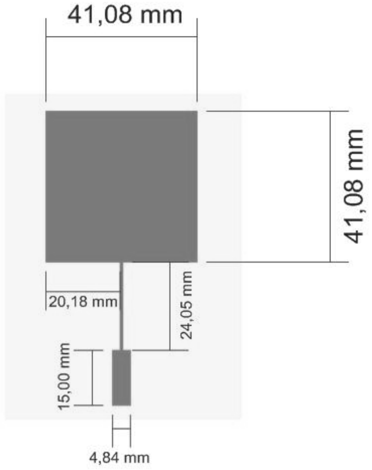


Figure 1. Microstrip antenna designed for εr=2.2.

The use of photonic structures is an alternative technique that although it was developed in research on optical media, can have a concept used in other fields and applied to a range of frequencies, including microwave bands and millimeter waves. This class of artificial materials regularly spaced and with specific electrical properties, can be used in planar devices that characterize the microstrip circuits. This paper presents a way to adjust the antenna by changing its substrate values. A PBG (Photonic Band Gap) or EBG (Electromagnetic Band Gap) [5]-[8] substrate at microwaves frequencies, formed by semiconductor materials with spaced cylinder gaps, are employed. New results are compared with the use of other substrates, where a modification of the same is made, to perform a reconfiguration of the antenna resonant frequency.

2 Theory

Periodic structures for operating in the optical range can be made of dielectric or metallic materials. These structures can generate frequencies bands forbidden. When photons are launched into a crystal of this type, their modes decay exponentially within the structure. This happens due to the wave number become complex (in whose case the modes are evanescent) and, therefore, light is strongly attenuated in all directions of the periodic arrangement. Thus, other structures can be designed to forbidden bands (band gaps) that prevent energy electromagnetic propagation; these structures are referred to as PBG (Photonic Band Gap).[9]-[11].

A periodic structure can have its periodicity in one, two or three directions [2]-[3]. Thus, it can be classified into one-dimensional when its constituent characteristics vary in one direction only, two-dimensional, has periodicity in two directions, or three-dimensional, while its structure is periodic in three directions, shown in Figure 2, with its reciprocal representations.

The homogenizing process of the substrate is made, slice by slice, as shown in Figure 3. The rods with permittivity ε_1 are embedded in a medium of permittivity ε_2. The procedure consists in dividing the structure into a superposition of homogenized layers.

This work was partially supported by CNPQ and INCT-CSF.

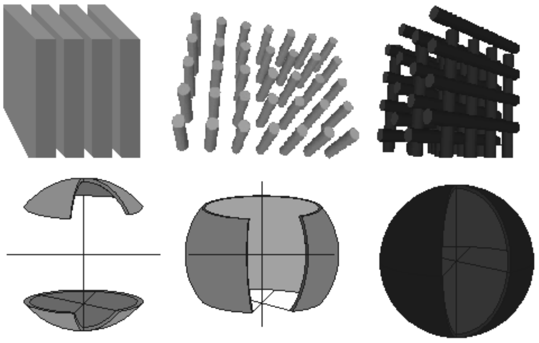


Figure 2. Real and reciprocal representations of PBG/EBG structures: one-dimensional, two-dimensional and three-dimensional.

The layers containing the rods are broken up into cells hose y size (resp. x size) is the diameter of a rod (2r) (resp. the period d).

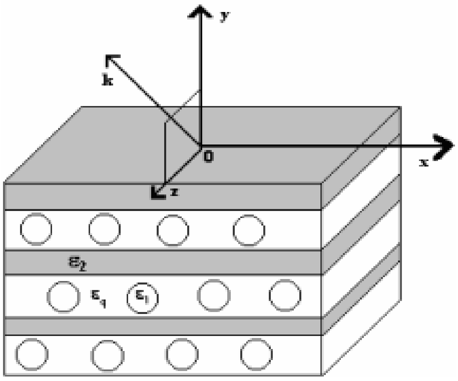


Figure 3. Homogenized bidimensional EBG crystal.

According to homogenization theory the effective permittivity depends on the polarization [9]. For the s and p polarization, respectively, we have (1) and (2):

$$\varepsilon_{eq} = \beta(\varepsilon_1 - \varepsilon_2) + \varepsilon_2$$

$$(1)\ \frac{1}{\varepsilon_{eq}} = \frac{1}{\varepsilon_1}\left\{1 - \frac{3\beta}{\frac{2/\varepsilon_1 + 1/\varepsilon_2}{1/\varepsilon_1 - 1/\varepsilon_2} + \beta - \frac{\alpha(1/\varepsilon_1 - 1/\varepsilon_2)}{4/3\varepsilon_1 + 1/\varepsilon_2}\beta^{10/3} + O\left(\beta^{14/3}\right)}\right\}$$

(2)

β is defined as the ratio between the area of the cylinders and the area of the cells, α is an independent parameter whose value is equal to 0.523.

3 Results

Thus, was analyzed the influences of changing the antenna's substrates with the shift of the resonant frequency. The antenna initially designed to εr = 2.2 has substituted its substrate to promote the reconfiguration frequency antenna. In Figure 4 has the return loss of antenna with substrate of 2.2; in Figure 5 has the radiation pattern plane E in red and H in blue to 2.36 GHz and Fig.6 the 3D diagram of radiation pattern.

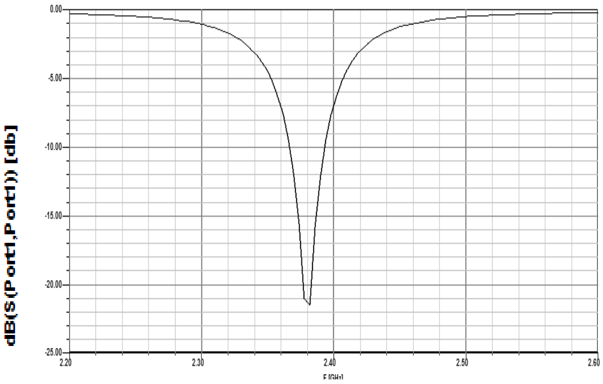


Figure 4. Return loss using substrate with εr = 2.2

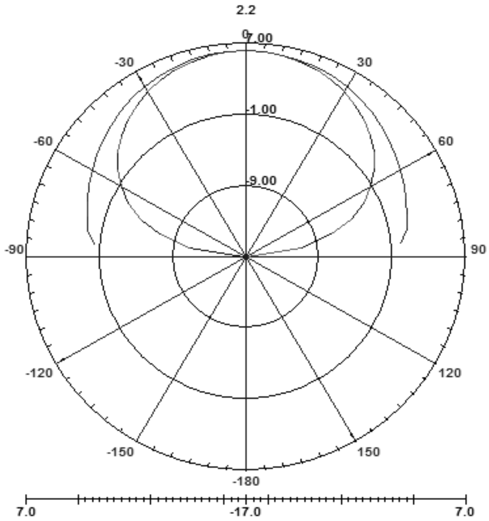


Figure 5. Radiation pattern, E plane the outside line and H plane the inner line, to 2.36 GHz.

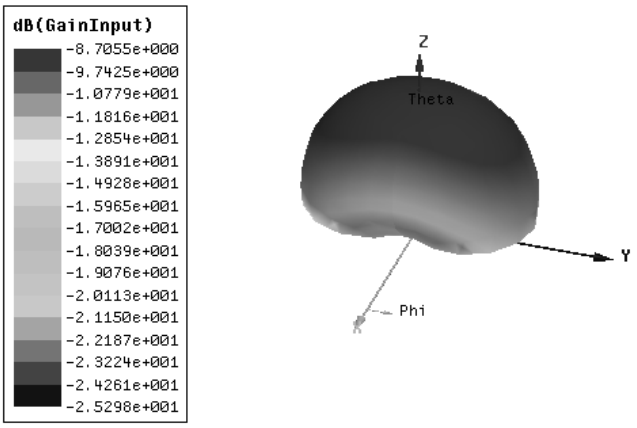


Figure 6. 3D pattern fields of the antenna using substrate with □r = 2.2

Then was replaced the antenna's substrate from 2.2 to 8.702 in order to check the displacement at a frequency of resonance. The Figure 7 present the antenna´s return loss with that substrate.

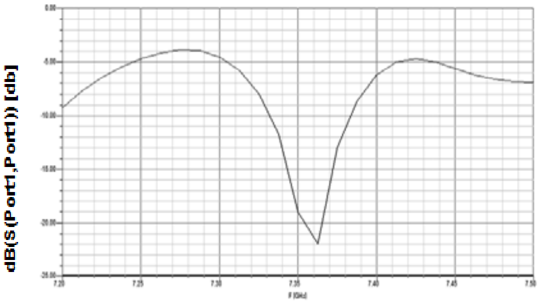


Figure7. Return loss using substrate with □r = 8.702

It was comproved the shift in resonance frequency by 2.36 to 7.36 GHz with the change of the substrate. The Figure 8 shows the radiation pattern to 7.36 GHz and in Figure 9 the radiation pattern in 3D.

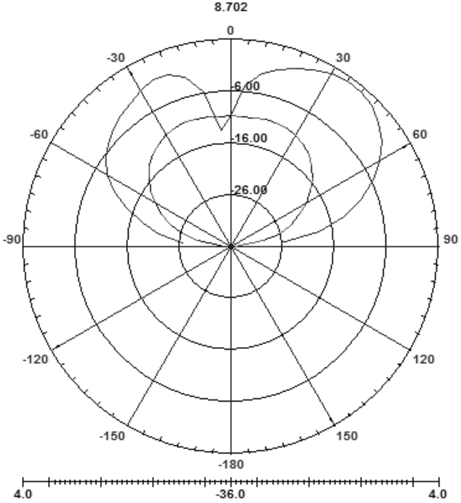


Figure 8. Radiation pattern, E plane the outside line and H plane the inner line, to 7.36 GHz.

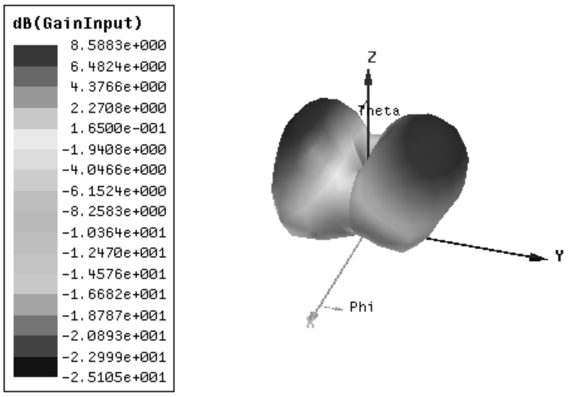


Figure 9. 3D pattern fields of the antenna using substrate with □r = 8.702 (p polarization)

Once again in order to verify the change in the frequency was modified the substrate to 10.233 to prove the effectiveness of the technique in the rewriting frequency of microstrip antennas. In Figure 9 we have the return loss of the antenna with this substrate.

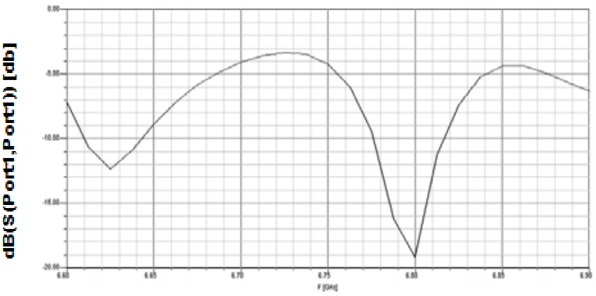


Figure 10. Return loss using substrate with □r = 10.233

Thus there is a shift in frequency to 6.8 GHz proving the change in resonant frequency of the antenna. In Figure 11 has the radiation pattern to 6.8 GHz and Figure 12 the 3D radiation pattern.

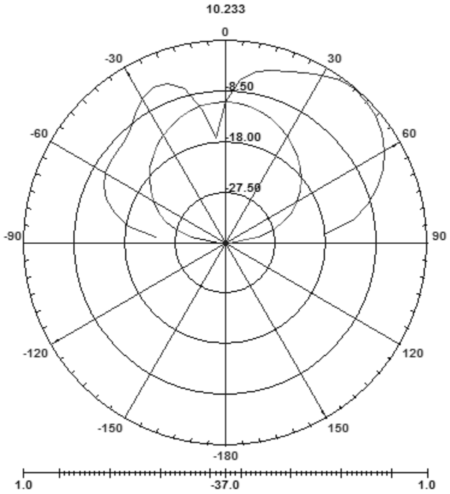


Figure 11. Radiation pattern, E plane the outside line and H plane the inner line, to 6.8 GHz.

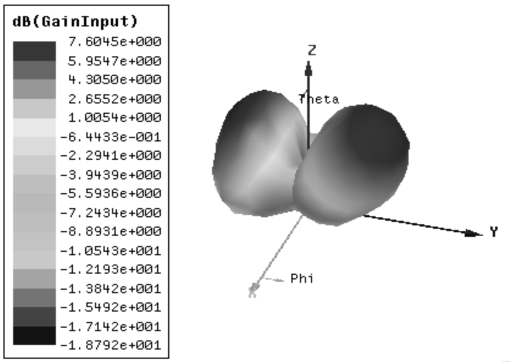


Figure12. 3D pattern fields of the antenna using substrate with □r = 10.233 (s polarization)

4 Conclusions

With the increased usage of wireless systems, several frequency spectra have been detected, and then it has become necessary to build systems adaptable to the environment in which they live. Thus, this research showed development of a reconfigurable antenna of changes in substrates. The antenna originally designed to operate at 2.4 GHz, εr = 2.2 has been changed to a frequency range of 6.8 and 7.35 GHz, with substrate PBG photonic altered, with relative permittivity of 10.233 (p polarization) and 8.702 (s polarization), respectively. Thus, the technique of changing the substrate is shown to provide effective frequency reconfiguration of microstrip antennas. Due to the periodicity of the EBG structures, eliminates surface waves and thereby increases the efficiency of the antenna. These antennas are employed in adaptive systems, and these types of structures have fundamental importance in addressing the problems of surface waves, wrapped in microstrip configurations.

5 References

[1] I.J. Bahl e P. Barthia, "Microstrip Antennas", Artech House, 1982.

[2] Radisic, Y. Qian, R. Coccioli e T. Itoh, "Novel 2-D Photonic Bandgap Structure for Microstrip Lines", IEEE Microwave and Guided Waves Letters, Vol. 8 Nº. 2, Feb. 1998.

[3] P. Damiano, J. Bennegueouche e A. Papiernik, "Study of multilayer microstrip antennas with radiating elements of various geometries", IEEE Proc. H, Vol 137, Nº.3, Jun. 1990.

[4] E. Yablonovitch e T. J. Gmitter: "Photonic band structure: the face-centered-cubic case", Journal of the Optical Society of America A, p. 1792 Sept. 1990.

[5] C. A. Balanis, "Antenna Theory – Analysis and Design". Harper & Row,Publishers, New York, 1982.

[6] J.S. Sun and G.Y Chen, "Efficiency of various Photonic Bandgap (PBG)" , 3rd Internacional Conf. on Microwave and Milimeter Wave Technology", Proceedings CD, Taiwan, China, 2002.

[7] Y. Qian and T. Itoh, "Microwave Applications of Photonic Band Gap (PBG) Structures", Microwave Conference, Asia Pacific, 1999.

[8] M. N. Mollah and N. C. Karmakar, "Planar PBG Structures and their Applications to Antennas", International Symposium on Antennas and Propagation, V. 2, pp.494, Jul, 2001.

[9] H.C.C. Fernandes, R. R. C. Franca and D. B. Brito, "Asymmetric Finline and Coupler at Millimeter Waves on PBG Substrate", Journal of Infrared, Millimeter, and Terahertz Waves, V. 32, Nº 1, pp. 116-125, Jan. 2011.

[10] R E. Centeno and D. Felbacq, "Rigorous vector diffraction of electromagnetic waves by bidimensional photonic crystals", J. Optical Soc. American, Vol. 17, No.2, pp.320-327, Feb. 2000.

[11] H. F. AbuTarboush, H. S. Al-Raweshidy and R. Nilavalan, "Bandwidth Enhancement for Patch Antenna Using PBG Slot Structure for 5, 6 and 9 GHz Applications", Wireless and Microwave Technology Conference, . IEEE 10th Annual, 2009.

[12] K. U. Osama, "Design of X-band 4x4 Butler Matrix for Microstrip Patch Antenna Array", TECNON / IEEE Conf., 2006.

Reputation-based Enforcement Schemes Tailored to Opportunistic Networks Design

A. Modarresi, I. Woungang, L. Reyhani and H. Razavi
Department of Computer Engineering, Ryerson University, Canada

Abstract—*Cooperation among nodes is one of the main concerns in Ad hoc network; and success or failure of a network depends on the rate of cooperation. In this paper, incentive patterns to stimulate nodes for cooperation is reviewed, then it is focused on reputation-based schemes and a survey on some of available techniques is performed. In the next step, the indicated schemes are evaluated based on common parameters in most incentive pattern schemes and finally a proper method tailored to opportunistic network is selected.*

Keywords: Opportunistic networks, reputation, Cooperation, Ad hoc networks

1. Introduction

Although opportunistic resource utilization network or *oppnets* is an innovative paradigm, it inherits many characteristics of its ancestor namely mobile ad hoc networks (MANETs). Even if all seed nodes, as main members of the network, show a cooperative behavior, considering such behavior is not realistic among helpers; therefore existence of selfish behavior among members of the network is inevitable. On the other hand, such networks can be successful in their mission when nodes cooperate with each other. In order to increase incentive pattern in the network, different mechanisms have been introduced in literature, ranging from discovering a selfish node and limiting them, to outstanding a cooperative node in the network. Regarding to Oppnet networks, selfish nodes must also be discovered and their communication should be limited. In this paper, it is tried to review some methods for controlling selfish behavior based on reputation scheme and tailored to Oppnet.

This paper is organized as follows. First, the structure of opportunistic network is explained. Then incentive patterns to stimulate nodes in unstructured networks are reviewed. In section three, a brief survey about reputation-based methods is provided. Section four covers the evaluation of methods introduced in section three. Finally, as the conclusion of this paper, a suitable reputation-based schemes tailored to opportunistic network is chosen.

2. Structure of the Opportunistic Resource Utilization Networks & Their Classifications

Presented recently, oppnets are a specialized form of ad hoc networks with the virtue of expansion. In this paradigm diverse systems which are not included in the original network can join the system dynamically, and provide certain tasks in a specific situation like an emergency incident. In this case, oppnets can increase their potential abilities to fulfill a specific task in a certain period of time. Oppnets are categorized into two main classes, namely class 1 and class 2 and a sub-class between the two main classes called class 1.5 [1]. Class 1 introduces opportunistic communication when devices are in each other range; while class 2 defines opportunistic expansion and opportunistic usage of resources gained by this opportunistic expansion. Class 1.5 is a specialized network facilitated for opportunistic data forwarding [2].

Oppnets are deployed by a set of initial nodes called seeds as an autonomous network; then other nodes, called helpers, can be discovered and joined to the main network. A distributed control center including a subset of seed nodes is created to control the expansion of the network by admitting helpers. Helpers can be invited or ordered to join. In the invited form, a helper can accept or reject the invitation, but in the ordered form a helper is enforced to join the network [1].

In spite of introducing the structure of oppnets, there are still many challenges to overcome. One of these challenges is the nodes stimulation for cooperation. In the following section, incentive patterns and their general characteristics are reviewed to choose a suitable one for oppnets.

3. Incentive Patterns and Its Taxonomy

In order to increase utilization in decentralized unstructured networks two related tasks must be fulfilled, namely stimulating nodes for better cooperation and limiting uncooperative nodes. Stimulating nodes for cooperation is known as an incentive pattern in the literature and isolating and limiting the activity of uncooperative nodes is called punishment. Effective incentive patterns can also decrease selfish behavior.

In [3] incentive patterns are categorized to two main groups, namely trust based incentive patterns and trade based incentive patterns. Prior to explaining each group, some general characteristics to differentiate incentive pattern are expressed. The characteristics plus other attributes are the main criteria to evaluate the reputation-based incentive schemes.

- Roles: it defines the role of an entity in the system. In general, if an incentive pattern stimulates an entity to act as a provider for a consumer, then the cooperation relationship is *asymmetric*. However, if an incentive pattern enforces a consumer to act as a provider at the same time, it is called *symmetric* relationship.
- Remuneration: the way a consumer rewards a provider is called remuneration. There are different types of remuneration like reputation or check. The type of remuneration varies in each incentive pattern.
- Uncooperative behavior: In [4] profitable and reasonable behavior are considered as the main criteria to distinguish the uncooperative entities. In an incentive pattern, misbehaviors are restrained while venial noncooperation are tried to ignore. These uncooperative behaviors can be categorized as malicious, selfish, lavish and venial behavior. *Malicious behavior* is not profitable for malicious entity; therefore, in an incentive pattern such entities are discovered and punished seriously. *Selfish behavior* is usually a one-way profitable relationship for providers. The effect of such misbehavior is controlled by remunerating the provider. *Lavish misbehavior* is a profitable relationship for the consumer and it is diminished by remuneration. *Venial noncooperation* is a reasonable uncooperative behavior and is usually ignored most of the time.
- Trust: regardless of the incentive pattern, a consumer must satisfy that a provider executes a respective service. In order to do so, the provider usually gathers proofs of his work. On the other hand, the consumer must show the validity of the provided remuneration. Regarding to the incentive pattern, the provider must trust either the consumer directly based on provided evidence or third parties' entities.
- Scalability: the number of entities who apply the incentive pattern represent the scalability.

According to characteristics declared above, the incentive patterns are grouped as follows.

Trust Based Incentive Pattern: this is a very straightforward pattern. A provider responds to a consumer, if it trusts the consumer. This group is sub-divided to two sub-groups as follows.

- The Collective Pattern: a set of entities with mutual trust and unconditional cooperation is called a collection of entities. In this category, since all entities are members of a collection, the incentive for cooperation is stimulated. The assumption is that, as long as an entity is a member of the collection, it shows the cooperative behavior. In ad hoc networks, a group of devices belong to one organization or one person represents a sample of a collection. Regarding to Oppnet networks, seeds can be categorized in this group.
- The community pattern: good reputation is the main requirement of this incentive pattern. A group of local entities who have gained reputation by providing services to other entities are called a community. In this pattern increasing local reputation of the provider is the benefit which is gained by executing a service for a consumer. A community is usually formed when devices from different organizations want to cooperate with each other. In Oppnet networks helpers can be categorized in this group.

Trade Based Incentive Pattern: in this pattern an explicit remuneration of the provider might be desirable. This remuneration is usually a service in return by the consumer on behalf of the provider. This service can be provided either immediately or deferred; therefore trade based incentive patterns can be divided to two sub-categories as follows.

- Immediate service in return: in some type of networks like highly volatile networks, the assumption of mutual trust can be too restrictive. In such cases, any future cooperation is not assumed; therefore a consumer must provide in return service as soon as it receives its desired service from provider. In other words, an entity is a provider and consumer at the same time and the mutual services occur simultaneously. In this incentive pattern mutual trust is not necessary and two entities can communicate each other anonymously. Moreover, this method can restrain selfish and lavish behaviors. This method is also called "*the barter trade pattern*".
- Deferred service in return: if a provider does not need a service immediately, "immediate service in return" will not be helpful. In such a case, the consumer hands over a bond to promise a service in return to provider. This incentive pattern is also called *"bond based incentive pattern"*. This pattern can be implemented in four different ways namely, bearer notes, bearer bills, banking pattern and bank note pattern. In all of these methods, an official document like a bearer note or a bank note

specify that a specific consumer must provide a service to the bearer at the time of presenting the document.

In summary, the main difference of "trust based incentive pattern" and "trade based incentive pattern" is that, in the former every entity who can trust other entities can request a service, but in latter, providing a service is limited to some specific entities who get their remuneration by requesting a service in return at once, or in deferred by presenting a bond. In other words, trust, cooperation and consequently selfish behavior in the second category, are more restricted than the first category.

In Oppnet networks which may be usually used in specific situation like emergency incidents, high cooperation is necessary; however, establishing an entity as trusted third party usually needs special hardware and longer time; therefore trust based incentive pattern are suitable as their incentive pattern.

4. Reputation-Based Incentive Schemes

In this section the following reputation-based incentive schemes are reviewed.

- RADAR: a ReputAtion-based Scheme for Detecting Anomalous Nodes [5]
- SORI: A Secure and Objective Reputation-Based Incentive Scheme [6]
- OCEAN: Observation-based Cooperation Enforcement in Ad hoc Network [7]
- Refaei et. al. [8]
- Zakhari et. Al. [9]
- LARS: a Locally Aware Reputation System for MANETs [10]

A. RADAR: a ReputAtion-based Scheme for Detecting Anomalous Nodes

In this scheme, reputation is used to detect anomalous nodes, a kind of intrusion detection technique. Reputation is defined here as opinion of a mesh node from other nodes. Similar to intrusion detection techniques, in anomalous detection, normal profiles must be created by sampling observed subjects in normal events. However, creating these profiles is not easy in Wireless Mesh Networks (WMN) due to the some difficulties like unreliable physical medium, signal interference and traffic congestion. To accomplish this task, reputation is used to quantify the behavior of each node by collecting relevant observable subjects. In this scheme, it is assumed that nodes are heterogeneous regarding to reputation, limited in mobility and autonomous in decision making.

In order to quantify nodes' reputation, nodes rate each other after each communication session. Particularly, each node monitors all its interesting nodes during a certain amount of time. Then local trust of all neighbors of node *i*, is calculated and integrated to represent the local trust of the node. Then, the global trust of each node is calculated using transitive trust. In this way, a number of trustworthy intermediate nodes which are not direct neighbors of node *i*, are involved in the calculation. These values are used to learn the network to detect future anomalous nodes.

RADAR can resist various types of network layer attacks, namely malicious collective, DoS and routing loop. The aim of malicious collective is to subvert reputation management, and the goal of routing loop attack is to compromise the routing protocol to create routing loop. RADAR can detect all nodes creating attacks with 20% false positive. Nodes launching DoS are detected, because their trust values decrease gradually. Nodes launching routing loop attack are discovered, because their trust value doesn't change significantly; and finally nodes creating DoS by spoofing address are punished.

In terms of scalability, all nodes in the system work as a detector to detect anomaly.

B. SORI: A Secure and Objective Reputation-Based Incentive Scheme

SORI quantifies reputation by objective measurement; moreover, this scheme focuses on securing propagation of reputation to resist malicious nodes by using one-way-hash authentication. In addition, SORI's nodes send the reputation values to direct neighbors to reduce communication overhead. Generally, sending reputation values to direct neighbors produce less network traffic and provide more trustable information than the methods using second hand information from intermediate nodes. Reputation inconsistency is one common problem in second hand reputation values, due to the fact that each node may have different views about the reputation of its neighbors. Discriminating trust value is quite confusing, when a node receives different values for a specific neighbor. On the other hand, having a consistent wide-network reputation value makes the decision making more trustable than one-hand values.

SORI assumes nodes are uncooperative in packet forwarding, but no conspiracy among nodes; therefore there is selfish behavior in the network but not malicious one. Moreover, nodes do not change their identity during their life time. The way of transmission is broadcast approach. In term of scalability, all nodes contribute in incentive pattern. Nodes are in a promiscuous mode; therefore each node can listen to every packet transmitted by its neighbors.

In order to calculate the reputation value, each node keeps track of two numbers for each of its neighbors, namely *request-for-forwarding* and *has-forwarded* values. The ratio of these two values defines a metric called confidence to consider reputation. Then each node propagates the confident value to all of its neighbors to increase the effect of punishment. Upon receiving all confident values from each neighbor, overall evaluation value is calculated based on the credibility that each neighbor has earned from perspective of the node. In this way, a selfish node is punished by its entire neighbors instead of one neighbor. The punishment is performed by dropping packets of punishing nodes probabilistically.

In term of security, SORI uses a one-way-hash chain for authentication mechanism. This mechanism keeps the integrity of messages and stops the attack of selfish nodes to impersonate a node with good reputation.

C. OCEAN: Observation-based Cooperation Enforcement in Ad hoc Network

OCEAN uses trade-based scheme for its incentive pattern. Every node has counters called *chipcount* for each neighbor. When a node forwards a packet on behalf of another node, it earns chips. Similarly, when a node asks other nodes to forward a packet, it loses chips. When a node wants to forward a request, it checks the chipcounts of its neighbor. If it is below the threshold, it denies forwarding the packet.

OCEAN calculates reputation from first-hand observation to reduce the burden of trust-management in second-hand information gathering and trustiness. When a node sends a packet to its neighbor, it listens to medium to find out whether the neighbor attempts to forward the packet. If it does not, the sender registers a negative event against the neighbor node; otherwise a positive event is registered. When a rating of a node falls below a faulty threshold, that node is added to the faulty list. Consequently, routes whose next hops are chosen from the faulty-list nodes are considered as bad routes. This list is propagated by Route-Request (RREQ) of DSR routing protocol to inform other nodes.

OCEAN resists against two types of routing misbehavior, namely misleading and selfish behavior. In misleading misbehavior, a node may respond to route request, but fails in actual packet forwarding. This misbehavior is detected when other nodes observe the lack of participation of misbehaved node. Then, this observation is reported to other nodes to increase the effect of punishment by refusing to forward the traffic of the misbehaved node. In selfish misbehavior, selfish node does not even respond to route request, but it sends its traffic to the network. In OCEAN direct observation is the solution to capture this behavior.

In order to increase security in OCEAN, nodes generate their asymmetric keys. The public key is propagated among neighbors and the private key is kept in the original node.

D. Refaei et. al. (M. T. Refaei 2005)

In this scheme, each node calculates the reputation of its neighbors autonomously based on the completion of the requested service. The principal is that when a node sends a packet to one of its neighbor, it holds the responsibility for the node to complete the correct delivery. Consequently, uncooperative nodes are detected, isolated and punished. In order to do so, each node must maintain a reputation table to keep the reputation index of its direct neighbors. When packet forwarding is successful, each node along the path to the destination increases the reputation index of the next hop on the path. On the other hand, unsuccessful delivery decreases the value of the index for each node along the path. In addition, the main criterion for a node to forward or drop a packet is the value of the reputation index of the sender. If this value falls below a designated threshold, receiver drops the packet. Falling down the reputation index below the threshold is a sign for selfish behavior. The amount of values for increasing or decreasing reputation index has a direct effect on the behavior of the network. Assigning higher values causes faster isolation but produces more false positive; therefore a trade off must be considered between accuracy and isolation.

This scheme is independent from routing algorithm; because its principal is based on the feedback returning from destination. Therefore, there is no need for each node to listens to the medium to find out if its neighbor has forwarded the packet. In this way communication can be done as directional form. Moreover, this method doesn’t have any communication overhead, since each node calculates the reputation independently and does not share it with other nodes.

E. Zakhari et. Al. (S. R. Zakhary n.d.)

This scheme has been designed for highly mobile nodes and sparse environment considering security for single and multiple black hole attack. In this model node’s reputation is categorized to multiple zones that provide better decision making and higher details depending on the required services like packet forwarding. Moreover, both direct information and second hand information are used to calculate the reputation. This method uses reputation discounting, a concept to fade out old reputation and provide a chance for new nodes to claim reputation. Moreover, considering degree of centrality for each node can reveal the most influential nodes to assist other nodes to build trust. Nodes with higher centrality have higher probability to getting in contact other nodes. In addition, nodes with high centrality and high reputation are considered as suitable sources for indirect reputation.

This scheme holds *N* neighbor reputation records, one for each neighbor, to represent reputation observation of that neighbor. Each node performs a selective deviation test to ensure unity of view with neighbors upon receiving indirect reputations. In this case, a received indirect reputation is compared with direct reputation of receiving node to detect all deviation greater than a specific threshold to eliminate it. Direct reputation is also calculated by the Eigen Trust algorithm [11]. Then, a global consistent reputation value is calculated at each node for all of its neighbors.

This method also uses adaptive expiration technique for observing the behavior of neighbors. In this way, a node can adjust observation based on neighbor reputation and network conditions. The observation expiration time for trusted neighbors is usually higher than non-trusted neighbors.

F. LARS -A Locally Aware Reputation System for Mobile Ad Hoc Networks

There are dynamically located nodes in Mobile Ad-hoc Networks. The nodes can behave selfishly due to preserving power. According to node selfishness, two concepts can be discussed; extreme selfishness and selective selfishness. The extreme selfishness is what we call when a node drops all the packets which are not for the node itself. However, the selective selfishness occurs when a node drops some packets selectively. A Locally Aware Reputation system (LARS) deals with node selfishness. LARS is able to deal with both extreme and selective selfishness.

In Lars, the reputation of a node can be defined as the perception of a node regarding to the performance of another node. In other words, the reputation of a node is derived from direct observation, while the second hand information is not allowed. In most reputation systems, the reputation of a node is globally shared in the network. However, the local reputation is based on the node’s direct observations of its neighbors. In reputation systems, different nodes may face different reputation values for the same node. There might be different reasons for the inconsistency node problem; nodes may calculate different reputation values, the global reputation values may differ and nodes may receive different indirect reputation.

Reaction to selfish node is done as follows;

If a node observes a reputation value of another node is below the allowed threshold, then that node is considered selfish and the first node will generate a warning message about the selfish node. If another node receives the warning message, it broadcasts an alarm to its neighborhood. The integrity of the generated alarm message can be ensured in the following two ways:

First, due to the inconsistent reputation problem discussed before, different nodes may have different reputation values for the same node. For instance, a node which is considered selfish by another node, may be considered non selfish by a third node. We assume that if *m* nodes in the neighborhood agree on the selfishness of a node, then it is considered to be selfish with high probability.

Second, by requiring *m* nodes to send the same WARNING message, we prevent nodes from false accusation (blame).

In LARS, two selfish node reaction behaviors can be considered. First, the routes of selfish nodes will be deleted and other routes bypassing the selfish nodes will be defined. Second, the punishment of selfish nodes is done by dropping their traffic.

5. Evaluation of Reputation-Based Incentive Schemes

In this section, we evaluate reviewed methods from various criteria, namely incentive patterns, security issues, implementation and performance, and routing and dissemination.

Table 1 shows the results of evaluation from incentive patterns point of view. In this table the following parameters are used for comparison and evaluation.

- Incentive scheme: it shows the main concept for cooperation among nodes. The valid values for this parameter are trust-based incentive pattern and trade-based incentive pattern.
- Trust Pattern: if trust accrues from membership it is called “collective pattern”, but if it adapts dynamically, it is called “community pattern”. In trade-based scheme, trust patterns are either “barter trade pattern” for immediate service or “bond-based pattern” for deferred service.
- Remuneration type: the form of remuneration exchanged between consumer and provider has a specific type. Reputation and check are two types of remuneration.
- Type of Trust: the possible values are “static” and “dynamic” for this field. The trust will be dynamic if it is based on direct or indirect experience of the entity. It is static, if there is a statement of trust like certificate.
- Roles: if a provider can be a consumer at the same time, the role of the entity is “symmetric”, otherwise it would be “asymmetric”.

TABLE 1
Evaluation of methods based-on characteristics of incentive patterns

	RADAR	**SORI**	**(Refaei)**	**Zakhary**	**OCEAN**	**LARS**
Incentive Scheme	Trust	Trust	Trust	Trust	Trade	Trust
Trust Pattern	Community /Collective	Community	Community	Community	bond	Community
Remunera -tion Type	Reputation value	Reputation value	Reputation value	Reputation value	Chip count	Reputation value
Type of Trust	Dynamic	Dynamic	Dynamic	Dynamic	Dynamic	Dynamic
Roles	Symmetric	Symmetric	Symmetric	Asymmetric	Symmetric	Symmetric

Table 2 illustrates the evaluation of reviewed methods based on security issues. The following parameters are used to compare the methods.

- Security: it indicates that whether a specific method supports any kind of security issues. N/A is used when there is no security procedure in the method.
- Integrity: it indicates whether the integrity of messages is kept during transmission. If there is an integrity method, its name is mentioned, otherwise N/A will be used.
- Misbehavior: it identifies the type of misbehavior that the method can resist. Selfish and malicious are two possible values for this field.
- Punishment: it indicates whether there is a punishment procedure in the method. “Yes” or “No” may be used for this part.
- Type of Punishment: it explains the procedure used to punish a misbehaved entity. The value in this field shows a summary of the procedure.
- Attack: it shows the type of attacks that a method has procedure to resist for. There are different types of attacks, but some common attacks are DoS, routing loops, black holes and rushing attack.

TABLE 2
Evaluation of methods regarding to security issues

	RADAR	SORI	(Refaei)	Zakhary	OCEAN	LARS
Security	IDS	Authentication	No	Replication	Public key	Integrity
Integrity	N/A	Hash-Chain	N/A	N/A	N/A	Yes
Mis-behavior	Malicious	Selfish	Selfish	Malicious	Misleading/ selfish	Selfish
Punish-ment	Yes	Yes	Yes	Yes	Yes	Yes
Type of Punish-ment	Reroute traffic from Malicious node	Reputation Propagation with overall Evaluation	Discarding Packets of Selfish nodes	Reroute traffic	Discarding Packets of Selfish nodes	Degrading the Reputation Value of misbehaved nodes
Attack	DOS Routing loop	Not specified	N/A	Single/ Multiple Black/Gray hole	Rushing attack	Routing attack

Table 3 shows the comparison among methods regarding to implementation and performance. The following parameters are the main criteria for this evaluation.

- Reputation Calculation: it shows in which manner reputation is calculated. Reputation for node can be calculated by considering direct local nodes or indirect nodes. Indirect methods include direct calculation. The possible values for this field are “direct” or “indirect”.
- Reputation Measurement: it summarizes the way of measuring reputation by a node.
- Mobility: it shows whether nodes in the network are mobile or not
- Network Model: it indicates the type of network which the method has been implemented on.
- Storage: it explains the type of storage to store the value of incentive pattern.

TABLE 3
Evaluation of methods regarding to implementation and Performance

	RADAR	SORI	(Refaei)	Zakhary	OCEAN	LARS
Reputation Calculation	Various Range-global trust	Direct Neighbors	Indirect	Direct & Indirect	Direct	Direct/Indirect
Reputation Measure-ment	Based on attributes of MAC and Routing Protocol	Packet forwarding	Completion of Requested Service	Centrality of reporting nodes & indirect Rep.	Packet forwarding	Direct Observation
Mobility	Limited-No	Highly mobile (20m/s)	Isolation Inversely proportional with speed	Highly mobile (20m/s)	Highly mobile (20m/s)	Highly Mobile
Network Model	Wireless Mesh Network	Mobile Ad-hoc Network	Ad-hoc	Mobile Ad-hoc Network	Mobile Ad-hoc Network	Mobile Ad-hoc networks
Storage	Neighbor node list (global,local)	Neighbor node list	Packet traces table with hash	Neighbors' reputation records	Neighbor node list	Neighbors Reputation Records

Table 4 indicates a brief comparison based on routing algorithms and the way of message dissemination. Two following parameters are considered for this comparison.

- Dissemination: “broadcast” or “unicast” are two possible values for this field. It indicates the way of transmitting the packets. Since wireless networks use omin-directional medium, the transmission usually occurs in broadcast form. If the destination is specifically identified, the value of directional is used as the value of the field.
- Routing: it shows the routing protocol which has been used in the method. If the method does not depend on routing protocol, the value of “independent” is mentioned. The name of routing protocol follows “Independent” shows the protocol used in the experiment.

TABLE 4
Evaluation of methods based-on routing and dissemination

	RADAR	SORI	(Refaei)	Zakhari	OCEAN	LARS
Dissemin ation	Broadcast	Broadcast	Directional	Broadcast	Broadcast	Broadcast
Routing	* (DSR)	DSR	* (AOVD)	AODV	* (DSR)	* (DSR)

* Independent from routing algorithm

6. Conclusion

One of the most important ideas in Oppnets is the expansion of the network opportunistically with support of available nodes in the area. This is usually fulfilled by adding helper nodes to the network. This expansion expands the network mainly into two aspects, namely size and potential ability. Since those helpers are not part of the original network, they may not be completely cooperative and trustable; therefore the first problem is motivating of new nodes. Another problem is security of information. This is due to the fact that, even those nodes connected to the network may show malicious behavior. In other words, in addition to selfish behavior malicious behavior is a serious concern in Oppnets. Mobility is another factor that must be concerned. Since helper nodes are not identifiable at first step, they may be highly mobile; therefore those incentive patterns which do not consider mobility are not suitable. Moreover, independency from routing protocol increases autonomy of Oppnets network.

Among those methods reviewed in this survey, SORI shows better compatibility with Oppnets; although some new characteristics must be added.

SORI is a trust-based incentive scheme with community pattern which is suitable for Oppnets; because all nodes are not the same and trustable. It supports high mobility, deals with selfish nodes with punishment procedure but not malicious nodes. Malicious nodes must be considered. It keeps the integrity of messages using hash function, but resisting to other attacks launched by malicious nodes is not supported. This current version of SORI has been implemented on DSR routing algorithm, but the independency from routing protocol provides higher scalability for Oppnets. After adding new features to SORI, it would be more compatible to Oppnet than other methods. Further simulation experiments are needed to reveal the correctness of this claim.

7. References

[1] Lilien L., Gupta A., Kamal Z., Yang Z. "Opportunistic Resource Utilization Networks - A New Paradigm for Specialized Ad Hoc Networks." *Elsevier*, 2009.

[2] L. Lilien, Z.H. Kamal, A. Gupta, I. Woungang, and E. Bonilla Tamez. "Quality of Service in an Opportunistic Capability Utilization." In *Mobile Opportunistic Networks: Architectures, Protocols and Applications*, by M. Denko et al. (Eds.). Auerbach Publications, Taylor & Francis Group, 2010.

[3] P. Obreiter, J. Nimis. "A Taxonomy of Incentive Patterns - The Design Space of Incentives for Cooperation." *Lecture Notes in Computer Science* 2872/2005 (2005): 89-100.

[4] Obreiter, P., K¨onig-Ries, B., Klein, M. "Stimulating cooperative behavior of autonomous devices - an analysis of requirements and existing approaches." *Second International Workshop on Wireless Information Systems (WIS2003).* 2003.

[5] Z. Zhang, F. Na¨ıt-Abdesselam, P. Ho, X. Lin. "RADAR: a ReputAtion-based Scheme for Detecting Anomalous Nodes in WiReless Mesh Networks." *WCNC.* IEEE Press, 2008.

[6] Q. He, D. Wu., P. Khosla. "SORI: A Secure and Objective Reputation-based Incentive Scheme for Ad-hoc Networks." *WCNC.* IEEE Press, 2004.

[7] S. Bansal, M. Baker. "Observation-based Cooperation Enforcement in Ad hoc Networks." *CoRR*, 2003.

[8] M. T. Refaei, V. Srivastava, L. DaSilva, M. Eltoweissy. "A Reputation-based Mechanism for Isolating Selfish Nodes in Ad Hoc Networks." *the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'05).* IEEE Press, 2005.

[9] S. R. Zakhary, M. Radenkovic,. "Reputation-Based Security Protocol for MANETs in Highly Mobile Disconnection-Prone Environments."

[10]J. Hu, M. Burmester. "LARS- A Locally Aware Reputation System for Mobile Ad Hoc Networks." *Proceedings of the 44th annual Southeast regional conference.* ACM, 2006.

[11] Schlosser, M.T. ""The EigenTrust Algorithm for Reputation Management in P2P Networks." *ReCALL.* 2003.

Radiation Efficiency of Rectangular Slot Resonator with Multilayer Photonic

Humberto Dionisio de Andrade, Anderson Max Cirilo Silva and Humberto Cesar Chaves Fernandes
humbertodionisio@yahoo.com.br cmaxander@hotmail.com humbeccf@ct.ufrn.br
Electrical Engineering Department - Technological Sector- Federal University of Rio Grande do Norte,
59078-970- Natal, RN, Brazil

Abstract *- This work has as main objective to analyze the rectangular slot resonator with four layers, with photonic materials - PBG, to obtain the complex resonant frequency and radiation efficiency of this structure. The analysis developed in this work was performed using the TTL - Transverse Transmission Line method. Numerical-computational results are presented in graphical form in three dimensions for all analysis performed, and the resonance frequency depending on the length and width of the slot and the radiation efficiency as a function of resonance frequency and height of the layers of substrate structure under study.*

Keywords: TTL Method, PBG-Photonic Band Gap, Four layer Slot antenna, Efficiency.

1 Introduction

The rectangular slot line resonator with four layers, consist of one rectangular slot line resonator, where there are two layers under and two layers over the patch.

This structure is shown in Figure 1, with width “w” and length “l”. With the analysis through the TTL method, the general equations to the electromagnetic fields are obtained.

The complex resonant frequency is calculated using double spectral variables, being the same, used in the elaboration of the efficiency and bandwidth's parameters.

By the usage of the system of Cartesian coordinates and the dimensional nomenclatures as presented in Figure 1 (perspective view), the equations of the electromagnetic fields are obtained, being considered despicable the thickness of the slot line.

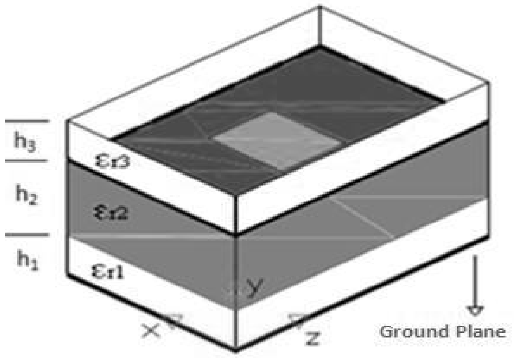


Figure 1. Perspective view of the four layers slot resonator.

2 PBG Structure

One of the problems when working with photonic material is the relative dielectric constant determination as the PBG is a non-homogeneous structure where the incident sign goes at the process of multiple spread.

A solution can be obtained through a numerical process known as homogenization. The process is based in the theory related to the diffraction of an incident electromagnetic plane wave, imposed by the presence of immerged cylinders of air in a homogeneous material.

For electromagnetic waves propagating in the xy plane these waves have the *s* polarization (E field parallel to the z axis) and *p* polarization (E field perpendicular to the z axis).

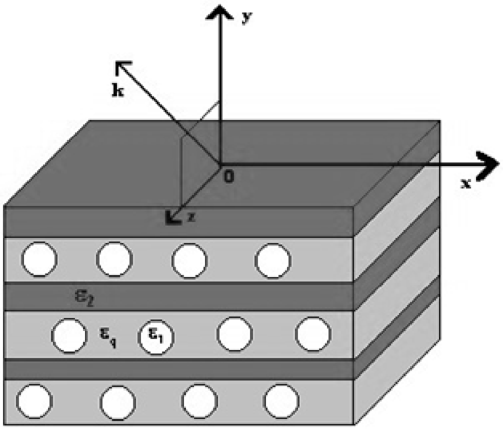


Figure 2. Homogenized PBG/EBG Crystal.

3 Fields Calculation

Due to the limitation in the length, the equations should be used for the analysis in the spectral domain in “x” and “z” directions as function.

Therefore the field equations are applied for double Fourier transformed defined as:

$$\tilde{f}(\alpha_n, y, \beta_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)\cdot e^{j\alpha_n x}\cdot e^{j\beta_k z}\,dx\,dz \qquad (1)$$

Where α_n is the spectral variable in the “x” direction and β spectral variable in the “z” direction.

After using the Maxwell’s equations in the spectral domain, the general equations of the electric and magnetic fields to the method TTL are obtained:

$$\tilde{E}_{xi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\alpha_n \frac{\partial}{\partial y}\tilde{E}_{yi} + \omega\mu\beta_k \tilde{H}_{yi}\right] \qquad (2)$$

$$\tilde{E}_{zi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\beta_k \frac{\partial}{\partial y}\tilde{E}_{yi} - \omega\mu\alpha_n \tilde{H}_{yi}\right] \qquad (3)$$

$$\tilde{H}_{xi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\alpha_n \frac{\partial}{\partial y}\tilde{H}_{yi} - \omega\varepsilon\beta_k \tilde{E}_{yi}\right] \qquad (4)$$

$$\tilde{H}_{zi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\beta_k \frac{\partial}{\partial y}\tilde{H}_{yi} + \omega\varepsilon\alpha_n \tilde{E}_{yi}\right] \qquad (5)$$

Where:

i = 1, 2, 3 and 4 are the dielectric regions of structure;

$$\gamma_i^2 = \alpha_n^2 + \beta_k^2 - k_i^2 \qquad (6)$$

Is the constant of the propagation in y direction; αn is the spectral variable in “x” direction and βk the spectral variable in “z” direction..

$k_i^2 = \omega^2\mu\varepsilon = k_0^2\varepsilon_{ri}^*$ Is the number of wave of ith term of Dielectric region;

$\varepsilon_{ri}^* = \varepsilon_{ri} - j\dfrac{\sigma_i}{\omega\varepsilon_0}$ Is the dielectric constant relative of the material with losses;

$\omega = \omega_r + j\omega_i$ is the complex angular frequency;
$\varepsilon_i = \varepsilon_{ri}^* \cdot \varepsilon_0$ is the dielectric constant.

The equations above are applied to the resonator being calculated, the fields Ey and Hy through the solution of the Helmoltz’s wave equations in the spectral domain [2]-[4]:

$$\left(\frac{\partial^2}{\partial y^2} - \gamma^2\right)\tilde{E}_y = 0 \qquad (7)$$

$$\left(\frac{\partial^2}{\partial y^2} - \gamma^2\right)\tilde{H}_y = 0 \qquad (8)$$

The solutions of Helmoltz’s equations for the four regions of the structure are given as examples:

Region 2:

$$\tilde{E}_{y2} = A_{2e}\cdot \operatorname{senh}\gamma_2 y + B_{2e}\cdot\cosh\gamma_2 y \qquad (9)$$

$$\tilde{H}_{y2} = A_{2h}\cdot \sinh\gamma_2 y + B_{2h}\cdot\cosh\gamma_2 y \qquad (10)$$

Region 4:

$$\tilde{E}_{y3} = A_{3e}\cdot e^{-\gamma_3 y} \qquad (11)$$

$$\tilde{H}_{y3} = A_{3h}\cdot e^{-\gamma_3 y} \qquad (12)$$

Substituting these solutions in the equations of the fields (2) to (5), as function of the unknown constants A21, A22, B21 and B22 are obtained, for example, for the region 2:

$$\tilde{E}_{x2} = \frac{-j}{k_2^2+\gamma_2^2}\begin{bmatrix}(j\omega\mu_0\beta_k B_{21}+\alpha_n\gamma_2 A_{22})\cosh(\gamma_2 y)+\\ j\omega\mu_0\beta_k B_{22}+\alpha_n\gamma_2 A_{21})\sinh(\gamma_2 y)\end{bmatrix} \quad (13)$$

$$\tilde{H}_{x2} = \frac{-j}{k_2^2+\gamma_2^2}\begin{bmatrix}(j\omega\varepsilon_0\beta_k B_{21}+\alpha_n\gamma_2 B_{22})\cosh(\gamma_2 y)+\\ j\omega\varepsilon_0\beta_k A_{22}+\alpha_n\gamma_2 B_{21})\sinh(\gamma_2 y)\end{bmatrix} \quad (14)$$

For the determination of the unknown constants, it is applied the boundary conditions to the 1, 2 and 3 regions:

For the regions 1 e 2: y = h1

$$\tilde{E}_{x1} = \tilde{E}_{x2} \quad (15)$$

$$\tilde{E}_{z1} = \tilde{E}_{z2} \quad (16)$$

$$\tilde{H}_{x1} = \tilde{H}_{x2} \quad (17)$$

$$\tilde{H}_{z1} = \tilde{H}_{z2} \quad (18)$$

For the regions 2 e 3: y =d- ; (g=h1+h2)

$$\tilde{E}_{x2} = \tilde{E}_{x3} = \tilde{E}_{xg} \quad (19)$$

$$\tilde{E}_{z2} = \tilde{E}_{z3} = \tilde{E}_{zg} \quad (20)$$

After several calculations are obtained, for region two:

$$A_{21} = \frac{\varepsilon_1\cosh(\gamma_2 y)}{\varepsilon_2\gamma_1\sinh(\gamma_1 g_1)\cosh(\gamma_2 g_2)+\gamma_2\frac{\varepsilon_1}{\varepsilon_2}\cosh(\gamma_1 g_1)\sinh(\gamma_2 g_2)}\left[j(\alpha_n\tilde{E}_{xg}+\beta_k\tilde{E}_{zg})\right] \quad (21)$$

$$A_{22} = \frac{\gamma_1\sinh(\gamma_2 y)}{\gamma_2\gamma_1\sinh(\gamma_1 g_1)\cosh(\gamma_2 g_2)+\gamma_2\frac{\varepsilon_1}{\varepsilon_2}\cosh(\gamma_1 g_1)\sinh(\gamma_2 g_2)}\left[j(\alpha_n\tilde{E}_{xg}+\beta_k\tilde{E}_{zg})\right] \quad (22)$$

$$B_{21} = -\frac{\sinh(\gamma_1 g_1)}{\omega\mu_0\sinh(\gamma_1 g_1)\cosh(\gamma_2 g_2)+\frac{\gamma_1}{\gamma_2}\cosh(\gamma_1 g_1)\sinh(\gamma_2 g_2)}\left[-\beta_1\tilde{E}_{xg}+\alpha_n\tilde{E}_{zg}\right] \quad (23)$$

$$B_{22} = -\frac{\gamma_1}{\gamma_2}\frac{\cosh(\gamma_1 g_1)}{\omega\mu_0\sinh(\gamma_1 g_1)\cosh(\gamma_2 g_2)+\frac{\gamma_1}{\gamma_2}\cosh(\gamma_1 g_1)\sinh(\gamma_2 g_2)}\left[-\beta_1\tilde{E}_{xg}+\alpha_n\tilde{E}_{zg}\right] \quad (24)$$

The general equations of the electromagnetic fields are obtained as function of the tangential electric fields on the antenna resonator.

4 Admittance Matrix Calculation

The following equations (25) and (26) relate the current densities on the sheets and the magnetic fields on the interface y = h1+h2:

$$\tilde{H}_{x2} - \tilde{H}_{x3} = \tilde{J}_{zt} \quad (25)$$

$$\tilde{H}_{z2} - \tilde{H}_{z3} = -\tilde{J}_{xt} \quad (26)$$

It has being done the substitutions of the magnetic fields equations, so after some calculations are obtained,

$$Y_{xx}\tilde{E}_{xg} + Y_{xz}\tilde{E}_{zg} = \tilde{J}_{zg} \quad (27)$$

$$Y_{zx}\tilde{E}_{xg} + Y_{zz}\tilde{E}_{zg} = \tilde{J}_{xg} \quad (28)$$

These equations are represented in the matrix form:

$$\begin{bmatrix}Y_{xx} & Y_{xz}\\ Y_{zx} & Y_{zz}\end{bmatrix}\begin{bmatrix}\tilde{E}_{xg}\\ \tilde{E}_{zg}\end{bmatrix} = \begin{bmatrix}\tilde{J}_{zg}\\ \tilde{J}_{xg}\end{bmatrix} \quad (29)$$

The "Y" admittance terms are the Green's dyadic functions to the antenna and they are represented by:

$$Y_{xx} = -\frac{j}{\varpi\mu 0(\gamma 2^2+k2^2)}\left[-\beta k^2\gamma 2.E+k2^2\alpha n^2.F\right]+ \frac{j}{\varpi\mu 0\gamma 3}\left[\alpha n^2 k3^2 E-\beta k^2\gamma 3^2.D\right] \quad (30)$$

$$Y_{xz} = \frac{-j\alpha n\beta k}{\varpi\mu 0\left(\gamma 2^{2}+k2^{2}\right)}\left[A+k2^{2}.(B)\right] - \frac{\alpha n\beta k}{\varpi\mu 0\gamma 3\left(k3^{2}+\gamma 3^{2}\right)}\left[k3^{2}.C+\gamma 3^{2}.D\right] \tag{31}$$

$$Y_{zx} = \frac{-j\alpha n\beta k}{\varpi\mu 0\left(\gamma 2^{2}+k2^{2}\right)}\left[A+k2^{2}.(B)\right] - \frac{\alpha n\beta k}{\varpi\mu 0\gamma 3\left(k3^{2}+\gamma 3^{2}\right)}\left[k3^{2}.C+\gamma 3^{2}.D\right] \tag{32}$$

$$Y_{zz} = \frac{j}{\varpi\mu 0\left(\gamma 2^{2}+k2^{2}\right)}\left[\alpha n^{2}.A-\beta k^{2}.k2^{2}.\mathbf{B}\right] - \frac{j}{\varpi\mu 0\gamma 3\left(k3^{2}+\gamma 3^{2}\right)}\left[\alpha n^{2}\gamma 3^{2}.C-\beta k^{2}\gamma 3^{2}.D\right] \tag{33}$$

Where,

$$A=\frac{\gamma 1\cdot\gamma 2}{\gamma 2tgh(\gamma 1\cdot h1)+\gamma 1tgh(\gamma 2h2)}+\frac{\gamma 2^{2}}{\frac{\gamma 2}{tgh(\gamma 2\cdot h2)}+\frac{\gamma 1}{tgh(\gamma 1\cdot h1)}} \tag{34}$$

$$B=\left(\frac{\varepsilon 1}{\gamma 1\cdot\varepsilon 2tgh(\gamma 1\cdot h1)+\gamma 2\cdot\varepsilon 1tgh(\gamma 2\cdot h2)}+\frac{\gamma 1\cdot\varepsilon 2}{\frac{\gamma 1.\gamma 2.\varepsilon 2}{tgh(\gamma 2.h2)}+\frac{\gamma 2^{2}.\varepsilon 1}{tgh(\gamma 1.h1)}}\right) \tag{35}$$

$$C=\left(\frac{\frac{\gamma 3\cdot\varepsilon 4}{\gamma 4\varepsilon 3}+tgh(\gamma 3h3)}{1+\frac{\gamma 3\varepsilon 4.tgh(\gamma 3h3)}{\gamma 4\varepsilon 3}}\right) \tag{36}$$

$$D=\left(\frac{\frac{\gamma 4}{\gamma 3}+tgh(\gamma 3h3)}{1+\frac{\gamma 4.tgh(\gamma 3h3)}{\gamma 3}}\right) \tag{37}$$

$$E=\left(\frac{\gamma 1+\gamma 2.tgh(\gamma 1h1).tgh(\gamma 2h2)}{\gamma 2tgh(\gamma 1h1)+\gamma 2.tgh(\gamma 2h2)}\right) \tag{38}$$

The tangential electric fields in the interface have being expanded using base functions [3], [5]:

$$\tilde{E}_{xg} = \sum_{i=1}^{n} a_{xi}\cdot\tilde{f}_{xi}(\alpha_n,\beta_k) \tag{39}$$

$$\tilde{E}_{zg} = \sum_{j=1}^{m} a_{zj}\cdot\tilde{f}_{zj}(\alpha_n,\beta_k) \tag{40}$$

Where a_{xi} and a_{zj} are constant unknown and the terms n and m are numbers integer and positive that can be done equal to 1, as seen in equations (41) and (42) following:

$$\tilde{E}_{xg} = a_x\cdot\tilde{f}_x(\alpha_n,\beta_k) \tag{41}$$

$$\tilde{E}_{zg} = a_z\cdot\tilde{f}_z(\alpha_n,\beta_k) \tag{42}$$

The Fourier transformed of the base functions chosen are [6]:

$$\tilde{f}_x(\alpha_n) = \pi\cdot J_o\left(\alpha_n\frac{w}{2}\right) \tag{43}$$

$$\tilde{f}_x(\beta_k) = \frac{2\pi l\cdot\cos\left(\frac{\beta_k l}{2}\right)}{\pi^2-(\beta_k l)^2} \tag{44}$$

$$\tilde{f}_x(\alpha_n,\beta_k) = \frac{2\pi^2 l\cdot\cos\left(\frac{\beta_k l}{2}\right)}{\pi^2-(\beta_k l)^2}\cdot J_o\left(\alpha_n\frac{w}{2}\right) \tag{45}$$

Where J_0 is the Bessel's function of first species and zero order. The Garlekin method is applied to (29), to eliminate current densities and the new equation in matrix form is obtained [5], [7].

$$\begin{bmatrix}K_{xx} & K_{xz}\\ K_{zx} & K_{zz}\end{bmatrix}\cdot\begin{bmatrix}a_x\\ a_z\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix} \tag{46}$$

Where,

$$K_{xx} = \sum_{-\infty}^{\infty}\tilde{f}_x\cdot Y_{xx}\cdot\tilde{f}_x^{*} \tag{47}$$

The solution to the characteristic equation of the determinant (14) it supplies the resonant frequency.

5 Radiation Efficiency

For rectangular slot resonator, the theoretical development to determine the radiation efficiency takes into account the losses of reflection, conduction and dielectric.

The radiation efficiency is defined as the ratio between the power delivered, to the radiation resistance and input power delivered to RL and Rr [8]:

$$\eta(\%) = \frac{R_t}{R_r}100 \tag{48}$$

The total resistance, Rt (Ω) is calculated from the following equation:

$$\frac{1}{R_t} = \frac{1}{R'_r} + \frac{1}{R_c} + \frac{1}{R_d} \tag{49}$$

The total resistance, Rt (Ω) is calculated from the following equation:

6 Results

The Figure 3 shows the 3D result of the resonant frequency as function of the slot length and width. The frequency increases when the width and the length decrease.

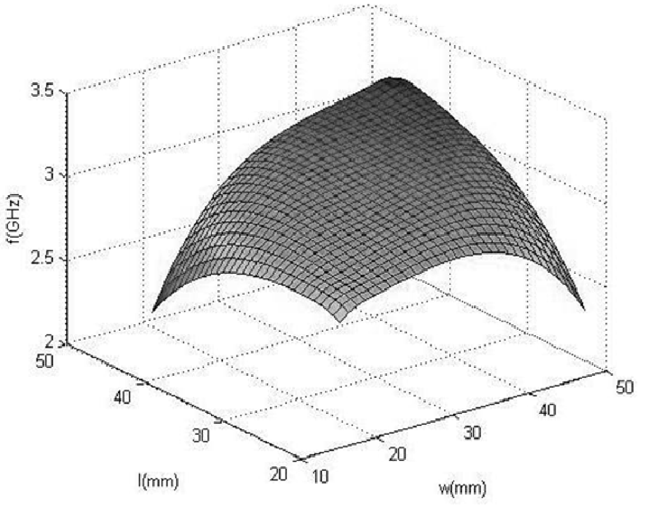


Figure 3. Frequency (GHz) as function of the width (mm) and length (mm).

Figure 4 shows a graphic comparing the curve of the radiation efficiency as a function of resonance frequency, for different thicknesses of substrate. For the results in Fig. 4 was considered permittivity as follows: εr1 = 12.0, εr2 = 8.702 (p polarization), εr3 = 12.0, εr4 = 1. The air is considered as the fourth layer. For the p polarization, was considered a height of substrate, h1 = 3.302 mm, h3=1.27 mm, h2 = 2.54 mm, h2 = 3.302 mm and h2 = 5.842 mm, respectively. The width 35.56 mm and the length is 71.12 mm. Therefore there is an improvement of radiation efficiency of the structure under study, and an increase in resonant frequency when the efficiency increases.

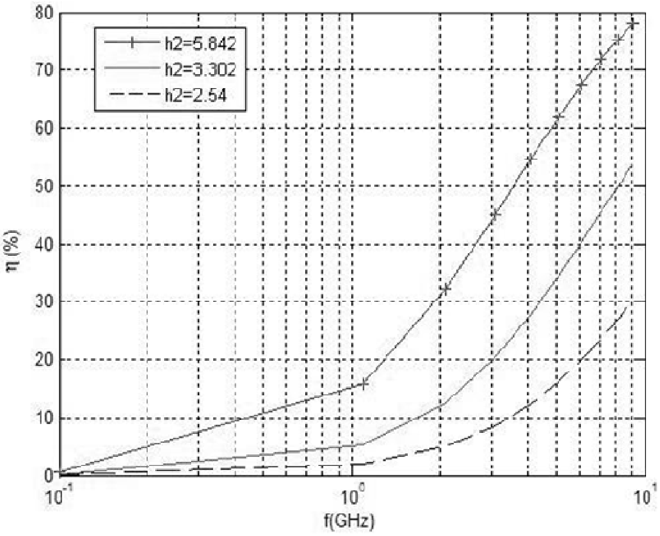


Figure 4. Radiation efficiency (η) as function of frequency for *p* polarization.

For the results in Figure 5 was considered permittivities for the following: εr1 = 12, εr2 = 10.233 (s polarization), εr3 = 12, εr4 = 1 (air). The thicknesses employed for this analysis are: h1 = 3.302 mm, h3=1.27 mm, h2 = 2.540 mm, h2 = 3.302 mm and h2 = 5.842 mm, respectively. The width is 35.56 mm and the length is 71.12 mm.

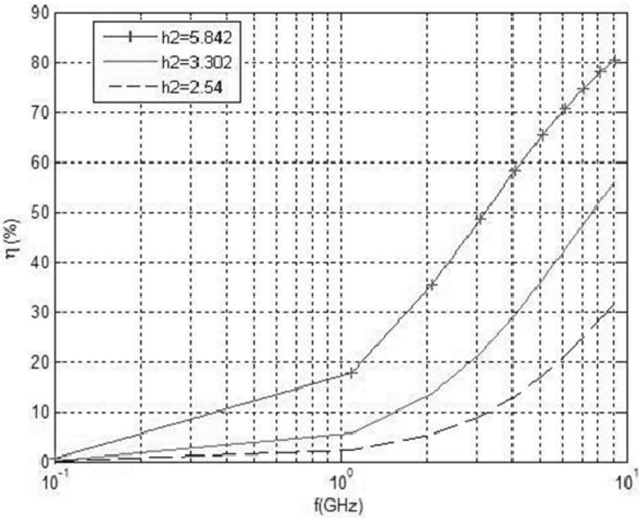


Figure 5. Radiation efficiency (η) as function of frequency for *s* polarization.

7 Conclusions

The Transverse Transmission Line – TTL method was used, in the analysis to obtain the numeric results of the four layers slot line resonator. According to the concise and effective procedures the calculus of the complex resonant frequency was obtained with accuracy. The rectangular slot resonator with four dielectric layers has its results from a complex resonant frequency. The development of the calculation of radiation efficiency as presented here is consistent with current literature.

ACKNOWLEDGMENT

This work was partially supported by CNPQ, CAPES, and INCT-CSF.

References

[1] A. K. Agrawal and B. Bhat, Resonant Characteristics and End Effects of a Slot Resonator in Unilateral Fin Line, Proc. IEEE, Vol. 72, pp. 1416-1418, Oct. 1984.

[2] H. C. C. Fernandes, S. A. P Silva and José P. Silva, "Coupling Analisys at the Coupler and Unilateral Edge-Coupled Fin Line," Int. Conference on Millimeter And Submillimeter Waves and Applications II, SPIE's 1998 International Symposium on Optical Science, Engineering and Instrumentation, San Diego, CA, USA. Conf. Proc. pp. 53-54, July of 1998.

[3] H. C. C. Fernandes and S. A. P. Silva, "Asymmetric and Unilateral Thick Edge-Coupled Fin Line and Coupling Analysis", in PIERS 1999 - The Progress in Electromagnetics Research Symposium, Taipei, Taiwan, Republic of China. Conf. Proc. 90, Mar 1999.

[4] H. C. C. Fernandes, S. A. P. Silva and O.S. D. Costa, "3D Complex Propagation of Coupled Unilateral and Antipodal Arbitrary Finlines", CBMAG'96- of Electromagnetism Brazilian Conference, Ouro Preto-MG, pp. 159-162, Nov. 1996.

[5] B. Baht and S. K. Koul, "Analysis, design and applications of finlines", Artech House, 1987.

[6] S. A. P. Silva and H.C.C. Fernandes, "Functions of basis in analysis of the coupled unilateral fin line coupler", IV SRET-Symposium of Research and Extension in Technology, Natal-RN, Annals pp. 79-81, Nov. 1998.

[7] H. C. C. Fernandes and M. B. Aquino Neto, "Four layers slot resonator", Proc. of the WSEAS2004, CSCC, CD-ROM, Athens, Greece, 5p. Jul. 2004.

[8] C. A. Balanis, "Antenna Theory: analysis and design" John Wiley & Sons, 1997.

[9] H. C. C. Fernandes, H. D. Andrade, M.B.L. Aquino and D.B. Brito, "Rectangular Slot Resonator with Four Dielectrics Layers". In: PIERS2007- Progress In Electromagnetics Research Symposium, PIERS2007, Beijing, China.. v. 1. p. 2121-2127, 2007.

[10] H. C. C. Fernandes, H.D. Andrade, "Photonic Rectangular Slot Resonator with Four Dielectrics Layers". In: ICAM – Int. Conf. on Advances Materials, Rio de Janeiro: SBPMAT, 2009. V. I. p. 301-301, 2009.

Novel Placement Mesh Router Approach for Wireless Mesh Network

Mohsen Rezaei[1], Mehdi Agha Sarram[2],Vali Derhami[3] ,and Hossein Mahboob Sarvestani[4]
Electrical and Computer Engineering Department, Yazd University, Yazd, Iran
[1]mohsen_rezaei@ stu.yazduni.ac.ir
[2]mehdi.sarram@yazduni.ac.ir
[3]vderhami@yazduni.ac.ir
[4]mahboob@stu.yazduni.ac.ir

Abstract - *Wireless Mesh Networks due their cost-efficient and fast deployment have had important networking infrastructure. Major problem in Wireless Mesh Networks, is Mesh router placement. An optimal Mesh router placement can ensure desire network performance in terms of network connectivity and coverage area. Since the problem is NP hard, to solve mesh router placement problem and achieve optimal solution with suitable quality, we can use heuristic approach. In this paper, we proposed a Genetic algorithm beside approach that has derived from circle packing problem. Circle packing problem consists in packing n non-identical circles without overlap inside the smallest containing circle C. Our model use of two objectives, maximization both network connectivity and coverage area. We have experimentally evaluated the proposed approach on instance network. The experimental results showed the efficiency of our approach for achieving high quality and optimal solutions of mesh router nodes placement in WMN.*

Keywords: Mesh router placement, Genetic algorithm, circle packing problem, network connectivity and coverage area.

1 Introduction

A wireless mesh network (WMN) is a communications network made up of radio nodes planned in a mesh topology. In Wireless mesh networks (WMNs) there are two types of nodes: mesh routers and mesh clients. A set of mesh routers (MRs), connecting to each other wirelessly and forming a backbone to serve set of mesh clients[1]. A few MRs with the Internet connections act as Internet Gateways (IGWs) to pass on the traffic between the Internet and the WMN. Low cost design nature and fast deployment of WMNs is that make them a cost-effective option to providing wireless Internet connectivity for mobile users at anytime and anywhere. These characteristics especially would be useful in developing regions or countries, avoiding costs of deployment and maintenance of wired Internet infrastructures.

The good performance and operability of WMNs largely depends on placement of mesh routers nodes in the geographical area to achieve network connectivity, stability and user coverage. The objective is to find an optimal and strong topology of the mesh network to support requirement services to clients. However, in a real deployment of WMN the automatic or purely random node placements produce poor performance WMN since the resulting placement could be far from optimal. Further, real deployment of WMNs may require taking into account specific restrictions and characteristics of real geographic area and therefore one need to explore different topologies for placing mesh routers. In fact, node placement can be seen as a crucial design and management issue in WMNs.

A practical MR placement scheme should determine the positions of MRs to satisfy the following three basic requirements: (1) Maximizing network coverage: MRs should be placed as far away from each other as possible to cover more areas. However, a MR placement only considering coverage requirement could lead to insufficient network connectivity. (2) Maintaining network connectivity: Each MR can communicate with one IGW through single or multi-hop wireless links. Should there are at least one path between an each two MRs. Therefore graph of network should be connective graph. (3) Adapting to network environment: MR placement is constrained by network environment. One important environmental limitation is geographical constraint and another is traffic distribution. The nature of the environment dictates where MRs could be placed. Traffic distribution influences the number of MRs. This scheme will become more important if do two following issue: (1) Determining optimal numbers of mesh routers to maximize coverage area and connectivity. (2) Determining optimal placement of mesh routers for maximization two above objectives.

Compared to the previous works, our approach makes the following improvement:

Firstly, do 2 issue above, that’s, determine number of mesh router needed to coverage environment of network. Do this with circle packing algorithm, that numbers of mesh router used to coverage network too many lower of other researches.

Secondly, due to low number of MRs used in network, cost of development and deployment in WMNs reduce proportionally.

The rest of the paper is organized as follows: In Section 2, we briefly discuss the existing work in the literature. In Section 3, the network model and problem formulation is described. Circle packing problem describe in section 4. In Section 5, we give a genetic algorithm based on circle packing problem for our problem. In Section 6, we evaluate our approach and finally, we conclude the paper in Section 7.

2 Related work

Unfortunately, node placement problems are shown to be computationally hard to solve to optimality [2-4], and therefore heuristic and meta-heuristic approaches are the *de facto* approach to solve the problem for practical purposes. Several heuristic approaches are found in the literature for node placement problems in WMNs [5-8]. In these papers, used of simulated annealing algorithm, Genetic algorithm, local search algorithm and neighborhood search methods respectively for addressing router nodes placements in wireless mesh network.

In [9], author *Partitioned the deployment area into grids*, *that* used for counting the number of users and measuring the signal strength received from each mesh node. For measurement and estimate the received signal strength at each position used of the channel pathloss model. The pathloss describes the attenuation experienced by the wireless signal as a function of distance. A local searching algorithm used to find the best locations through candidate locations and its method based on probability model.

In [10] the problem is addressed under a constraint network model in which the traffic demand is non-uniformly distributed and the candidate positions for MRs are pre-decided. Authors proposed a heuristic algorithm to obtain a close-to-optimal solution to reduce complexity of determining the locations of MRs while satisfying the traffic constraint.

A 2-stages multi objective evolutionary optimization algorithm which tries to optimize the two objectives by means of genetic algorithms, where individuals or solutions are represented by network graphs proposed in[11].In the first stage of MOGAMESH, candidate network topologies are found by letting a population of graphs evolve. In the second stage, a link elimination algorithm further reduces the number of links of the network.

A virtual force based MR placement algorithm (VFPlace) presented in [12] which Given a certain number of mesh routers, VFPlace targets to determine the positions of these mesh routers to maximize their overall coverage and maintain a certain number of neighbors for each mesh router, while satisfying geographic and traffic constraints of a specific WMN. VFPlace tried to dynamically avoid placing mesh routers in prohibitive regions, favor preferential regions and balance the distance between mesh routers**.**

3 Network Model and Problem Formulation

3.1 Network Model

The MR placement problem can be described as a way to determine appropriate positions for a number of MRs in a network area while satisfying environmental and traffic constraints. The area to be covered by a WMN backbone is modeled as a two-dimensional disk with a radius *R* in a two dimension coordinate plane. The center of disk is located at (0, 0), the origin of the coordinate system. At first, define a node set of MRs name*d* V_N. $V_N=\{v_0,v_1,\dots,v_{n-1}\}$ that each node represent a mesh router. We show each mesh router with a circle. We consider transmission range of each mesh router as radius of related circle. Then we denote geographical constraints area set with $V_C=\{v_n,v_{n+1},\dots,v_{n+c}\}$ that each node represents a circle area, which MRs can't be placed inside in. Given any node$v_i \in V_N$, its position is represented with the coordinate(x_i, y_i). For a MR node(x_i, y_i) is the position where the MR is situated. In other words, *(x_i, y_i)* represent a center of v_i circle. Since we consider one IGW in this paper, is located in (0, 0).

Two mesh router nodes are connected if the Euclidean distance between them is no longer than the MR's transmission range. In other words, two mesh router (circle) connected if have overlap with each other. The coverage area of a node C (v_i), *prime of circle v_i.*

It should be also noted that routers are assumed to have different radio coverage and higher radio coverage router assume to have more capacity and powerful addresses.

3.2 Problem Formulation

Determination placement of MR nodes has to minimize the required number of MR nodes needed to meet the maximize coverage as possible, full connectivity, and environmental constraints for network. The minimum number of MRs has great effect to maximally reduce the investment cost imposed by MR hardware. Based on the above network, we formulate the MR placement problem is to find the optimal placement by following inequalities:

$$\bigcup_{i=0}^{v_i} C(v_i) \approx \pi R^2 \text{ , } v_i \in V_N \qquad (1)$$

$$|R_{ij}| > 0 \ \forall v_i \text{ , } v_j \in V_N \qquad (2)$$

$$(x_i,y_i) \notin (x_c,y_c), \ \forall \ i \in V_N, \forall \ c \in V_C \qquad (3)$$

Inequality (1) says the set of selected MR nodes provide maximize coverage as possible of the network domain and inequality (2) models the full connectivity of all nodes. Full connectivity implies that every node to be connected to the IGW by at least one path and exist one path at least between any two nodes in network where $|R_{ij}|$denote distance a path between node v_i , v_j. Inequality (3) to satisfy geographical constraints says the position of mesh nodes shouldn't inside setV_C.

4 Circle Packing Problem

Cutting and Packing (C&P) problems are scientifically challenging problems with a wide spectrum of applications [13–17]. They are very interesting NP-hard combinatorial optimization problems; i.e., no procedure is able to exactly solve C&P problems in polynomial time. They generally consist of packing a set of items of known dimensions into one or more large objects or containers as to minimize the unused part of the objects or waste. The items and objects can be rectangular, circular, or irregular. In this paper, we use of the problem of packing a set of circular items into the smallest circle. The circular packing problem (CPP) consists of packing a set $N=\{1,2,\ldots,n-1\}$ of non-identical circles C_i without overlap into the smallest containing circle C where each $C_i$is characterized by its radius r_i The goal is to search for the best packing of the n circles into C, where the best packing minimizes waste. Instance of circle packing problem showed in Fig. 1.

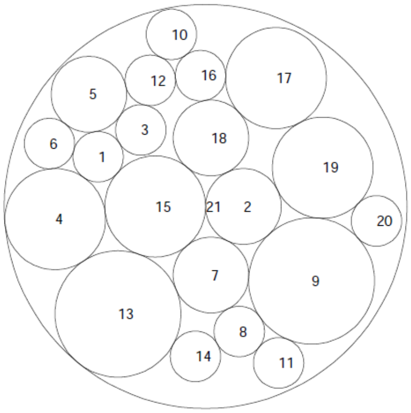


Fig. 1- Instance of Circle Packing Problem

CPP is equivalent to finding the coordinates (x_i, y_i) of every circle C_i , i ∈N and the radius r and coordinates (x,y) of C, such that no pair $(C_i,C_j)\in N\times N$, and $i\neq j$ overlap. Formally, the problem can be stated as finding the optimal level of the decision variables r, (x, y), and (x_i , y_i), i ∈N, that

$$\text{(CPP)} \left\{ \begin{array}{l} \qquad\qquad \text{Minimize } r \\ \text{Subject to } \sqrt{(x_i-x)^2+(y_i-y)^2} \ \le \ r-r_I,\ i\in N \text{ , } \quad (4) \\ \sqrt{(x_i-x)^2+(y_i-y)^2} \ \ge r_i+r_j,\ \ i\in N\times N,\ j<i \end{array} \right.$$

CPP has a linear objective function but non-linear non-differentiable constraints. The first set of constraints states that any C_i , i ∈Nis totally contained within C. There are n of these constraints, one for each C_i. The second set reinforces the no overlap constraint of any pair of distinct circles (C_i,C_j); that is, the Euclidean distance between the centers of C_i and C_j must be greater than or equal to (r_i+r_j).

In circle packing problem as mentioned above, one of conditions, is *circles shouldn't have overlap* with each other. in order that we can use this algorithm for our problem, need to change this condition such that each circle (router) overlap with at least one other circle (router) to satisfy network connectivity condition. Thus we modify algorithm in determine position such that any circle overlap with two other circles.

The resulted pattern has the following key properties:

- The final network is formed by *n* circles. Each MR is placed in the center of each circle.
- Each circle is surrounded by at most six other circles to cover the disk without gaps.
- The distance between any two neighboring MR nodes i and j is smaller than the sum of transmission range's, r_i+r_j, so that the network connectivity is guaranteed.

Our approach based on circle packing pattern fills the network plane with limited overlaps and no gaps. Overlaps are necessary to satisfy the connectivity requirement, limited overlaps essential for minimizing the number of MRs in the disk; no gap is needed to satisfy the coverage requirement.

5 Our Genetic Algorithm Based on Circle Packing Problem

In our approach, we present modified version of circle packing algorithm and solve it with genetic algorithm so that use for placement mesh routers problem in WMNs.

5.1 Encoding

We use a permutation of (1,2,…,n) to code a individual for encoding. M individuals are randomly generated to form the initial population. We run GA with different value of n. for obtain optimal number of routers, we call genetic algorithm with different value for n, and calculate coverage area for any pattern obtained with these value of n. each chromosome denotes a router.

5.2 Fitness function

The fitness function is of particular importance in GAs as it guides the search towards most promising areas of the solution space. Furthermore, in our case, although we face an optimization problem with multiple criteria, including connectivity of MRs and coverage area and therefore the fitness function in our particular case can be expressed in different ways. Since positioning algorithm, placed routers in positions that have overlap at least with one other router, thus we can use ratio of total coverage area each individual to evaluate fitness of it. This means that proportion total coverage area to number of routers used to achieve this coverage. If two patterns have same total coverage area, we calculate the numbers of router in pattern and pattern with lower router has greater fitness.

5.3 Crossover operators

We apply a crossover operation that retains the validity of permutations [18]. For convenience, we assume K is an even number, so we have K/2 pairs of parents. For each pair of parents, we apply the crossover operation to generate two children. In this paper, we apply crossover as follows: $child_1$ and $child_2$ area pair of parent solutions. Suppose q is a random integer, where $1 \leq q \leq n$. The first child is generated by taking routers1…q from $child_1$ and appending to this subsequence any missing routers in the order in which they appear in$child_2$. The second child is obtained in the same way, only with first subsequence taken from $child_2$, and the remainder being made up from $child_1$.

5.4 Mutation operators

We randomly select two routers from its sequence and exchange their positions as mutation operator.

6 Simulation Result

In this section, we present the simulation results to show the effectiveness of our approach. The simulation is done using the Matlab. For compare our result with [12], used of same scenario network in simulation, that's: The radius of the network domain is 8 units (R = 8). The transmission range of each MR node is 2 units (r = 2). But transmission MRs can be different. The number of MR nodes to be placed determine by our approach. The IGW located at (0, 0), the center of the network disk and doesn't shows in Figures. As mentioned above and described in next section, we consider value R^2/r^2 for initial and during algorithm increase its value.

Fig. 2 shows the result of placement our algorithm in the case of geographical constraints doesn't consider. At compare with Fig. 4 that show random placement and use of 38 nodes, our result can get to coverage area close to 95% with *only 24 nodes* (router) rather than 38 nodes.

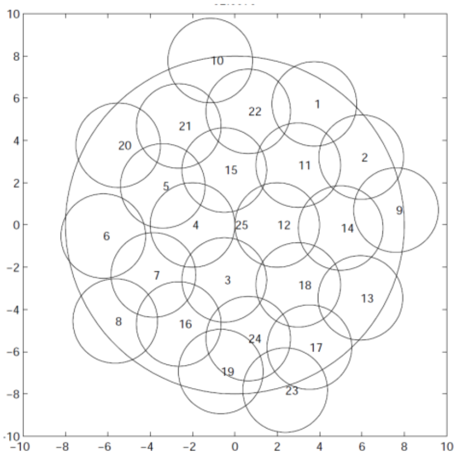


Fig. 2-MR placement without geographical constraints

Geographical constraints, places that cause to condition node can't be placed in these regions as shown in Fig. 3, considered and coverage area *90%* obtained again with 24 nodes. Rather than Fig. 5 that with 38 nodes coverage network, our approach with 24 nodes, this mean with *14 nosed less*, almost entire network coverage. This reduce in number of router needed to coverage network, Significantly decrease cost of network, consist of installing, deployment, maintenance. In others word, cost of hardware of mesh network *36% has reduced than [12]*.

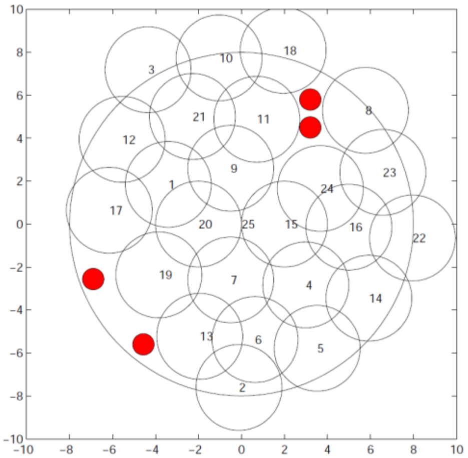


Fig. 3- MR placement with considering geographical constraints

In terms of network connectivity, since our algorithm use of method that each node at least has overlap with another node, full connectivity 100% obtain and between all of the nodes exist path. While in Fig. 4 random placement presented in [12], full connectivity doesn't exist.

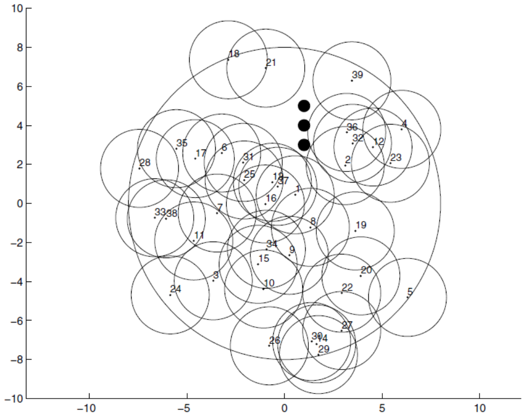


Fig. 4-Random Placement

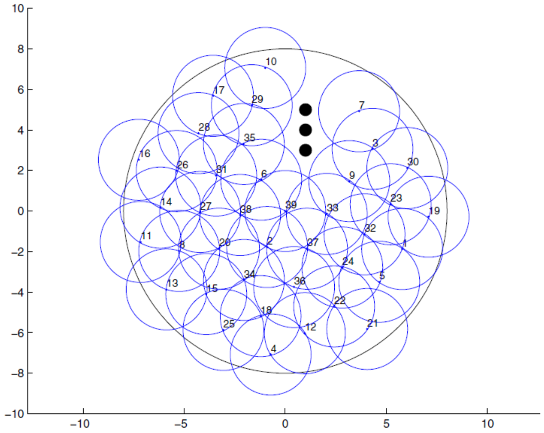


Fig. 5-MR placement by [12]

6.1 Finding Optimal Number of Mesh Routers

If we consider transmission range of mesh router (radius circle) identical, and assume any two mesh router don't have overlap with each other, at least R^2/r^2 MRs required to coverage network area (if position of circles determine with circle packing problem) but it is clear don't get full coverage in this case because the circles don't have overlap with each other and thus exist gap in network disk. We in our approach to find the optimal number of MRs to coverage network consider two cases: 1) all of mesh routers have same transmission range and 2) routers have different transmission range. In case 1, algorithm with R^2/r^2 MRs that have overlaps started and one by one add MR to network till total coverage of pattern close to 100%, because the area related to geographical constraint set prevent to get to full coverage in some case based on position of member set VC. If we use of routers with different coverage radio, problem different partly. In our algorithm as mention in genetic algorithm section, have two phases.

In first phase consider all if routers with identical and based on placement done. After placement, we calculate traffic loads that each router must passed and if traffic loads one router beyond its capacity, algorithm alters this router with powerful router that have more capacity and higher transmission range. In phase continue to replace weaker router with powerful router and thus the position of these routers to be constant. Since position of these routers determined, algorithm restarts for other nodes and adds MRs one by one to network domain till get total coverage close to 100% and re-determine their final position of these nodes. Thus number of MRs according to this approach can be obtained that consist of some routers with distinct transmission range.

In Fig. 6 shows the case of, router's number 20 (in Fig. 3), after calculation traffic load, exchange with powerful router, that's No. 18 router.

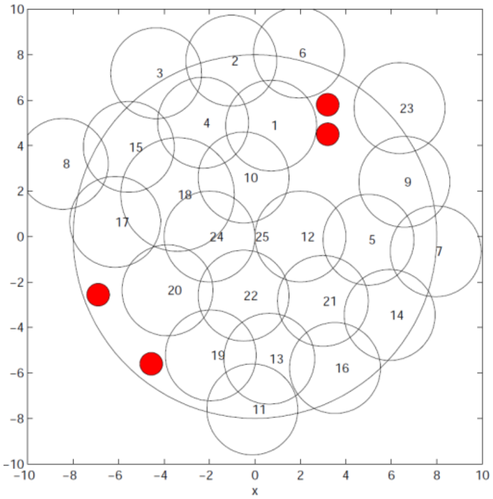


Fig. 6-Find number of routers after traffic load phase

7 Conclusion

In this work we have presented Genetic Algorithm based on circle packing problem for the problem of mesh router nodes placement in Wireless Mesh Networks (WMNs). We have considered two metric for evaluation our solution: network connectivity, and ration of network coverage area. Experimental results showed in network connectivity, obtained 100% connectivity, and in coverage area with many lower number mesh routers, coverage area close to 100% obtained. In fact, with reduce number of needed mesh router; cost of setup, investments decrease significantly. The proposed approach has practical usefulness for designing and deploying of real WMNs. In our future work we would like to improve our approach to get 100% coverage.

8 References

[1] I. F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. Computer Networks **47**(4) (2005) 445-487.

[2] E. Amaldi, A. Capone, M. Cesana, I. Filippini, F. Malucelli. Optimization models and methods for planning wireless mesh networks. Computer Networks **52** (2008) 2159-2171.

[3] M.R. Garey and D.S. Johnson.Computers and Intractability –A Guide to the Theory of NP-Completeness. Freeman, San Francisco, (1979).

[4] A. Lim, B. Rodrigues, F. Wang and Zh. Xua. k−Center problems with minimum coverage. Theoretical Computer Science **332** (2005) 1-17.

[5] F. Xhafa, A. Barolli, C. Sánchez, L. Barolli. A simulated annealing algorithm for router nodes placement problem in Wireless Mesh Networks. Simulation Modelling Practice and Theory, In Press, 2010.

[6] Xhafa, F., C. Sánchez, and L. Barolli, Genetic Algorithms for Efficient Placement of Router Nodes in Wireless Mesh Networks in Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on p. 465-472.

[7] Xhafa, F., C. Sánchez, and L. Barolli, Locals Search Algorithms for Efficient Router Nodes Placement in Wireless Mesh Networks, in Network-Based Information Systems, 2009. NBIS '09. International Conference on p. 572-579.

[8] Xhafa, F., C. Sánchez, and L. Barolli, Ad Hoc and Neighborhood Search Methods for Placement of Mesh Routers in Wireless Mesh Networks Distributed Computing Systems Workshops, 2009. ICDCS Workshops '09. 29th IEEE International Conference on p. 400-405.

[9] Franklin, A.A. and C.S.R. Murthy.Node Placement Algorithm for Deployment of Two-Tier Wireless Mesh Networks.in Proceedings of IEEE GLOBECOM 2007, IEEE Global Communications Conference.

[10] Wang, J., et al., Efficient Mesh Router Placement in Wireless Mesh Networks, in Mobile Adhoc and Sensor Systems, 2007. IEEE Internatonal Conference on p. 1-9.

[11] De Marco, G, MOGAMESH: A multi-objective algorithm for node placement in wireless mesh networks based on genetic algorithms, in Wireless Communication Systems, 2009. ISWCS 2009. 6th International Symposium on p. 388 – 392.

[12] Wang, J., W. Fu, and D.P. Agrawal, An Adaptive Router Placement Scheme for Wireless Mesh Networks in GLOBECOM Workshops, 2008 IEEE on p. 1-5.

[13] E. Bischoff, G. Wa¨scher, Cutting and packing, European Journal of Operational Research 84 (1995) 503–505.

[14] K.A. Dowsland, Palletisation of cylinders in cases, OR Spektrum 13 (1991) 171–172.

[15] H. Dyckhoff, G. Scheithauer, J. Terno, Cutting and packing (C&P), in: M. Dell'Amico, F. Maffioli, S. Martello (Eds.), Annotated Bibliography in Combinatorial Optimization, John Wiley and Sons, New York, 1997.

[16] A. Lodi, S. Martello, M. Monaci, Two dimensional packing problems: A survey, European Journal of Operational Research 141 (2002) 241–252.

[17] P.Y. Wang, G. Wa¨scher, Cutting and packing, European Journal of Operational Research 141 (2002) 239–240.

[18] Y.C. Xu, R. B. Xiao, and M. Amos .A novel genetic algorithm for the layout optimization problem. Proceedings ofthe 2007 IEEE Congress on Evolutionary Computation (CEC07), IEEE Press, on p. 3938-3942.

Rectangular Microstrip Antennas Linear Array with Superconductor at High Temperature

Humberto César Chaves Fernandes, Hugo Michel Camara Azevedo Maia and Leonardo Martins Caetano
Electrical Engeneering Department, Federal University of Rio Grande do Norte, 59078-970-Natal, Rio Grande do Norte, Brazil
humbeccf@ct.ufrn.br , humychel@yahoo.com.br , leoteleco@yahoo.com.br

Abstract - *The analysis and new results of the resonance frequency and pattern fields of microstrip antennas array, with superconductor patch, for different very high critical temperatures, are presented. The linear superconducting rectangular microstrip antennas array use the new materials $Sn_5InBa_4Ca_2Cu_{10}O_y$ at temperature of 212 K, conductivity of $1.88x10^5$ S/m and $Tl_5Ba_4Ca_2Cu_{10}O_y$ at temperature of 233 K with conductivity equal $2.0x10^5$ S/m. The concise full wave Transverse Transmission Line (TTL) method is used in the analysis. New results as functions of the temperature, and resonant frequency as functions of the various antenna parameters, for different superconductor are presented.*

Keywords: High Temperature; microstrip antenna; superconductor; linear antenna array.

1 Introduction

The superconductivity is a phase, a state of matter observed only in some solids, mostly metals.

Regarding the study of superconductivity, there is a phenomenon that was discovered in 1933 by researchers H. Meissner and R. Ochsenfeld, which reports the perfect diamagnetism, which means that lines of magnetic flux are completely expelled from the superconductor and there is a force that repels the superconducting away from magnetic fields. This phenomenon became known as the Meissner effect [1-3].

A theory widely used for superconductor is the BCS theory, developed by Bardeen, Cooper and Schrieffer. The macroscopic theory uses the two fluids model and the London equations. In this work are used microstrip antennas with recent very high superconductor patch. The structure is shown in Figure 1.

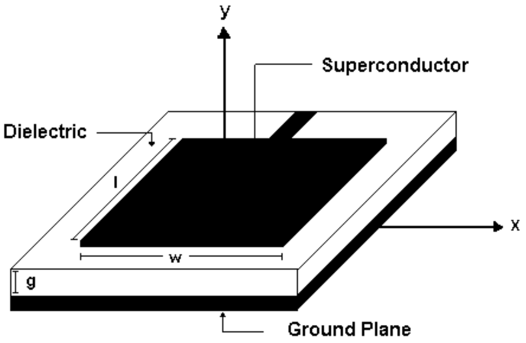


Figure 1. Superconducting microstrip antenna with patch of width, w, and length, l.

Considering the microstrip antenna resonator of Figure 1, the equations that represent the electromagnetic fields in the x and z directions as function of the electric and magnetic fields in the y direction are obtained, applying the TTL method [4]-[10].

2 Theory

Starting from the Maxwell's equations, after various algebraic manipulations, the general equations for the antenna in the FTD - Fourier Transformed Domain are obtained, for the x and z directions as [8]-[10]:

$$\tilde{E}_{xi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\alpha_n \frac{\partial}{\partial y}\tilde{E}_{yi} + \omega\mu\beta_k \tilde{H}_{yi}\right] \tag{1}$$

$$\tilde{H}_{xi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\alpha_n \frac{\partial}{\partial y}\tilde{H}_{yi} - \omega\varepsilon\beta_k \tilde{E}_{yi}\right] \tag{2}$$

$$\tilde{E}_{zi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\beta_k \frac{\partial}{\partial y}\tilde{E}_{yi} - \omega\mu\alpha_n \tilde{H}_{yi}\right] \tag{3}$$

This work was partially supported by CNPQ and INCT-CSF.

$$\tilde{H}_{zi} = \frac{1}{\gamma_i^2 + k_i^2}\left[-j\beta_k \frac{\partial}{\partial y}\tilde{H}_{yi} + \omega\varepsilon\alpha_n\tilde{E}_{yi}\right] \qquad (4)$$

Where:

i = 1, 2 are the dielectric regions of structure;

$\gamma_i^2 = \alpha_n^2 + \beta_k^2 - k_i^2$ Is the propagation constant in **y** direction;

α_n is spectral variable in **x** direction;

β_k is spectral variable in **z** direction;

$k_i^2 = \omega^2\mu\varepsilon = k_0^2\varepsilon_{ri}^*$ is the wave number of the dielectric region;

$\varepsilon_{ri}^* = \varepsilon_{ri} - j\frac{\sigma_i}{\omega\varepsilon_0}$, is the relative dielectric permittivity of the complex material.

$$\begin{bmatrix} Z_{xx} - Z_S & | & Z_{xz} \\ ---- & | & ---- \\ Z_{zx} & | & Z_{zz} - Z_S \end{bmatrix} \cdot \begin{bmatrix} \tilde{J}_x \\ -- \\ \tilde{J}_z \end{bmatrix} = \begin{bmatrix} \tilde{E}_x^{out} \\ -- \\ \tilde{E}_z^{out} \end{bmatrix} \qquad (5)$$

Where :

$$Z_S = \frac{1}{\sigma_S t} \qquad (6)$$

Zs is the superconductor impedance, σ_s is the conductivity and "*t*" is the thickness of the superconductor patch.

The Moment method is used to eliminate the electric fields in (5), and to obtain the homogeneous matrix equation (7) for the calculation of the complex resonant frequency.

$$\begin{bmatrix} K_{xx} & K_{xz} \\ K_{zx} & K_{zz} \end{bmatrix} \cdot \begin{bmatrix} a_x \\ a_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (7)$$

The determinant of this K matrix, witch is the inner product of the Z matrix and the basis functions, provides the real and imaginary resonant frequencies, which are numerically calculated using computational programs developed in this by the authors, in Fortran Power Station Language.

2.1 BCS Theory

The quantum theory of the superconductivity was launched in 1957 by the work of Bardeen, Cooper and Schieffer. This theory includes:

An attractive interaction between electrons can be conduced to a ground state, separated from an excited state by an energy gap, which separates the superconducting electrons below the gap of the normal electrons. The critical field, the thermal and electromagnetic properties are many other consequences of this energy gap.

The penetration depth,λ, appear as natural consequences of the BCS theory [3]. The London equation is obtained at the magnetic fields, which vary slowly in space. Thus the Meissner effect is obtained naturally.

The BCS theory predicts the critical temperature of an element or alloy. There is a paradox: the greater resistivity at room temperature is the likelihood that this metal is a superconductor when cooled.

The London equations are used [7],

(3.1)

$$e\vec{E} = m\frac{d\vec{v}}{dt} \qquad (8)$$

$$\Lambda\frac{\partial\vec{j}}{\partial t} = \vec{E} \qquad (9)$$

Where, Λ is London constant, m is the electrons number, e is the electron charge and v is the Fermi velocity.

3 Antenna Arrays

The antenna array [4] consists of a finite number of identical irradiants elements, which combines the induced signals in these antennas, to form the array. The maximum beam direction is controlled, adjusting the phase of the sign in elements of different spaces. The phase induced in the several adjustments in the elements, so that the sign obtain maximum directivity and gain.

The antenna array can be classified as linear and planar. Figure 2 shows the linear array of microstrip antenna with four elements. The linear array factor is,

$$FA_n = \left[\frac{\operatorname{sen}\left(\frac{N}{2}\psi\right)}{\left(\frac{N}{2}\right)}\right] \qquad (10)$$

Where:

$$\psi = kd\cos\theta + \beta \tag{11}$$

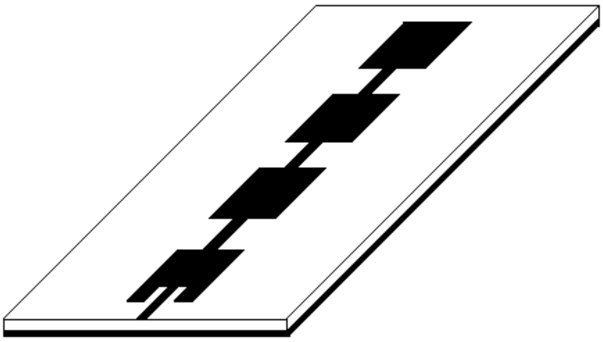


Figure 2. Linear antenna array with four superconducting patch.

4 Results

In this work the Transverse Transmission Line (TTL) method was applied, using double Fourier Transform [10].

Computational algorithms were developed in Matlab and Fortran PowerStation languages.

Figure 3 shows the frequency resonance as function of the patch length for different critical temperatures. The parameters are w = 25 mm, l = 30 mm, ε_{r1} = 10.233 (RT DUROID 6010), ε_{r2} = 1; and the $(Sn_5In)\ Ba_4Ca_2Cu_{10}Oy$ [8] at the superconducting temperature of 212 K.

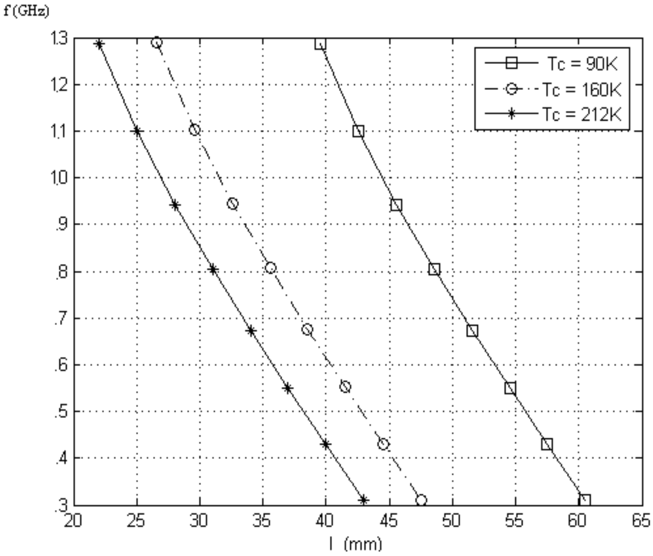


Figure 3. Resonance frequency in GHz as functions of the patch length, at critical temperature of 90 K, 160 K and 212 K.

Figure 4 shows the frequency resonant in function to the patch length for different critical temperature. The parameters are w = 25 mm, l = 30 mm, ε_{r1} = 10.233, ε_{r2} = 1; and the $(Tl_4Ba)Ba_2Ca_2Cu_7O_{13+}$ [8] at the superconducting temperature of 254 K.

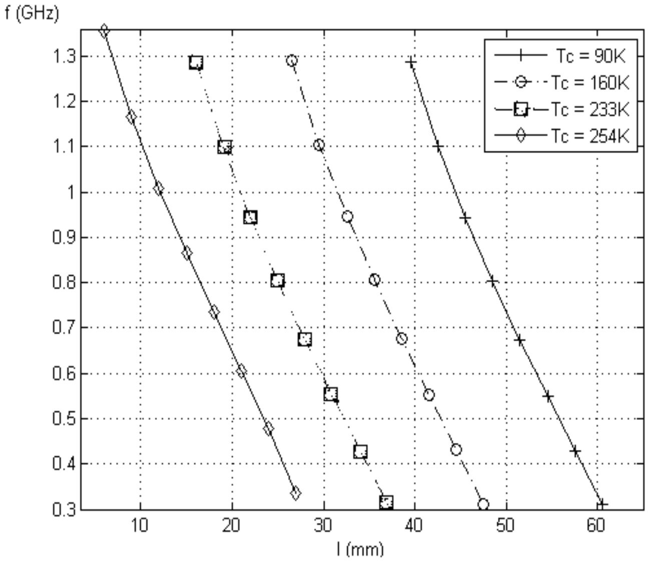
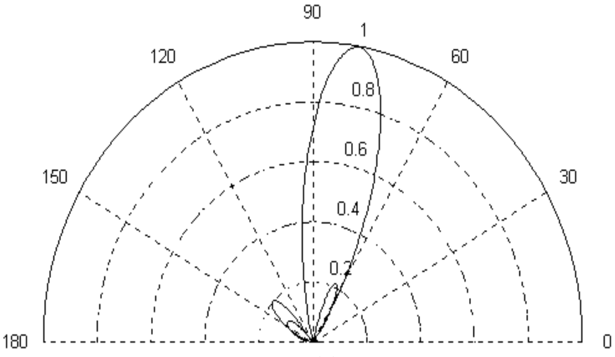


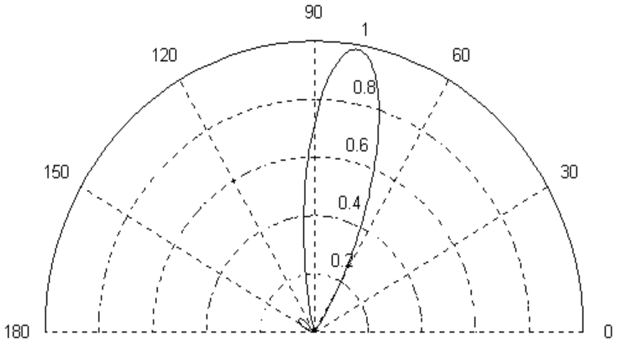
Figure 4. Resonant frequency in GHz as functions of the patch length, at critic temperature of 90 K, 160 K, 233 K and 254 K.

The results in the Figure 5.a and 5.b shows the pattern fields in E-Plane and H-Plane, for a linear array with 4 elements spaced λ/2 for an angle of irradiation of 80° , resulting in a phase of β = 31,25°.

Finally the results in Figure 6.a and 6.b shows the pattern fields in E-Plane and H-Plane for a linear array with 4 elements spaced λ/2 for an angle of irradiation of the 90° .

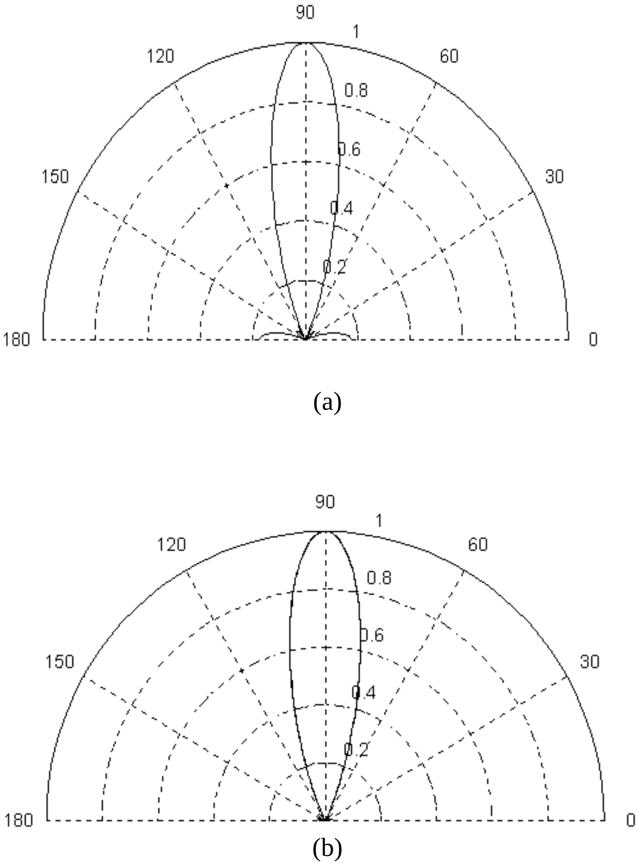


(a)



(b)

Figure 5. Irradiation Diagram for a linear array with superconductor patch at Tc = 254 K and $\theta = 80^o$, for plane-E (a) and plane-H (b).



(a)

(b)

Figure 6. Pattern field for a linear array with superconductor patch at Tc = 254 K and $\theta = 90^o$, for plane-E (a) and plane-H (b).

5 Conclusions

The superconductor theories have been presented at new superconductor materials. Numerical results of the resonance frequency, as functions of the linear antenna array including pattern fields of the E-plane and H-plane, were presented. The results obtained from having a good metric conformity and this is seen when Tc increases, the dimension of the antenna reduces. New very high critical temperature material presented in the literature was used in this new application.

6 References

[1] A.C. Rose-Innes and E.H. Roderik, "*Introduction to Superconductivity*", 2^a Edition, Pergamon Press, 1978.

[2] E. A. Linton, "*Superconductivity*", London: Mathuen & Co. LTDA, Ney York: John Wiley & Sons Inc. 1964..

[3] E. B. Eckholm and S. W. Mcknight, "Attenuation and Dispersion for High-Tc Superconducting Microstrip Lines", IEEE-MTT, Vol. 38, pp. 387-395, 1990.

[4] D. Nghiem, J. T. Williams and D. R. Jackson, "A General Analysis of Propagation Along Multiple-layer Superconducting Stripline and Microstrip Transmission Lines", IEEE-MTT, Vol. 39, No 9, pp. 1553-1565, Set.1991.

[5] J. M. Pond, C. M. Krowne and W. L. Carter, "On Application of Complex Resistive Boundary Conditions to Model Transmission Lines Consisting of Very Thin Superconductors", IEEE-MTT, Vol. 37, No 1, pp. 181-189, Jan. 1989.

[6] Z. Cai and J. Bornemann, "Generalized Spectral-Domain Analysis for Multilayered Complex Media and High-Tc Superconductor Application", IEEE-MTT, Vol. 40, No 12, pp. 2251-2257, Dez. 1992.

[7] Zhi-Yuan Shen, "High-Temperature Superconducting Microwave Circuits", Artech House, Inc. 1994

[8] H.C.C. Fernandes and L. P. Rodrigues, "Double Application of Superconductor and Photonic Material on Antenna Array", Trans. on Communications, Athens, Greece, pp. 425-432, Jul. 2004.

[9] H.C.C. Fernandes, J.R.S. Oliveira and A.J. Giarola, "Dispersion in Finlines on Semiconductor and the S-Parameters of a Step Discontinuity", International Journal of Infrared and Millimeter Waves, Vol. 12, N^o 5, pp. 505-519, Plenum Press, USA, May, 1991.

[10] H.C.C. Fernandes and H. M. C. A. Maia, "New Antenna with Superconductor at Critical Temperature of 212". IMOC2009-SBMO/IEEE MTT-S Inter. Microwave and Optoelectronics Conference, Belém – PA, Brazil, paper 59111. pp. 197-199, Nov. 2009.

Improving Network Lifetime with Fuzzy Sink Mobility Scheme in Wireless Sensor Network

Hossein Mahboob Sarvestani[1], Mehdi Agha Sarram[2],Vali Derhami[3] ,and Mohsen Rezaei[4]
Electrical and Computer Engineering Department, Yazd University, Yazd, Iran
[1] mahboob@stu.yazduni.ac.ir
[2]mehdi.sarram@yazduni.ac.ir
[3]vderhami@yazduni.ac.ir
[4] mohsen_rezaei@stu.yazduni.ac.ir

Abstract - *As wireless sensor networks (WSNs) have increasing and high capabilities in broad range of application, recently have earned remarkable attention. Generally, sensor nodes use battery as power source, to this end the important issue for concern is how to reduce the energy consumption of nodes, so that the network lifetime can be prolonged to reasonable times. A major reason affecting the lifetime of WSNs is unbalanced energy consumption in different parts of the network. This unbalanced energy consumption is a direct result of having a static sink. Nodes near the sink frequently participate in forwarding data of other nodes to the sink. Therefore they are more subject to premature energy depletion. This condition called funneling effect [1]. The main contribution of this article is a mobile sink approach based on fuzzy logic and expert system that can analyze network environment, and make best decision for extending the network lifetime automatically.*

Keywords: Wireless Sensor Networks, Mobile sink, Network lifetime, Funneling effect

1 Introduction

A wireless sensor network consists of sensor nodes deployed over a geographical area for monitoring physical phenomena in wide range of commercial, scientific, health, surveillance, and military applications. Typically, a sensor node is a tiny device that includes three basic components: a sensing unit for data acquisition from the physical surrounding environment, a processing unit for local data processing and storage, and a wireless communication unit for data transmission. In addition, a power source supplies the energy needed by the device to perform the programmed task. This power source often consists of a battery with a limited initial energy. On the other hand, the sensor network should have a lifetime long enough to fulfill the application requirements. However, as a result of the limited energy supply for sensor nodes, extending the lifetime of Wireless Sensor Networks (WSNs) has been a basic target for a significant amount of research during the last decade.

Consider a static sensor network that sensors are randomly deployed. As mentioned a sensor node has a wireless communication interface through which it can communicate with other devices in its vicinity. Due to the limitation of the energy source and due to the fact that communication is the overcoming power consumer in a sensor node, the transmission range of these nodes is limited for energy-efficiency purposes. Sensor nodes that are spatially far from the sink node use multi-hop forwarding to deliver data to the sink. Multi-hop communication results unbalanced energy expenditure in the different parts of the network; nodes around the sink deplete their energy much faster than distant nodes. Not only does this stop those nodes near the sink from functioning, it also renders the sink unreachable by other nodes. In this case, the sensed data cannot be successfully delivered to the sink. The lifetime of the sensors close to the sink becomes the bottleneck for the network lifetime.

1.1 Related work

One way to balance the load over the network is to deploy more nodes in areas closer to the sink. The authors in [2] proposed a non-uniform node distribution strategy that divides the sensing field into a number of coronas and gives the ratio in node densities between two consecutive coronas. One problem of this approach is the exponential growth of the total number of sensor nodes in the network. Moreover, since the area around the sink will have too many nodes, there will be a need for a complicated MAC protocol to control the access of these nodes to the wireless channel and/or to manage their duty cycles.

Other way to extend the network lifetime is to exploit the node mobility in mobile WSNs such that to balance the energy consumption. Actually Mobility of sensor nodes is feasible, and it can be accomplished in different ways [3]. For example, sensors can be equipped with mobilizers for changing their location. Mobility-based energy conservation schemes can be classified depending on the nature of the mobile element: a mobile sink (MS) or a mobile relay (MR) that our scheme belongs to mobile sink [4]. For example of recently proposed schemes the Greedy Maximum Residual

Energy (GMRE) is introduced [5]. According to GRME, the MS selects as the new location (among feasible sites) the one which is surrounded by nodes with the higher residual energy. In order to obtain information about the residual energy, a special sentinel node is selected around each feasible site. Sentinels get the energy information from the surrounding nodes and answer the query coming from the MS. The MS uses this information to decide whether or not it should move. Another heuristic-based relocation scheme is considered in [6], where the MS selects its new location in proximity to the nodes with the higher traffic generation rates.

Saad et al. [7] present a solution to the problem of mobile sink in hierarchical structure sensor networks. The mobile sink starts at a fixed position and follows a well-planned moving path, which ends with the sink returning to the start position. Sensor nodes are randomly deployed in the area. Sensors are organized into clusters and cluster heads are selected. A cluster head has the role of gathering information from the nodes in its cluster, saving data in a buffer, and then communicating data to the mobile sink when it gets in the range.

Many other approaches proposed in the literature about sensor networks with mobile sinks (MSs) rely on a Linear Programming (LP) formulation which is exploited in order to optimize parameters such the network lifetime and so on. For example, in [8] the authors propose a model consisting of a MS which can move to a limited number of locations (sink sites) to visit a given sensor and communicate with it (sensors are supposed to be arranged in a square grid within the sensing area). During visits to nodes, the sink stays at the node location for a period of time (sojourn). Nodes not in the coverage area of the sink can send messages along multi-hop paths ending at the MS and obtained using shortest path routing. The authors derive a LP formulation in order to obtain the optimal sojourn times at each sink site. A similar approach, exploiting multiple MSs, is proposed in [9].

In this paper, we argue for using one or more mobile sink and we propose an approach based on fuzzy logic for placing these sinks in a way that balances the energy expenditure and increases the lifetime of the network. Fuzzy logic is a powerful technique for dealing with human reasoning and decision-making processes. A fuzzy expert system uses fuzzy logic instead of Boolean logic. In other words, a fuzzy expert system is a collection of membership functions and rules that are used to reason about data. Unlike the previous work we considering several important and influential parameter simultaneously to decision making about change location of mobile sink, then we design fuzzy system for select best new location.

The rest of this paper is organized as follows. Section 2, we describe network model and present our scheme. Section 3 shows the Simulation results. Finally, in Section 4, Conclusion and pointing out some related future research directions.

2 Network model and proposed scheme

We consider a WSN consisting of N sensor nodes and one mobile sink, those sensor nodes collects data from the vicinity environment and use multi-hop forwarding to deliver data to the closer mobile sink. We assume the transmission range, which is modeled as a disk, of all sensor nodes is fixed. We make other assumptions regarding the network model, such :

- The sink has sufficient power supply while nodes are battery-powered batteries with limited energy.
- All nodes have similar characteristics (e.g., range of radio coverage, energy of batteries, etc.).
- Nodes do not move after they are deployed, and each node has at least one route, consisting of wireless links, to the sink (i.e., the network is not partitioned).
- For each sensor node n_i , the location $loc_{\ i}$, the residual energy e_i, and the data generation rate t_i are known. The location of a sensor node can be obtained using the GPS or other Position estimation techniques [10]. The values of loc_i, e_i and tr_i are estimated at the node itself and in form of special packet send to sink.

There are several definition criterions for network lifetime in WSNs, for example, the time until the first sensor node dies or the time until a particular proportion of the sensors die. We can use both of them in our approach.

2.1 Fuzzy sink mobility scheme

In our method, sensing field is divided to several regions and the central point of all regions select as a candidate location set, that new location of mobile sink can be selected among this set. The number of these regions that we choose is enough suitable that guarantees select among candidate location set is near optimal decision for new location of mobile sinks.

In proposed scheme, we use fuzzy system for decision making about new location of mobile sink. We choose several influential parameters as inputs of system and the output of this fuzzy system shows the score of each regions as a new location of mobile sinks. Finally, the region with the highest score has the most priority to be chosen as the next position of the mobile sink.

Before we proceed to our fuzzy system, we describe how to calculate the inputs of system. First, we consider a timer T1 in each node. At intervals t_1 every node n_i generate special control packet and send it to sink. This packet contains values of node traffic (tr_i), residual energy of node (e_i) and node locations (loc_i). There is another timer T2 in the sink. At intervals t_2 that t_2=1.5t_1 , sink using the values of received control packets calculate the inputs of fuzzy system. At first, according to the value of loc_i, sink estimate the number of node in each region (num_j), then average traffic (tr_j) and average residual energy (e_j) of each region is calculated as follows:

$$tr_j = \frac{\sum tr_i}{num_j} \quad \text{that } n_i \text{ is in } region_j \qquad (1)$$

$$e_j = \frac{\sum e_i}{num_j} \quad \text{that } n_i \text{ is in } region_j \qquad (2)$$

Finally, we normalize these values and use of them as inputs of fuzzy system and calculate the score of each region as a candidate new location of mobile sink. Normalization means that every variable is normalized in the same range, e.g. [0, 1]. Fuzzy expert system is composed of four parts: fuzzification, fuzzy inference engine, Fuzzy Rule Base, defuzzification. Fig. 1 shows the architecture of the proposed system.

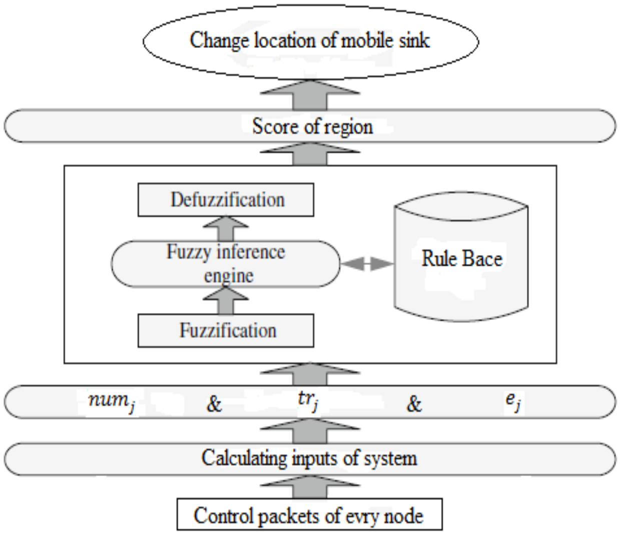


Fig. 1.architecture of the proposed system.

Each input crisp variable value needs to be first fuzzified into linguistic values before the fuzzy decision processes with the rule base. Suppose that X = $\{x_1 x_2 \ldots x_n\}$ is the problem domain of fuzzy model, fuzzification is responsible of maps x into the interval [0 1] using of a membership function μ(x). We use triangular fuzzification as a most commonly used membership function that is shown in Eq. (3) , and Fig. 2 presents the membership function for inputs of our system.

$$\mu(x) = \text{Triangular }(x;\, a,b,c) = \text{Max}\left(\text{ Min}\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \qquad (3)$$

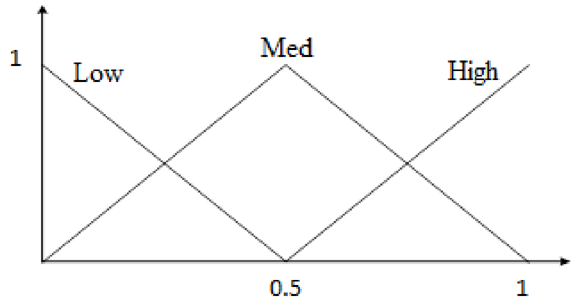


Fig. 2. Membership function of inputs

Knowledge base stores the rules which are used by the fuzzy inference engine to get a new fact from them. Fuzzy logic system uses fuzzy IF-THEN rules. Table 1 show some of the rules in our fuzzy system.

Table 1. Some rules in fuzzy mobility sink scheme

IF			THEN
num_j	tr_j	e_j	y
Low	Low	-	0
Low	-	Low	0
Low	Med	Med	0.12
Low	Med	High	0.25
Med	Med	High	0.62
Med	High	Med	0.5
Med	High	High	0.75
High	Low	Low	0.12
High	Med	High	0.87
High	High	Low	0.12
High	High	Med	0.62
High	High	High	1

We use of Sugeno-style fuzzy inference that does not need to defuzzification. In proposed system And method is a T- norm such as min, or Prod. According to the value of x for each input and membership function, degree of compatibility is calculated. Then using of Prod and degree of compatibility values, firing Strength (α_i) obtained. Implication method is multiplication and final output are achieved from the following equation:

$$y = \frac{\sum \alpha_i . y_i}{\sum \alpha_i} \qquad (4)$$

We consider final output as a score of region and the region with the highest score is chosen as a new location and mobile sink will change his location to there.

3 Simulation results

In this section, we analyze and compare the performance of our proposed scheme that we named it FMS. We perform our simulation in NS-2 simulator. In our simulation, nodes are divided into sensor nodes and mobile sink node, then several properties added to each other as follow:

- In every node add one filed as a node type whereby the ordinary node is distinct from the sink.
- In sensor nodes add timer T1 as a Scheduler for sending control data and one function to generate control packet and send it to mobile sink.
- In mobile sink add timer T2 as a scheduler for receiving and processing control packet and one function to receive control packet from sensor nodes. In addition, we have implemented our fuzzy system in mobile sink.

The sensors are deployed randomly in area and we use standard parameters of the channel and radio model. Used values for the parameters are given in Table 2. We compare our proposed scheme with two other schemes: a static scheme where sink node are static, a mobile scheme that sink follow pre-defined path for collecting data from sensor nodes.

Table 2. Simulation parameters

Parameter	Value
Simulator	NS2
MAC	IEEE 802.15.4
Number of node	100
Area size	150*150 m
Channel bandwidth	2Mbps
Application traffic	CBR (Variable in time and different nodes)
Simulation time	1000s
Propagation mode	Free space

3.1 Comparison criteria

We evaluated and compared our scheme with the mentioned methods base on four criteria that each of them represents network lifetime view. Fig. 3 show average residual energy in the network. As you can see in Figure, our method lead to higher average energy in network.

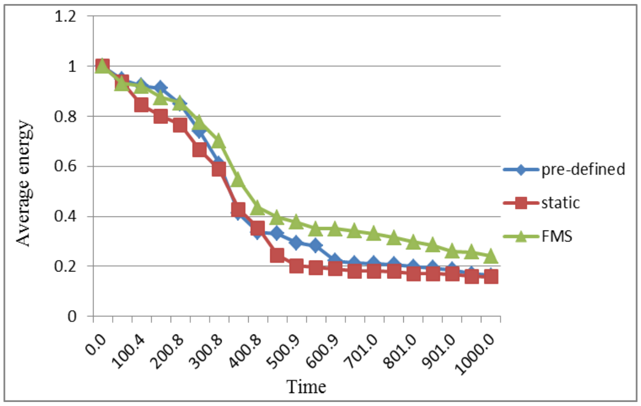


Fig. 3. Average residual energy

Clearly, higher average energy means better performance in energy conservation, but this is not enough singly. Balanced energy consumption in different parts of the network is another criterion beside the higher average residual energy that effects on network lifetime. For this purpose we use variance of energy consumption that shows residual energy of all nodes in half-time of simulation. Fig. 4, Fig. 5 and Fig. 6 show in our scheme energy consumption is more balanced.

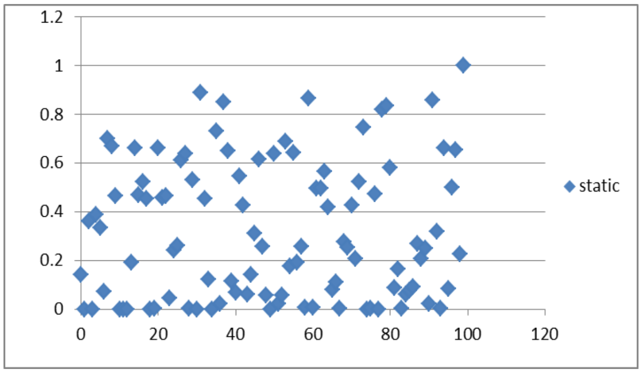


Fig. 4. Energy of nodes in static sink scheme

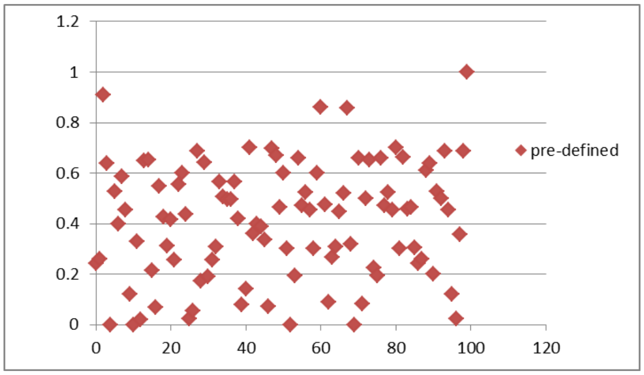


Fig. 5. Energy of nodes in pre-defined sink mobility

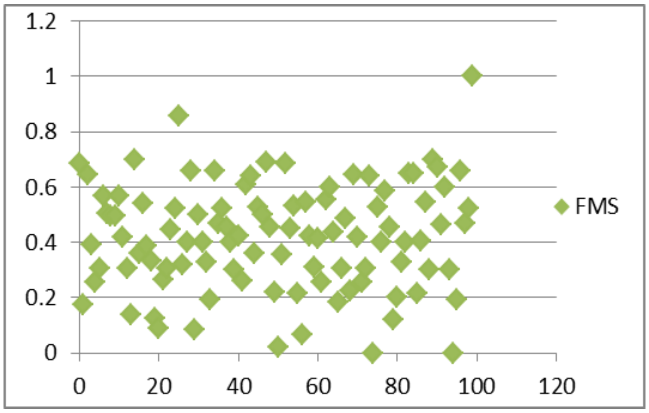


Fig. 6. Energy of nodes in fuzzy mobility sink scheme

Fig. 7 is another comparison between different methods that show number of dead nodes over time simulation. As you can see in proposed scheme nodes started to die later and total number of dead nodes is less than previous work.

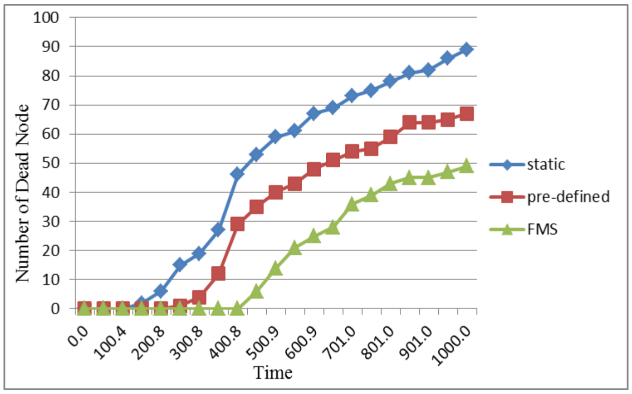


Fig. 7. Number of dead nodes

Finally, in Fig. 8 is shown network overhead that our scheme worst acts in. We consider overhead in network by calculating number of control packets and routing overhead packets in various time.

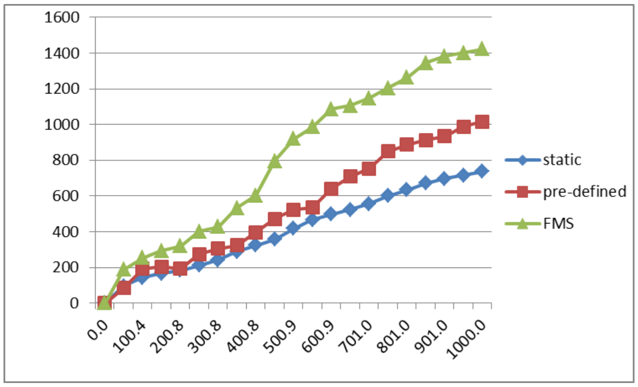


Fig. 8. Network overhead

4 Conclusion

Unbalanced energy consumption in different parts of the environment in WSNs is major limitation for network lifetime. In this paper, we proposed mobile sink approach based on fuzzy logic and expert system that selects best location in network environment for mobile sink as a extending the network lifetime. Simulation results show our scheme significantly prolong network lifetime. As a future work we can dynamically determine set of candidate location for mobile sink, instead of dividing the network into different regions, also we want improve our fuzzy system and combine it with multiple mobile sink scheme. We expect these improvements lead to more increasing in network lifetime.

5 References

[1] Anastasi, G., et al., Energy conservation in wireless sensor networks: A survey. Elsevier, 2009: p. 537–568.

[2] X. Wu, G. Chen, S. Das, On the energy hole problem of nonuniform node distribution in wireless sensor networks, in: Proceedings of the IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS), October 2006.

[3] I.F. Akyildiz, I.H. Kasimoglu, Wireless sensor and actor networks: research challenges, Ad Hoc Networks Journal 2 (4) (2004) 351–367.

[4] Cardei, M., Y. Yang, and M.I. Fonoage, Improving network lifetime with mobile wireless sensor networks. Elsevier, 2010: p. 409–419.

[5] S. Basagni, A. Carosi, E. Melachrinoudis, C. Petrioli, Z.M. Wang, Controlled sink mobility for prolonging wireless sensor networks lifetime, ACM/Elsevier Journal on Wireless Networks (2007).

[6] K. Akkaya, M. Younis, ‘‘Energy-aware to mobile gateway in wireless sensor networks”, in: Proc. IEEE Globecom 2004 Workshops, November 29–December 3, Dallas, United States, 2004, pp. 16–21.

[7] E.M. Saad, M.H. Awadalla, R.R. Darwish, A data gathering algorithm for a mobile sink in large-scale sensor networks, in: The Fourth International Conference on Wireless and Mobile Communications, 2008.

[8] Z.M. Wang, S. Basagni, E. Melachrinoudis, C. Petrioli, Exploiting sink mobility for maximizing sensor networks lifetime, in: Proc. 38th Annual Hawaii International Conference on System Sciences (HICSS'05), Hawaii, January 03–06, 2005.

[9] Alsalih, W., H. Hassanein, and S. Akl, Placement of multiple mobile data collectors in wireless sensor networks. Elsevier, 2010: p. 378–390.

[10] T. Banerjee, B. Xie, J.H. Jun, D.P. Agrawal, Increasing lifetime of wireless sensor networks using controllable mobile cluster heads, Wireless Communications and Mobile Computing (2009).

[11] Marta, M. and M. Cardei, Improved sensor network lifetime with multiple mobile sinks.Elsevier, 2009:p. 542_555.