SESSION

RADIO NETWORKS + RADAR + RFID - UTILIZATION AND APPLICATIONS

Chair(s)

TBA

FPGA-Based Implementation of a Hybrid DS/FFH Spread-Spectrum Transceiver

Stephen M. Killough[1], Mohammed M. Olama[2], Teja Kuruganti[2], and Stephen F. Smith[1]
[1]Measurement Science & Systems Engineering Division
[2]Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
Email: {killoughsm, olamahussemm, kurugantipv, smithsf}@ornl.gov

***Abstract*—In recent years there has been great interest in using hybrid spread-spectrum (HSS) techniques for commercial applications, particularly in the Smart Grid, in addition to their inherent uses in military communications. This is because HSS can accommodate high data rates with high link integrity, even in the presence of significant multipath effects and interfering signals. A highly useful form of this transmission technique for many types of command, control, and sensing applications is the specific code-related combination of standard direct-sequence modulation with "fast" frequency-hopping, denoted hybrid DS/FFH, wherein multiple frequency hops occur within a single data-bit time. In this paper, we present the efforts carried out at Oak Ridge National Laboratory toward exploring the design and implementation of a hardware prototypic hybrid DS/FFH spread-spectrum radio transceiver using a single Field Programmable Gate Array (FPGA). The high integration within a single FPGA allows the various subsystems to easily communicate with each other and thereby maintain tight synchronization. Experimental results are presented to show the receiver sensitivity and jamming-rejection capability of the implemented hybrid DS/FFH spread-spectrum system under widely varying design parameters.**

Keywords—Hybrid spread-spectrum; direct-sequence; frequency-hopping; jamming; Field Programmable Gate Array (FPGA); Phase-Locked Loop (PLL)

I. INTRODUCTION

Hybrid spread-spectrum (HSS) systems, which combine direct-sequence (DS) and frequency-hopping (FH) spread-spectrum (SS) techniques, are attractive for their strong multiple-access capabilities, resistance to multipath fading and intentional/unintentional jamming, and the security they provide against eavesdroppers [1]-[6]. In recent years there has been great interest in using HSS systems for commercial applications, particularly in the Smart Grid (SG).

Based on the hopping rate, an HSS system is classified into a hybrid direct-sequence/slow frequency hopping (DS/SFH) system or a hybrid direct-sequence/fast frequency hopping (DS/FFH) version. In hybrid DS/FFH systems, multiple frequency hops occur within a single data-bit time. Specifically, each bit is represented by chip transmissions at multiple frequencies. If one or more chips are corrupted by multipath or interference in the RF link, statistically a majority should still be correct. Standard or slow frequency hopping, in contrast, transmits at least one (and usually several) data bits in each hopping interval. DS/FFH systems have not been previously widely implemented in many commercial or industrial applications since fast frequency-hopping rates were limited by the technology of frequency synthesizers. Today's extremely fast hopping speed direct-digital synthesizers (DDSs) [7] are rapidly becoming an alternative to the traditional frequency-agile analog-based phase-locked loop (PLL) synthesizers. Output frequencies with micro-Hertz resolution and sub-degree phase tuning capabilities can thus be readily achieved using a single integrated circuit (IC).

Most of the works related to HSS in the literature have addressed evaluating its performance under different modulation techniques [1], channel conditions [2], multi-user interference [3], jamming [4], and their combinations [5], [6]. These works have shown that hybrid DS/FFH outperforms the existing standard DSSS and FHSS methods on wireless networks. In this paper, we present the efforts carried out at Oak Ridge National Laboratory toward exploring the design and implementation of a hardware prototypic hybrid DS/FFH spread-spectrum radio transceiver using a single Field Programmable Gate Array (FPGA). The high integration in a single FPGA allows the various subsystems to easily communicate with each other and thereby maintain tight synchronization. The hybrid DS/FFH prototype is optimized for a typical SG utility application. We present the challenges we faced in the design and implementation stages and how we overcome them. Experimental results are presented to show the receiver sensitivity and jamming-rejection capability of the implemented hybrid DS/FFH spread-spectrum system under widely varying design parameters.

II. TECHNIQUES FOR HYBRID DS/FFH IMPLEMENTATION

The hardware implementation of a hybrid DS/FFH system requires more advanced programing techniques such as Software Defined Radio (SDR) to allow the various subsystems to be implemented in a single Field Programmable Gate Array (FPGA). This high integration allows the various subsystems to easily communicate with each other and maintain good synchronization. Implementation on a single

FPGA also allows the various local oscillators and other timing circuits to be coherently locked in phase thus insuring proper phase alignment for the radio signals. This is especially important for the various circuits that turn on and off between frequency hops.

Although the FPGA maintains good alignment among the various segments within the transmitted packet, the analog circuitry associated with the radio causes phase discontinuities when the radios hop to a different frequency. This is because the antennas, analog filters, and outside terrain all have a phase-versus-frequency characteristic that will cause the radio signal to have a different carrier phase relationship compared to operation on the previous frequency. Although it is technically possible to calibrate for this effect, the hybrid DS/FFH system is specially designed to use a modulation method that does not depend on a consistent phase relationship between frequency hops. An additional advantage to this methodology is that the carrier phase only needs to be consistent within a single DS sequence and not long term, therefore circuitry to maintain long term phase coherence, such as a Costas loop [8], is not necessary.

There are two major hardware methods for implementing the FH portion of the system: **(1)** those being a conventional receiver with a hopped local oscillator (LO) and conventional detector, or **(2)** a fixed LO and wide receiver with the channel separation and detection performed in software. There are particular cost and performance advantages with each technique. The hopped LO enables a conventional radio except for the agile LO frequency. Off-channel interference can be thoroughly rejected with additional analog filtering and dynamic range is only limited by the linearity of the input stage transistors. Achieving rapid switching to a new LO frequency precludes the use of a single Phase Lock Loop (PLL) since the loop cannot lock to a new frequency quickly enough. An alternative is to have two separate PLL oscillators that hand off the LO task to each other, with one oscillator performing the LO task while the other one is locking to the next frequency. Another alternative is to use a direct digital-synthesis oscillator because of its rapid switching frequency. Another advantage of the conventional radio approach is that the intermediate frequency can be lower, which would enable a slower analog-to-digital converter to be used. However, the slower sample rate would not significantly reduce the size or speed of the FPGA, since the computational limitation of the FPGA is from the correlation algorithms that are required for both of the two FH methods. A significant disadvantage of the hopped-LO approach is that the receiver will only be able to listen on one frequency at a time. Although a specific frequency can be prearranged for the radios to make their initial contact via the packet preamble, there would be no provision for making contact on another frequency if the intended frequency is being jammed. However, if a very precise time reference is available on both the transmitter and receiver, it would be possible to coordinate a changing initial contact frequency.

We decided to use the SDR methodology because of its flexibility in changing the system to evaluate new concepts. The methodology has also proven to be very powerful in that the vast majority of the signal processing components can be placed in a single FPGA, which enables tight synchronization and communication between the subsystem components. The entire HSS band is down-converted to an intermediate frequency, digitized, and sent to the FPGA. Within the FPGA, look-up-table based local oscillators down-convert the individual FH channels to baseband. These baseband signals are then decoded using DS correlators and stored in a buffer for subsequent delivery to a host computer.

Software implementation of the detection and second down-conversion algorithms enable very stable and consistent performance between the individual FH channels. Although phase consistency between FH channels is not required at this time, the availability of this consistency would be useful for higher performance versions of hybrid DS/FFH in the future. The SDR implementation also allows the receiver to receive more than one radio at the same time. This is useful for high-throughput systems, but this has also been a crucial feature on the present radio implementation because it allows redundant detection of the packet preamble. To provide jamming resistance, the receiver must listen on several channels at once, since any prearranged channel could be jammed.

Because of the wider bandwidth required to digitize the entire band, the SDR system requires a higher speed analog-to-digital converter and extra circuitry in the FPGA to perform digital filtering that would normally be performed in analog hardware. Since digital computing hardware is continually becoming more cost effective, the SDR implementation will not necessarily be more expensive than a traditional analog intermediate frequency system. SDR implementations still have fundamental limitations in that the dynamic range and interference rejection capability of the system will be limited by the resolution of the analog-to-digital converter. Conversely, analog systems can add more filtering to obtain very high overall performance levels.

III. ORNL Specific Hybrid DS/FFH Implementation

The hybrid DS/FFH prototype was designed to demonstrate the fundamental advantages of the HSS system, such as jamming resistance, difficulty of unwanted interception, robust performance, and reasonable cost. The prototype operates in the unlicensed 902-928 MHz ISM band, although target applications such as the SG may ultimately use a dedicated frequency band.

The work in [9] discusses the optimal selection of hybrid DS/FFH parameters, such as DS code length, frequency hopping rate, and packet length. These parameters can be optimized with respect to jamming resistance, channel capacity, interference to other users, and difficulty in eavesdropping. The parameters chosen for the hybrid DS/FFH prototype are considered to be nearly optimal at this time, based on the available ISM bandwidth and FPGA capabilities, although more optimum values may be chosen in the future.

As shown in Fig. 1, the HSS unit splits the 902-928 MHz band into ten separate FH channels, each of which sends a DS spread spectrum signal with a 1.25-MHz chipping rate. An analog mixer converts these frequencies up or down for the transmitter or receiver, respectively, for use by the digital-to-analog or analog-to-digital converters. The SDR algorithms work over a designated 12.5-35.0 MHz frequency range.

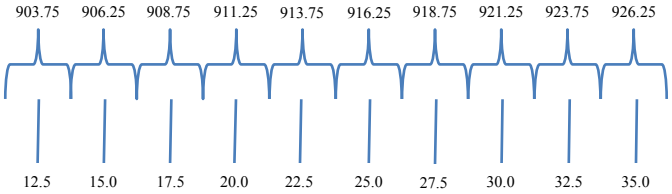


Fig. 1. The prototype hybrid DS/FFH FPGA radio frequencies in MHz.

Each DS signal is a 63-bit length Maximal Length Linear Feedback Shift Register code, although more advanced Gold or Kasami codes could also be used. After each 63-bit length code is transmitted, the system hops to a new frequency. The same data in the DS signal is repeated three times on three different frequencies, and at the receiver a two-of-three majority voting decision scheme determines the correct information even if one of the frequencies is completely blocked.

Of particular interest is the methodology for modulating the DS signal. Traditional PSK modulation requires a preamble at the beginning of the packet to determine the reference phase and a Costas Loop or similar mechanism to maintain this phase reference. With HSS in multipath channels, this phase reference is lost after each frequency hop, so HSS performs its DS modulation by shifting the start time of the code. The incoming signal is correlated with local copies of the shifted code pattern and an early-late voting system determines the amount of shift of the received signal. The correlation algorithm is independent of the carrier phase of the signal. The number of bits that can be encoded by this method is demonstrated by the early-late diagram described in Fig. 2.

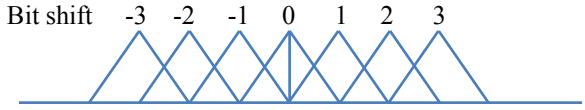


Fig. 2. Code-Phase-Shift-Keying modulation for the DS signal.

The bit-shift number refers to the amount of bits that the local DS code has been shifted for performing the correlation. To prevent ambiguous results from a correlation being between two bits, only every other bit position is used, which results in 31 positions available for each code word. The HSS prototype has a separate in-phase (I) and quadrature (Q) channel within each DS sequence, with a different DS code used for the I and the Q phases. For convenience, only 16 of the 31 positions are used for each of the I and Q. This results in an even 8 bits, per DS sequence. The I and Q channels are combined in an offset QPSK arrangement to provide a near constant-envelope signal. Four bytes of blank data are sent at the beginning of the packet as a preamble to set the reference DS start time.

A different interpretation of this methodology would be that the DS code is shifted because of a different time-of-flight, similar to GPS or continuous wave radar. Similar to the way GPS can achieve precise time-of-flight resolution, it can be expected that this methodology can be further developed to obtain higher bit capacity. Ref. [10] explores this methodology for multiple users occupying a channel simultaneously.

The HSS channel capacity is calculated by dividing the chip rate, or 1.25 MHz, by the 63-bit code length to get 19,841 DS sequences per second. Since the data is replicated three times for redundancy, the actual throughput is 6,613 DS sequences per second. Since each DS sequence contains 8 bits of data, the data throughput is 52,910 bits per second. The HSS prototype is optimized for reading household utility meters for SG applications and thus only requires 32 bytes, although the system has operated successfully with 256-byte packets.

IV. SDR Implementation

The prototype hybrid DS/FFH system is based on a Xilinx Virtex-4 FPGA for performing the digital signal processing. The hardware setup is described in Fig. 3. The FPGA, A/D, and D/A operate synchronously together at 100 MHz to allow operation on analog signals to a practical limit of 40 MHz. The D/A has 16-bit resolution for a dynamic range of 96 dB, and the corresponding A/D has 14-bit resolution for a dynamic range of 84 dB. The microcomputer loads and unloads data to the FPGA and communicates with sensors and other computers using Ethernet, RS232 or analog signals.

Fig. 4 describes the transmitter portion of the FPGA code, which consists of the data buffer, modulator, and ten local oscillators for generating the hopping carriers. Raised-cosine waveshaping is used to reduce the spectral sidebands. The receiver uses the same local oscillators for detecting signals, and all ten channels must be simultaneously received to detect the preamble during jamming situations. To acquire the packet preamble, a spread-spectrum correlator continually looks for the initial DS pattern on all channels. Once the preamble is detected, an internal timing sequence compares the signal with shifted copies of the DS code via a simple correlator. The shifted copy of the DS code that provides the strongest correlation then demodulates the actual data.

The preamble-detection section of the receiver is shown in Fig. 5. To make the signal detection independent of the carrier phase, both phases of the carrier (I and Q) are correlated with the preamble's DS code. However, the phase relationship must remain consistent during the duration of the DS sequence.

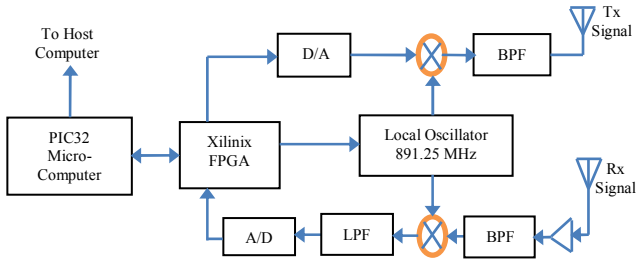


Fig. 3. Hardware setup for the hybrid DS/FFH prototype.

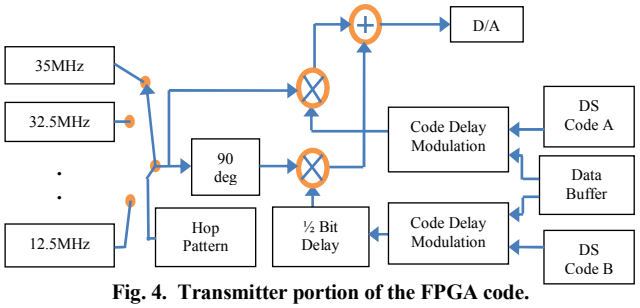


Fig. 4. Transmitter portion of the FPGA code.

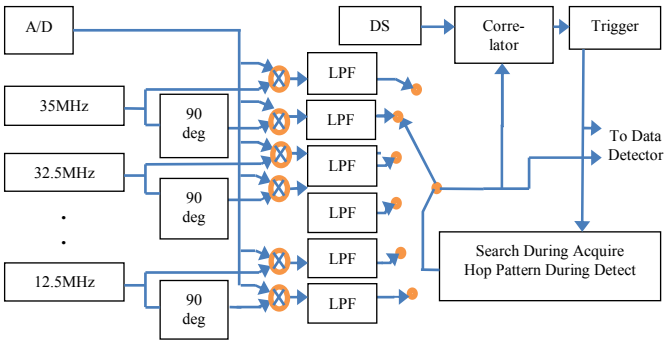


Fig. 5. Preamble-detection section of the receiver.

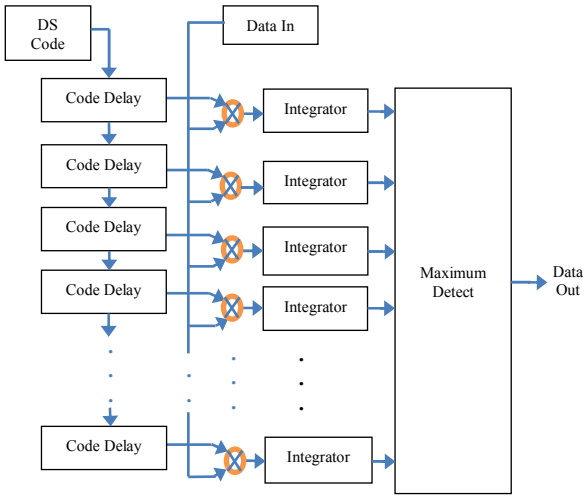


Fig. 6. Data-detection section of the receiver.

A key limitation of the radio's selectivity is the digital low-pass filter (LPF) implemented in the FPGA. Because we were limited to integer arithmetic in the FPGA, the filter was implemented as a simple square-window FIR LPF, with four of the filters connected in series. A future implementation of HSS could use a newer generation FPGA with floating-point arithmetic to achieve a filter with better rolloff characteristics and higher ultimate rejection. Fig. 7 is an analytically generated plot of the low-pass filter response, superimposed on the frequency spectrum of the spread-spectrum signal. The ultimate rejection level of 70 dB will be apparent in the experimental results presented in the next section.

Once the packet start has been established, the receiver begins listening on specific channels instead of all channels. A simple multiply-and-integrate correlator system is used for signal detection. This system is described in Fig. 6.

V. Experimental Results

Four bi-directional hybrid DS/FFH radio transceivers have been built and are performing well. The hardware prototype is shown in Fig. 8. The sensitivity for the units is –110 dBm to produce an approximately 80% success rate at the packet level. This is 5 dB less sensitive than theoretically possible, but it is expected that the detection algorithms in the SDR could be significantly improved for better overall sensitivity.

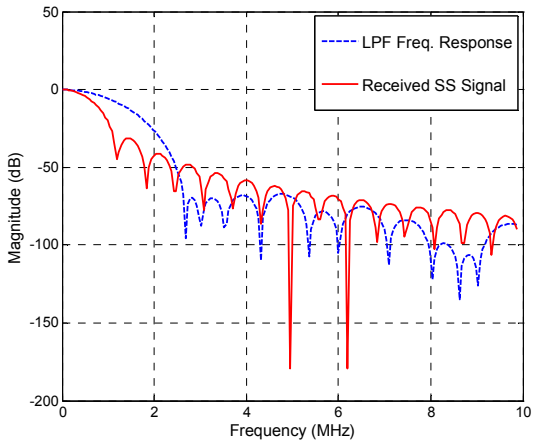


Fig. 7. The frequency response of the digital LPF implemented in the FPGA, superimposed on the frequency spectrum of the received SS signal.

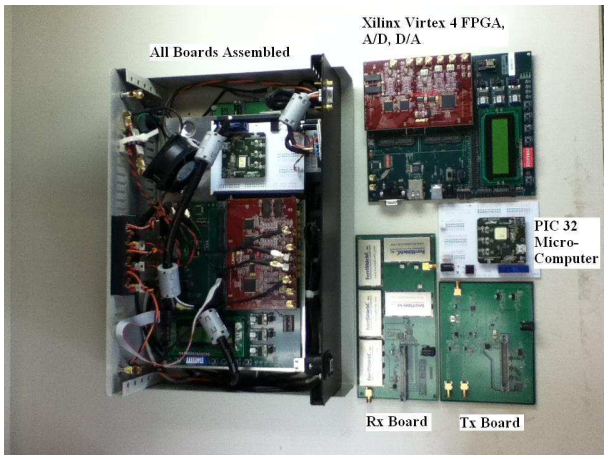


Fig. 8. The implemented hybrid DS/FFH prototype.

The jamming performance of the system was measured directly with laboratory equipment. The testing method used for the HSS evaluation is shown in Fig. 9. The square-wave generator is used at 20 kHz to modulate the signal generator at 100% AM modulation. The test procedure consists of initially transmitting data from the transmitter to the receiver with the signal generator turned off and the attenuator adjusted such that the receiver is operating at an 80% success rate. The attenuator is then reduced 20 dB so the system has a 20-dB margin. Then the signal generator is turned on and ramped up in power until the receiver has degraded to an 80% success rate. The difference in power between the signal generator (jamming) and the transmitter and attenuator combination (at the 20-dB margin point) is then recorded. This is repeated for signal generator frequencies from 902 to 928 MHz. Versions of the test are performed with and without the AM modulation. This methodology stresses the radio by exposing clipping and other non-linear effects that are expected in the A/D converter, SDR arithmetic, and analog front-end components.

A very interesting discovery during the tests was that the system performed better when the analog automatic gain-control (AGC) function was turned off. Normally the AGC

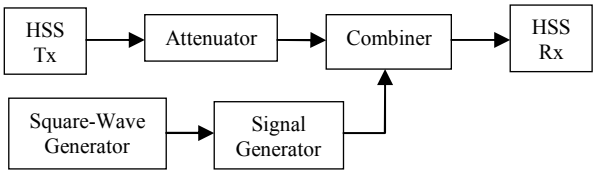


Fig. 9. Experimental setup for testing hybrid DS/FFH jamming resistance.

sets the signal strength such that the full range of the A/D is being used. This occurs since the AGC responds to the stronger interfering signal, which causes an undesired amplitude modulation of the desired signal. It is fortunate that the SDR system has enough dynamic range to detect weak signals when they are not boosted by an AGC amplifier. The following tests were thus conducted with the AGC turned off.

Since an AGC is not being used for this version of HSS, it is important to choose a proper amount of amplifier gain in the receiver. To reach a compromise between sensitivity and non-linear distortions caused by strong signals, two gain versions of the HSS were evaluated for performance. The difference in gain between the low-gain and high-gain version is 5 dB, and eventually an automatic adjustment will be developed to choose the best value for a particular environment.

The first test involved operating the HSS with the hopping feature turned off, so that the filtering capability of the SDR could be measured independently from the hopping benefits. For this test the intermediate frequency was always 12.5 MHz, which also allowed us to insert an analog 12.5 MHz, 3-pole bandpass filter (BPF) in line. This filter lets us operate the radio as a standard analog radio and allows us to do a direct selectivity comparison between the analog and SDR approaches. This comparison was made with the lower-gain version of the radio and the generator AM modulation turned off. The net results are shown in Fig. 10.

From the filtered version of the results, we still see the dynamic range limitations of the analog components ahead of the filter, which include the front-end amplifiers, surface acoustic wave (SAW) bandpass filters, and first mixer.

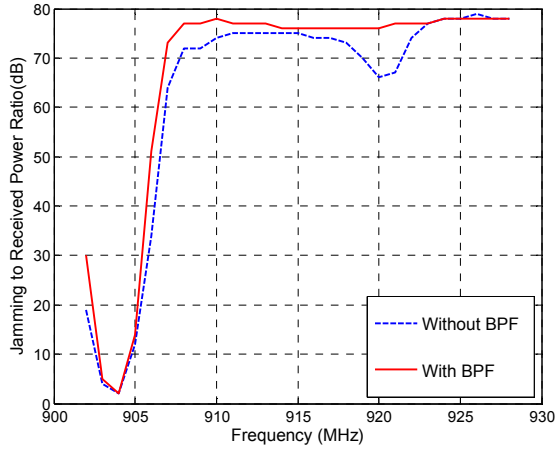


Fig. 10. The hybrid DS/FFH prototype performance while the frequency-hopping feature is disabled and with no jamming.

As a side note, the weak performance at 920 MHz can be explained by the corresponding IF signal being at 29 MHz. If this signal is strong enough to cause clipping in the A/D, the third harmonic at 87 MHz will produce a signal at 13 MHz and will thus jam the desired signal. For the SDR results, we can see that the SDR portion of the system has a dynamic range capability that is comparable to the analog portion, thus the SDR methodology causes only moderate performance degradation as compared to an analog system.

Also noteworthy is the 2-dB result (at 903.75 MHz) when the jammer is directly on the operating frequency. Typical QPSK systems require about a 6 dB signal-to-noise ratio (SNR) to operate, but since the HSS system works with random carrier phase it cannot reject the noise from the other quadrature phase, so the noise is doubled. This means that the HSS will require a 9-dB SNR. The process gain from the 63-bit length spread spectrum code is 16 dB; therefore, the HSS should theoretically tolerate a signal 7 dB stronger than the intended signal. Thus the HSS is within 5 dB of the theoretical.

Fig. 11 demonstrates the effect of AM modulation on the jamming signal. Peak values of the jammer signal are used for the comparison. In general, the modulation makes the radio 10 dB more susceptible to jamming. Although the analog gain stages do not use an AGC, the preamble detection correlator sets a threshold based on the overall signal strength and would thus have some susceptibility to AM jamming.

Sensitivity curves were generated for both the low- and high-gain versions of the radio and are shown in Fig. 12. The results are the percentage success rate at the packet level, averaged over 150 packets, with no error correction used.

These curves are unusual compared to typical digital radios because of their abrupt change from failure to success over a narrow power level range. This is due to the spread spectrum nature of the signal and in particular because of the asynchronous correlator used to detect the packet preamble. Typically for HSS, if the preamble is found, the rest of the packet is received error-free. Determining thresholds for the preamble detection was a particular challenge, and this is an area where there is potential for improving the HSS design.

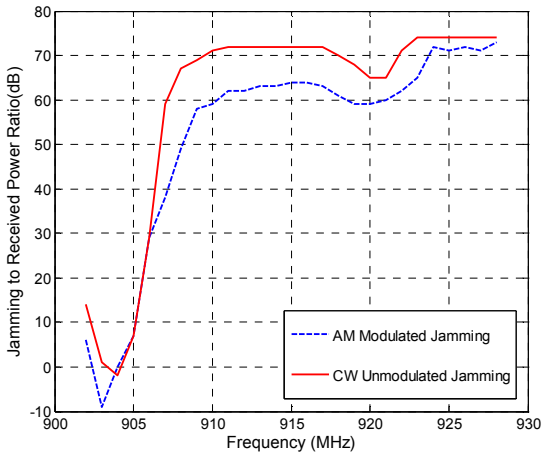


Fig. 11. The hybrid DS/FFH prototype performance while the frequency hopping feature is disabled and in the presence of jamming.

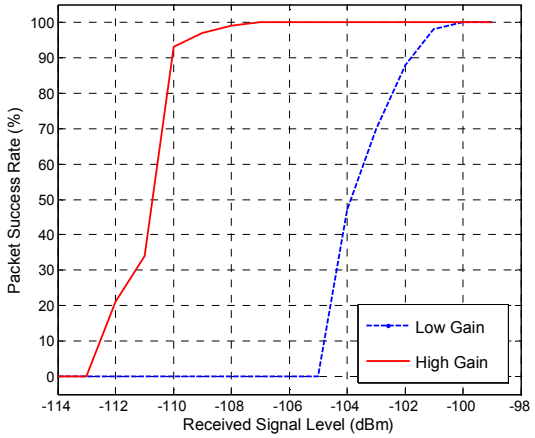


Fig. 12. Receiver sensitivity of the hybrid DS/FFH prototype performance.

The main test for HSS is to show that its FH will make the system jam-resistant at all jamming frequencies. Experiments showed that the hopping frequencies have to be judiciously chosen such that within a redundant triplet, no two of the three frequencies would be near each other, since this would let a single jammer jam both frequencies. Therefore the pattern could not be truly random but would need somewhat of a trend. Another limitation was caused by the characteristics of the analog first mixer. Since it was a double-balanced mixer, the second and fourth harmonics in the output were suppressed but the third and fifth harmonics were significant. For example, when the jamming frequency is 10.8 MHz at the A/D (902 MHz radio frequency), both the 12.5-MHz channel and the 32.5-MHz channel would be jammed with this single frequency. This issue will be solved in future HSS versions, but at this time the HSS is set to not use the top two channels. Using the analog version of HSS with the hopped LO would be a potential solution to this issue.

Fig. 13 shows the hybrid DS/FFH jamming susceptibility versus frequency. It is noticed that the smaller signal has less distortion and is able to better reject the undesired frequencies.

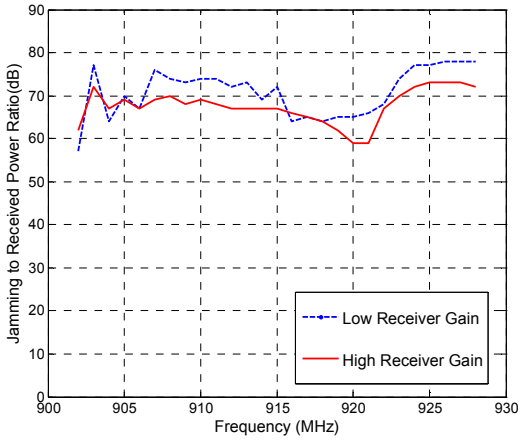


Fig. 13. The hybrid DS/FFH prototype performance in the presence of jamming.

VI. Conclusion

A hardware FPGA-based hybrid DS/FFH prototype was implemented successfully and optimized for a typical Smart Grid utility application. Experimental results indicate that high resistance of hybrid DS/FFH systems to other jamming signals allows the possibility of intentionally operating several HSS radios in the band simultaneously. For Smart Grid applications, this would enable a base station to service several clients at the same time, provided the system arranged for different clients to use different hop patterns and DS codes, and possibly even coordinated transmission time windows. The absence of an AGC in the receiver and the wide dynamic range also indicates the system will have quite good near-far performance.

Acknowledgment

This paper has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] E. A. Geraniotis, "Noncoherent hybrid DS-SFH spread-spectrum multiple-access communications," *IEEE Transactions on Communications*, vol. 34, no. 9, pp. 862-872, 1986.

[2] J. Zhang, K. C. Teh, and K. H. Li, "Error probability analysis of FFH/MFSK receivers over frequency-selective Rician-fading channels with partial band noise jamming," *IEEE Transactions on Communications*, vol. 57, no. 10, pp. 2880-2885, 2009.

[3] M. P. Pursley, "Direct sequence spread spectrum communications for multipath channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 3, pp. 653-661, 2002.

[4] J. H. Lee, B. S. Yu, and S. C. Lee, "Probability of error for a hybrid spread spectrum system under tone jamming," *Proc. of the IEEE Military Communications Conference (MILCOM'90)*, pp. 410-414, 1990.

[5] M. M. Olama, S. F. Smith, T. Kuruganti, and X. Ma, "Performance study of hybrid DS/FFH spread-spectrum systems in the presence of frequency-selective fading and multiple-access interference," *Proc. of the IEEE International Workshop on Communications Quality and Reliability (CQR)*, pp. 1-5, May 2012.

[6] M. M. Olama, X. Ma, T. Kuruganti, S. F. Smith, and S. M. Djouadi, "Hybrid DS/FFH spread-spectrum: A robust, secure transmission technique for communication in harsh environments," *Proc. of the IEEE Military Communications Conference (MILCOM'11),* pp. 2136-2141, Nov. 2011.

[7] A Technical Tutorial on Digital Signal Synthysis, Technical Report, Analog Devices, Inc., 1999.

[8] D. Taylor "Introduction to synchronous communications, a classic paper by John P. Costas," *Proc. of the IEEE*, vol. 90, no. 8, pp. 1459–1460, Aug. 2002.

[9] X. Ma, M. M. Olama, T. Kuruganti, S. F. Smith, and S. M. Djouadi, "Determining system parameters for optimal performance of hybrid DS/FFH spread-spectrum," *Proc. of the IEEE Military Communication Conference (MILCOM'12)*, pp. 1-6, Nov. 2012.

[10] Y.-R. Tsai, "M-ary Spreading-Code-Phase-Shift-Keying modulation for DSSS multiple access systems," *IEEE Transactions on Communications*, vol. 57, no. 11, pp. 3220-3224, Nov. 2009.

An Innovative Design of Infant Rooming-in Tracking Mechanism - The Experience of Cathay General Hospital in Taiwan

Lun-Ping[1] Hung, Kuo-Chung Chu[2], Shu-Hui Weng[3], Tuan-Ting Lu[4]
[124]Dep. of Information Management, National Taipei University of Nursing and Health Sciences
No.365, Ming-te Road, Peitou District, Taipei, Taiwan, R.O.C.
[3]Dep. of Information System, Cathay General Hospital, Taipei, Taiwan, R.O.C.
e-mail: lunping@ntunhs.edu.tw

Abstract - *The promotion of Baby-Friendly Hospital Initiative (BFHI) was launched in 2001 in Taiwan. To increase the ratio of requesting rooming-in, baby-friendly hospitals should prevent the mistake of identifying puerpera and infant and assure the safety of newborn babies. To reach the goal of facilitating the rooming-in care, we propose a tracing system using active RFID-tag and RSSI method to identify and monitor neonates. This system is implemented in Cathay general hospital in Taipei. With the integration of wireless devices and information technology, this system can effectively avoid the situation like stolen baby and switched baby and up to the standard certified by BFHI. The nursing department can easily arrange routine nursing works and increase the quality of nursing cares. The hospitals are benefit from using the system with increased ratio of 24 hours infant rooming-in care and fulfill the requirement of the baby-friendly hospital more effectively.*

Keywords: RFID, RSSI, infant rooming-in, BFHI

1 Introduction

The Baby-Friendly Hospital Initiative (BFHI) was launched in the 1990s by the World Health Organization (WHO) and United Nations Children's Fund (UNICEF) as a global effort with hospitals health services. The promotion of BFHI is intended to improve the service quality in the hospital's department of gynecology and obstetrics and create a breastfeeding friendly environment in the hospital to support and encourage breastfeeding. Breastfeeding is the best starting point to newborn babies. The infant rooming-in care can not only boost the maternal bonding with the newborn baby, but also can build mother's confidence of being required and relied by the infant. Due to the previous advantages, medical institutions have gradually obtained the certificate of BFHI and start practicing 24 hours infant rooming-in care. [1]

When performing 24 hours rooming-in care, newborn babies stay in the ward with the mother. However, there are few occasions like bathing, injection, or other nursing care that require to move newborn babies out of the ward. Babies might be stolen or miss placed during transportation. These kinds of tragedies are seriously harmful to both families and medical institutions. The traditional way of identifying newborn babies is to compare the identification band wearing on baby's wrist or ankle and the identification card hanging on baby's hospital bed. However, newborn babies look similar and are not easy to distinguish their differences which lead to possible identification error. This is the main reason why most parents with newborn babies refuse to accept rooming-in care and this is the main obstacle of promoting rooming-in care.

The issue of patient safety has been paid attention by many countries including England, United States, Australia, etc. In the report titled "National Patient Safety Goals, NPSGs" published by the Joint Commission on Accreditation of Healthcare Organization (JCAHO), correctly identifying patient is the first and primary goal. [2] The issue of patient identification has gradually gained attention by public. Among the various topics discussed about patient identification, the identification of newborn babies is especially important because newborn babies can not "confirm" their identity which leads to the importance of developing the mechanism to identify newborn babies. [3,4]

Along with the growing of the necessity of efficient wellness systems, there is a mounting demand for new technological solutions able to support remote and proactive healthcare. Wireless transmission through mobile devices combining with information technology can break through the blind spots occurred while objects are moving. Radio Frequency Identification (RFID) technologies can assist the construction of baby rooming-in environment in the baby-friendly hospital.

RFID technologies are composed of three components: reader, tag, and software system. Through the microchip on the tag to transmit ID information to the back-end database, objects like babies can be identified, traced, and confirmed. The advantages of using RFID are non-contact reading,

updated information, massive amount of data storage, better data safety, and capable of reading multiple objects simultaneously. [5,6] With the help of electronic tagging device and the Real Time Location Systems (RTLS), RFID mechanism can monitor baby's location in real time. Electronic tagging device contains batteries that allow it to actively detect the signals sent by readers in surrounding area and to transmit data to readers. RFID technologies are mainly used in the obstacle-occupied environment and the longest transmission distance can be reached more than 100 meters. RLTS is a system that can locate the position of specific targets promptly using wireless transmission technology in a confined space. Currently, technologies using RTLS to perform locating services are divided into four categories: Angle of Arrival (AOA), Received Signal Strength Indication (RSSI), Time of Arrival (TOA), and Time Difference of Arrival (TDOA). [7] TOA and TDOA locating system are both based on time as the base of measurement and both require the measurement of the time for transmitting signals precisely to receive correct results. Comparing with previous two methods, RSSI and AOS locating systems use signal strength as the base of measurement. However, these two methods can receive incorrect results because of the interruption of obstacles and multiple routes.

To provide a safe baby rooming-in environment and prevent the mistake of switched or missing babies, this paper proposes an infant rooming-in tracking mechanism using RFID technologies which can produce a statistical index record approved by BFH. The rest of this paper is organized as follows: section 2 gives a detailed presentation of the methodology of building the infant rooming-in tracking mechanism and section 3 shows the experiment results. We conclude the paper in section 4 with remarks on future work.

2 Methodology

The infant rooming-in tracking system framework includes the active RFID positioning mode and RSSI positioning method. The rooming-in rate plays an important role in the designing of the system and is discussed in this section as well.

2.1 The System Environment Framework

As shown in Fig. 1, the mobile area refers to the place for the baby's bath, vaccinations, or other nursing measures requiring the mother to go back and forth from the baby-friendly maternity ward (hereinafter referred to as the ward) to the baby room within the 24-hour rooming-in period. Inside the mobile area, active RFID readers are installed in important locations. When signals sent out by the RFID tag worn by the baby are received by the reader located closest to the tag, the reader sends out information, including the reader MAC address, the tag ID, the received signal strength and so on. This information is sent back to the middleware server of the RFID control center.

The RFID control center includes four components: Middleware server, system database, Hospital Information System (HIS), and rooming-in tracking system. When the middleware server receives the tag data sent back by the reader, through positioning calculation and processing, the multiple information content in the data is converted into interpretable original data and is stored in the system database. Once the rooming-in tracking system receives a query request from the remote services, the system will obtain positioning information and patient related information from its own database and the HIS. Then, through information integration and logic processing, a response is sent back to the device interface where the user sends out the request of the remote services. Remote services refer to remote user devices that sends out request commands to the rooming-in tracking system, including desktop computers used by nursing personnel, mobile devices for puerpera.

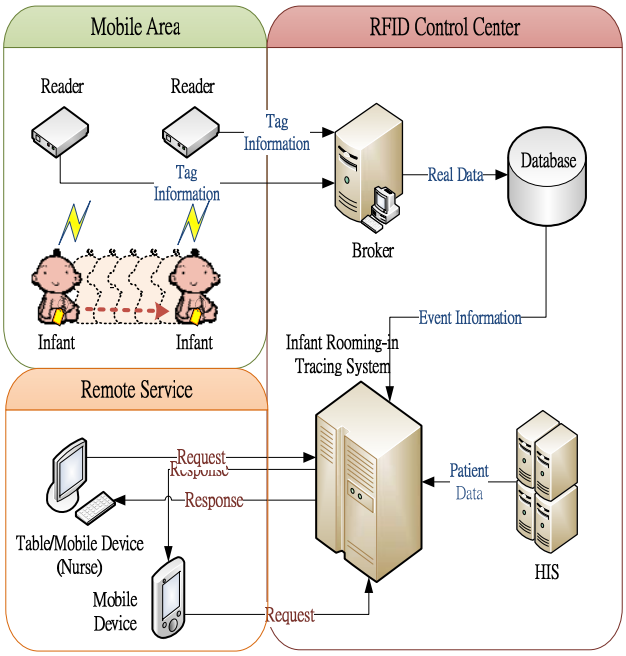


Figure 1. System environmental framework

2.2 The deployment of the active RFID positioning system

In this study, the 2.45 GHz active RFID tag was worn around the ankle of newborns for positioning and tracking because newborns often wave their hands near their faces. As shown in Fig. 2, it is relatively safer and less likely to be detached if wearing the tag around the ankle. 4 hours after the birth of the newborn, the tag was put on the newborn, and it was removed before the newborn was discharged from the hospital. Before putting on and after taking off the tag, the newborn is exposed under the risk of identification "window period". The active RFID tag is enclosed in a waterproof case. Rinsing and disinfecting with alcohol are allowed. Therefore, recycled uses are acceptable.

Active RFID network readers are located at each ward door on the fifth floor, entrances and exits of the baby room on the second floor, and the newborns' mobile spaces and ceilings of elevator entrances on the second and fifth floor, which are used to collect newborn positioning and tracking related information. The reading range of a reader can reach above 100m. This distance is adequate for the newborn safety during transportation.

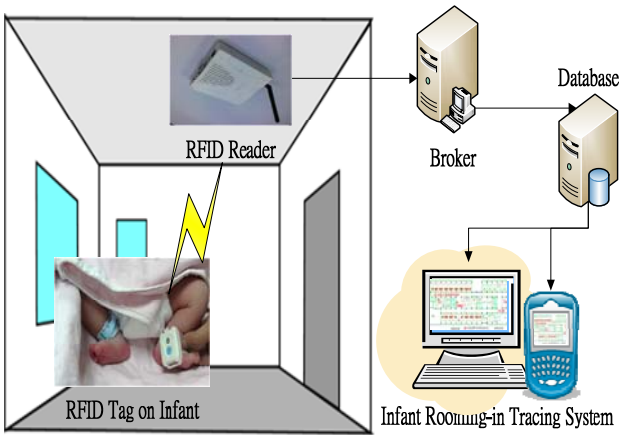


Figure 2. The deployment of the active RFID positioning system

2.3 The Active RFID Positioning Method and RSSI Positioning Method

During the 24-hour rooming-in period, the newborn underwent bath, vaccinations, and other nursing measures. The newborn is necessary to go back and forth from the ward to the baby room; therefore, in this study, active RFID readers were installed in important indoor locations within the mobile spaces. When the signals sent out by the active RFID tag are received by the reader closest to the tag, the reader will send its own information such as the reader MAC address and information carried by the tag, such as tag ID, RSSI received signal strength, and so on back to the intermediary software for the process of positioning. Then, the intermediary software converted the multiple data into interpretable original positioning data for storage in the system database. Once the rooming-in tracking system received a query request sent by the remote devices, the system automatically obtained related information from its system. After processing, it responds to the user's device that sent out the request, such as desktop computers, smart phones, etc.

3 Experimental results

Based on the medicinal and clinical on-site environment, integrated information monitoring management system was developed, as shown in Fig. 3, to provide medical management personnel a user-friendly interface and real-time location monitoring information. As for the path tracking under special circumstances, detailed information for queries was designed, as shown in Fig. 4. In addition, based on the statistical model of the rooming-in standard construct, the server received information was converted into standard rooming-in records for storage. In Fig. 3, the functions of statistical index related records are listed in the area covered with an oval. The matching list of mother's name and newborns' name is in the area covered with dot-line. Newborns' current location is shown in the area covered with black line.

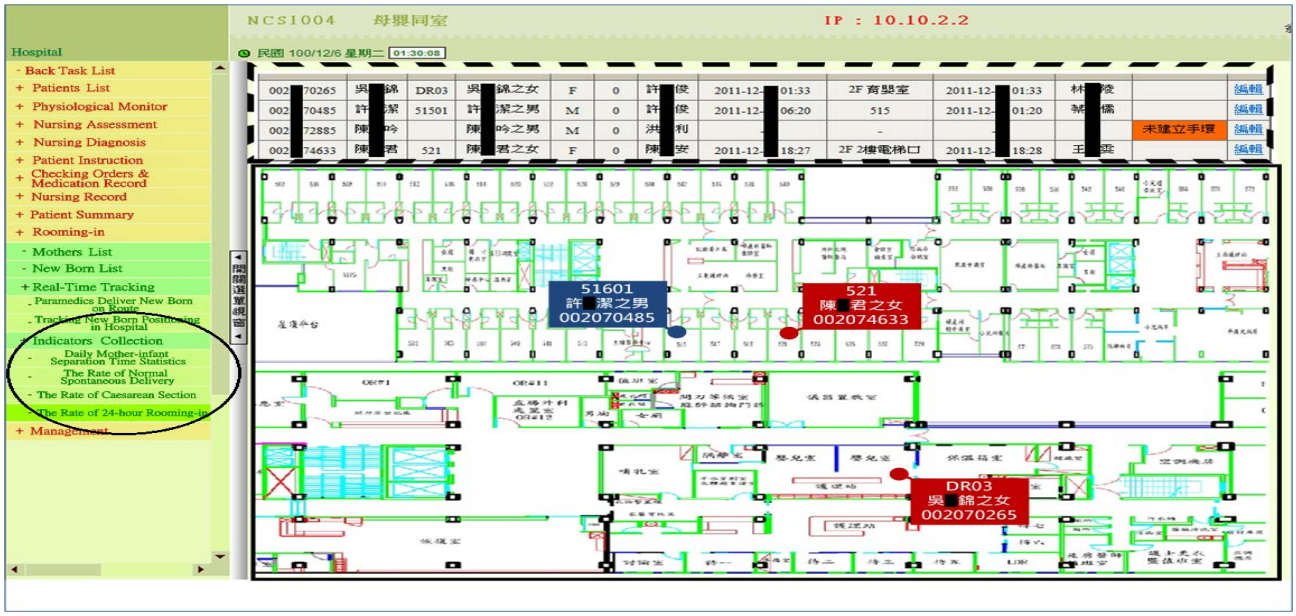


Figure 3. Rooming-in monitoring system

3.1 Development Tools

In view of the J2EE platform, the web-based application system with MVC as the framework was developed and constructed. At the front-end, the Java Servlet & JSP technology was mainly adopted as the basis, coupled with the SVG dynamic mapping technology to present a visually user-friendly interface. At the back-end, the RFID intermediary software-based MS SQL Server database and the HIS-based IBM DB2 database were connected. Finally, under the safe and flexible IBM WebSphere environment, the complete application program was constructed, deployed, and executed.

3.2 Statistical Definitions

According to Measurement 7 of the criteria certified by baby-friendly hospitals in 2011: the description of the "baby-friendly maternity ward implementation" assessment and the project content of baby-friendly hospitals under the Taiwan Association of Obstetrics and Gynecology [8], the following rooming-in statistical definitions and formulas were introduced.

3.2.1 The 24-hour Rooming-in Rate

The so-called vaginal delivery (caesarean session) pregnant women's 24-hour rooming-in rate refers to the rate of pregnant women that undergo vaginal delivery (caesarean session), which is calculated by dividing the number of people engaged in 24-hour rooming-in by the number of pregnant women opting for normal delivery. Among them, the number of 24-hour rooming-in are expressed as the number of 24-hour rooming-in babies delivered through vaginal delivery (caesarean session), while the number of pregnant women opting for vaginal delivery (caesarean session) among the pregnant women opting for normal delivery is expressed by the number of babies delivered through vaginal delivery (caesarean).

$$\text{24H Infant Rooming-in Rate of Vaginal Birth} = \frac{(\text{Vaginal Birth Infant} \cap \text{24H Infant Rooming-in})}{\text{Vaginal Birth Infant}} \%$$

3.2.2 The Implementation of the 24-hour Rooming-in Certified Criteria

The certified criteria for implementing the 24-hour rooming-in include: three months before the on-site certification, at least 10% of the pregnant women hospitalized for vaginal delivery (normal newborns) implemented the 24-hour rooming-in during the period of hospitalization, and at least 5% of the pregnant women among the pregnant women opting for caesarian session (normal newborns) implemented the 24-hour rooming-in during the period of hospitalization.

$$\text{Certificate Standard of 24H Baby Friendly} = \left(\frac{\text{24H Infant Rooming-in}}{\text{Rate of Vaginal Birth}} \geq 10\%\right) \cap \left(\frac{\text{24H Infant Rooming-in}}{\text{Rate of Caesarean Birth}} \geq 5\%\right)$$

3.3 The Display of Information Monitoring System

Fig. 4 shows the route and related information presented by the tracking and query of escort locations during the escort process. The detailed information about the path tracking is shown in the area blocked with black line. The red connected part consists of the path and the frame produced from the message feedbacks, which were effectively detected by the RFID. When the tag reached the elevator, the RFID reader automatically sent a message to the post-end server. Valid message requests continued to be sent, and new messages continued to be updated to present the real-time

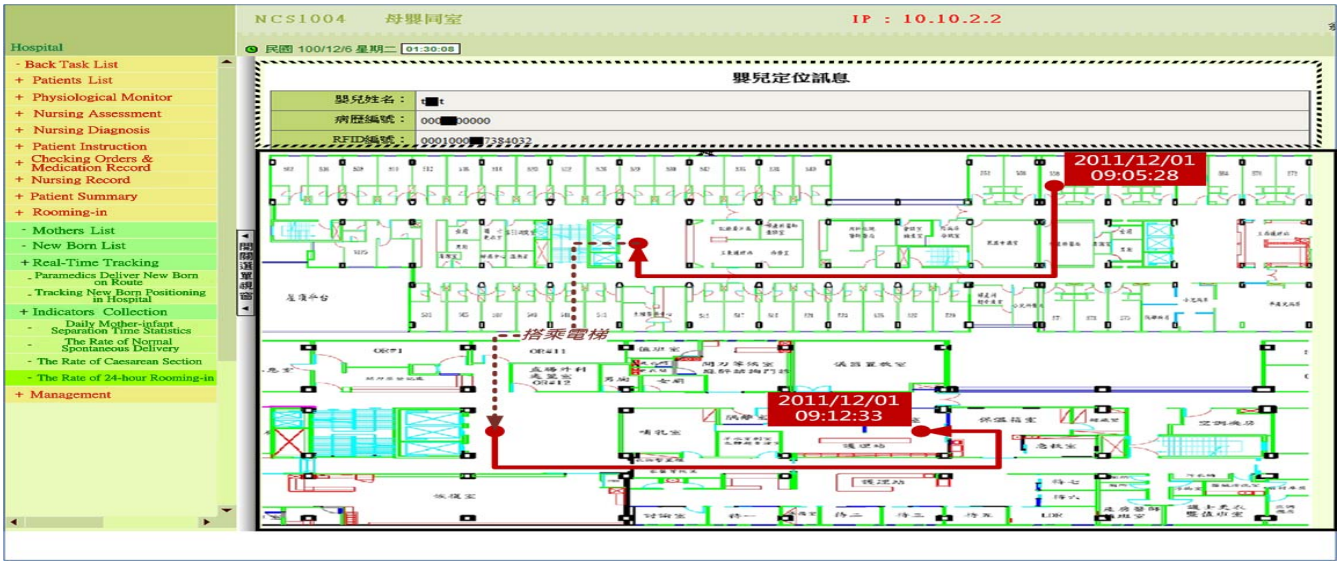


Figure 4. Detailed information for path tracing

4 Conclusions

This paper propose an infant rooming-in tracking system that provides a friendly and promptly tracking interface, traces newborn babies' delivery path and time, presents the distribution of infants, and calculates the total amount of newborns precisely. This system can identify newborns and their location. It also keeps records of their location, delivery path, and staying time to effectively avoid the mistakes of identification error. It also sends alarms when abnormal situation occurred to prevent the incident of missing babies.

In terms of management effectiveness, the proposed system can enhance the quality of the process control for 24 hours rooming-in service and can produce various indexes and statistical reports approved by BFHI that were used to be collected manually. By using the proposed system, hospital management can be performed more effectively. In the future, we intend to develop visualized tracking mechanism which combines cloud computing technology and smart phone as the base of the designed mechanism and allows nursing personnel to conduct routine nursing tasks and the same time to acquire prompt information about rooming-in services.

Acknowledgements

We are grateful to the Cathay General Hospital for providing resources and support. This research was supported by Cathay General Hospital, Taiwan, R.O.C. under contract number CMRI-10104.

5 References

Number in square brackets ("[ ]") should cite references to the literature in the main text. List the cited references in numerical order at the very end of your paper (under the heading `References'). Start each referenced paper on a new line (by its number in square brackets).

[1] D. J. Karl, J. A. Beal, C. M. O'Hare, and P. N. issmiller, "Reconceptualizing the nurse's role in the newborn period as an "attacher", American Journal of Maternal Child Nursing, vol. 31, no. 4, pp. 257-262, 2006.

[2] 2012 Hospital National Patient Safety Goals, [Online], http://www.jointcommission.org/assets/1/6/2012_NPSG_HAP.pdf

[3] World Health Organization (WHO). (2011) [Online]. http://www.who.int/nutrition/topics/bfhi/en/index.html

[4] R. T. Mercer, "Predictors of maternal role attainment at one year post birth", Western Journal of Nursing Research, vol. 8, no. 1, pp. 9-32, 1986.

[5] K. Finkenzeller, RFID Handbook - Fundamentals and Applications in Contactless Smart Cards and Identification. New York: John Wiley & Sons, 2003.

[6] L. Figarella, K. Kikirekov, H. Oehlmann, "Radio frequency identification (RFID) in healthcare: benefits, limitations, recommendations:, Health Industry Business Communications Council (HIBCC), Phoenix, Arizona, A HIBCC white paper, 2006.

[7] Cisco Systems, Inc. (2008) Wi-Fi Location-Based Services 4.1 Design Guide - Location Tracking Approaches. [Online]. http://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/wifich2.html

[8] Taiwan Association of Obstetrics and Gynecology. (2011) [Online]. http://www.taog.org.tw/content_04-3.htm

Towards Spectrum Resource Management in Cognitive Radio Networks via Intrusion Detection and Response Model

Obeten O. Ekabua
Department of Computer Science
North-West University, Mafikeng Campus,
Private Bag X2046, Mmabatho 2735, South Africa
(obeten.ekabuao@nwu.ac.za)

Ohaeri, Ifeoma Ugochi
Department of Computer Science
North-West University, Mafikeng Campus,
Private Bag X2046, Mmabatho 2735, South Africa
(23989688@nwu.ac.za)

Abstract: *Cognitive Radio Network (CRNs) was innovated as a means to solve the problem of spectrum scarcity. CRNs are able to detect and utilize vacant spectrum by means of dynamic spectrum access (DSS) without interfering to the primary licensed users. This has led to the increase in vulnerabilities and threats to the CRNs. The steady increase in attacks against cognitive radio network and its resources has caused a necessity to protect to these valuable assets. Among existing problems in CRNs resource management is the issues of security and privacy. CRNs are wireless in nature, they face all common security threats found in the traditional wireless networks and other new security threats and challenges that have arisen due to their unique cognitive (self-configuration, self-healing, self-optimization, and self-protection) characteristics. Traditional security measures would be inadequate to combat these challenges. There is a need to advance and improve the security standard and level for cognitive radio networks. Therefore, this research paper proposes an intrusion detection and response model (IDRM) to enhance and advance security for spectrum resource management in cognitive radio networks. IDRM monitors all the activities in order to detect the intrusion. It searches for security violation incidents, recognizes unauthorized accesses, and identifies information leakages. Unfortunately, system administrators neither can keep up with the pace that an intrusion detection system is delivering responses or alerts, nor can they react within adequate time limits. Therefore, an automatic response system has to take over this task by reacting without human intervention within the cognitive radio network. This research paper can be used as an introduction to the IDRM and analysis. It is hoped that the design and analysis of this new approach will facilitate future study and implementation of this novel security architecture for spectrum resource management in CRNs.*

Keywords: Network Management, security, authentication, authorization, access control.

1. Introduction:

Traditionally, specific bands are allocated to particular services operating under license according to spectrum allocation static policy. The expansion in wireless technology and the enormous increase in wireless devices and application have led to the lack of spectrum for emerging services. The spectrum is largely underutilized according to the Federal Communication Commission (FCC). However, CRNs are presented as solution by utilizing the vacant spectrum bands that are underutilized by the licensed users [1].

Cognitive radio network was first defined by Joseph Mintola as a network of cognitive radio. The CRNs components consist of two groups such as; the primary network and the cognitive radio network. The primary network group or licensed network is referred to as an existing network. The primary users are those who have the license (the right of access) to operate within the spectrum band of the existing network. They are given the first priority in access to spectrum. If primary network is infrastructure based, primary users activities are controlled via primary base stations. This is also called centralized network [2].

The cognitive radio network is a secondary network and does not have the license to operate in a choice spectrum band so an additional component is required to enable them share the licensed spectrum band conditionally. They can also use base stations with a single hop transmission, connecting the cognitive radio networks in their different locations. They are referred to as unlicensed users or secondary user. The secondary users are to identify or discover the white spaces (vacant bands) and select a suitable portion in other to operate without interfering to the licensed primary users. In essence, whenever the primary user presence is detected in the operational channel, the secondary users using the space switches to another band that is free. This process is referred to as spectrum handoff. They take the next priority in access to spectrum band. The cognitive radio network is also called decentralized network due to the

fact that they can cluster in different geographical locations [1].

Generally, cognitive radio networks are dynamic in its concept. This dynamic nature makes them vulnerable to attacks, which results into huge security challenges. However, spectrum brokers are deployed to distribute the spectrum resources in other to dynamically and effectively manage the flow of data and information in CRNs, and also efficiently control access to spectrum bands to avoid collision of users, and network failures.

However, this promising technology has been overtaken by several attacks which replicates on daily basis causing inefficiency in network delivery and reduces the quality of services provided. Consequently this research paper introduces intrusion detection system as suitable security mechanism to combat these attacks [3].

The structure of this paper is as follows: Section two describes firewalls as a first line of defense against attacks in CRN as appeared in literature. In section three, we specified the common intrusions to CRN. Next in section four we description of the structure of IDS and its supporting components [3].

2. Fundamental Security Objectives for Cognitive Radio Networks

Security objective differ depending on the application environment. Different combinations of these features are required based on the networks configuration, service and networks policy. However, common objectives exist that provide basic security controls in cognitive radio network environments and other wireless networks due to their operation on wireless media. Cognitive radio network is a system that employs and embraces a more complex set of heterogeneous users sharing spectrum resources, and the readiness to share is encouraged using effective and efficient protocol measures. Cognitive radio automatically detects unutilized, vacant spectrum and dynamically forms suitable number of channels in order to optimize spectrum usage, increase and improve spectrum efficiency and reduce interference [4]. It is able to adapt to service environment and adjust channels bandwidth, while considering locally used traffic distribution. Cognitive radio networks (CRNs) has three aims: to innovate, improve, and maintain existing wireless communication network [5].

However, these aims cannot be realized when the security concepts are breached. Cognitive radio network security is a customizable level of security that enables any system to organize its structure and it is able to conform to requirement changes. This security system secures monitors and analyses traffic and data packets to ensure that CRNs intrusions threats or attacks are detected, and access is granted to only the right users. It enables audit trails and keeps records of previous attacks and changes to indicate, where, when, how and who made the changes [6]. These objectives basically form the fundamental principles of any network security. The goal of this research work emphasizes mainly the intrusion detection and response model as security mechanism against all forms of CRNs intrusions or attacks to enhance security in CRNs [7].

In CRNs, reliability is achieved by applying security principles and access control measures, which involve hardware, software, applications and protocols, logical and physical policies. If specific security conditions are applied to all users of a network, information systems, and information resources, using stipulated rules, then reliability and quality of service is guaranteed. The security requirements for a reliable CRN include: availability, identification, confidentiality, integrity, authentication, authorization and non-repudiation.

(i) Availability

One of the basic objectives and aim for building a stable communication system is availability and robustness. If a network is not available, it is not usable and the objective is defeated. Security data and service profile information should be available for easy confirmations. Wireless transmission medium should always be available. The spectrum should be available for both primary and secondary users. Secondary users should not interfere or disrupt primary users by occupying the spectrum when needed. Security measures are to ensure that attacks are prevented.

(ii) Identification

This is a verification of security data and service profile information. It is a basic security objective for any communication device. It is also the process of establishing the identity of the users and other entities involved in the operations. It associates the user with a unique name. An equipment identity is assigned to all mobile devices in cellular and wireless networks called internal mobile equipment identifier (IMEI). However, tamperproof identification measure inbuilt in secondary devices is a security requirement in CRNs [20].

(iii) Confidentiality

A secured communication network such as CRN should be private and confidential for effective data and information management. This is a security requirement that ensures that only the sender and the receiver (parties and entities) involved are able to understand the communication flow. Confidentiality entails privacy and

trust relationship. This means that transmission and management of data and information (communication) among users and devices in cognitive radio network must be confidential and the entities involved must ensure a mutual agreement of trust to guarantee quality of service.

(iv) Integrity

Data packets can be intercepted or modified in transit by attackers for malicious use. Therefore, a secured communication in cognitive radio network requires integrity in order to establish effective transmission of data packets and management of data and information to achieve quality of service. Integrity ensures that data and information are not changed or modified in transit. Any change or modification must be done by the explicit consent of the entities involved. The receiving end or entity must be assured that the data packets or information received or is receiving is exactly what was transmitted from the transmitting end. Therefore, this objective ensures privacy of authorized user data and control information in cognitive radio network for effective data and information management.

(vi) Access Control

This restricts network's resources to authorized users or devices only. It ensures that every user or device in a network has the explicit right to access the resources requested for and also the privileges to perform certain tasks in a network. This objective forms the basis for validating any security mechanism.

3. Intrusions to CRN

During the design and analysis of this security model, a key aspect that should not be left out is the performance of threat analysis, of various types of threats profiles which may threaten to harm the proposed model. The high and quality performance of CRN technology lies majorly on an effective security mechanism. It guarantees the availability and robustness of network service and resources against the security challenges (threats and attacks). The following are some threats and attacks that may transpire in this CRN environment:

(i) Denial of Access

This is an unauthorized use of the spectrum band resulting into the primary (licensed) users loosing access to the network resources and services. Most times the network is been hijacked by these malicious users for selfish use and personal gains. When cognitive radio node emits power in an unauthorized spectrum, it makes primary users to lose access and malicious entities takes advantage of this nature to intrude and seize network.

(ii) Eaves dropping of cognitive messages.

Cognitive radio messages can be intercepted by a malicious user who can make use of the information to lunch several other attacks on the primary users of the network or the network itself.

(iii) License user emulation

Licensed users can be emulated by malicious users impersonating their details, camouflaging some trusted nodes, causing other node to join the network undetected, sending false routing information [21]. Transmitted packets can be intercepted while on transit by malicious users thereby having access to cognitive messages to their advantage. Malicious cognitive users can exchange or alter cognitive messages for ulterior motives and as well change cognitive radio nodes causing interference and internal node failure which can result into network failures.

(iv) Jamming of cognitive radio channels

Cognitive channels that transmit messages can also be made to jam in other to disrupt the messages passing through the network. The cognitive control channels (CCC) are made to transmit wrong messages or right messages in wrong forms. This makes the network fall short of the quality of service (QoS) assurance.

(v) Attacks to Cooperative Sensing

Usually, in CRN, cooperative sensing permits taking a decision about the primary users' presence in a particular channel. This is based on the reports provided by the cognitive radios. The secondary users sense the spectrum separately or individually and share the results with the other secondary users so as to improve the detection probability. Consequently, this can give rise malicious and selfish behaviors which include; deliberate report of false measurement by malicious node, leading to false positives, negatives or a selfish node. This does not cooperate in other to save energy. These attacks often times aims at improving the successful chances of a primary user emulation attacks.

3. IDRM Algorithm

The IDRM is designed and configured to monitor and analyse the activities of the CR Network as displayed in Figure 1.

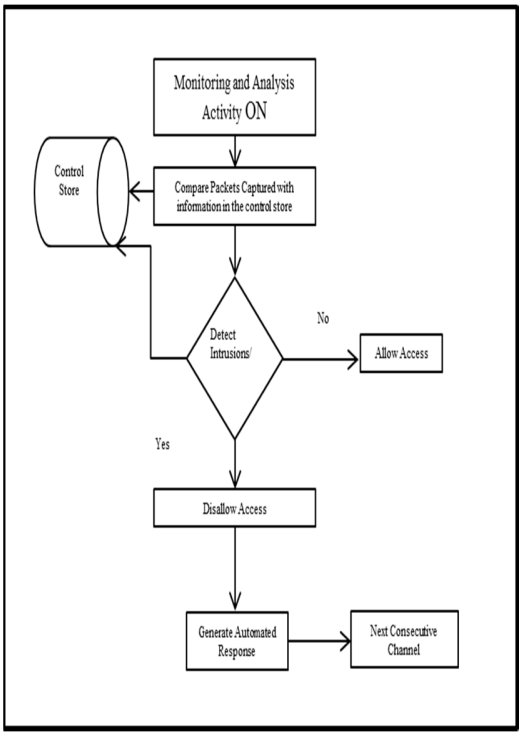


Figure 1: IDRM Algorithm

IDRM acts as a packet monitoring and analyser to monitor and analyse the nodes in connection and communication with each other within the network. It monitors the coordinators and the routers in real time. The packets that are captured are compared to the database or knowledge base of the intrusion specification, service and security policy that are previously configured on the bases of intrusions which are common to the network.

Fundamentally, the IDRM is set up to detect or discover framed packets that have formats not compatible with the networks configuration. The IDRM which is steadily tuned on starts monitoring and analysing the network traffic immediately the network connection is formed. It is connected to all the required devices or appliances (network layers, routers, coordinators, nodes, protocols) for comprehensive checking.

Transmitted packets are captured and compared with the information in the control store (access, service and security policy,) and the networks CRN database. The corresponding intrusions detected are stored in the intrusion detection database. To react to any intrusion detected or discovered, the IDRM is required to generate an automated response suitable for the attack and a form of warning to the intruder and stores the intrusion pattern into the intrusion detection database before moving to the next consecutive channel. If the IDRM detects an ID conflict, it automatically disables itself from the coordinators and performs an active scan to select the new appropriate ID. As soon as this happens, a channel message from the channel master is sent across, and promptly the operating channel is changed to the new ID and the monitoring activity is tuned on for further checking.

4. Intrusion Detection and Response Model (IDRM)

Figure 2 is an IDRM design as a security mechanism to enhance security in CRN by providing secure communication, enabling efficient resource allocation, effective spectrum usage, and access control to the limited and scarce resources. The IDRM that was designed achieved those characteristics by identifying computing or network activities that are considered as intrusions, malicious, or unauthorized. It is configured to properly scrutinize all packets at different protocol layers of the network such as; physical layer, link layer, network layer and transport layer due to accessing the spectrum dynamically.

The IDRM design shows the layout of the model and its associated components. The IDRM operates as a network-based model and provides maximum security to enhance the networks productivity and quality of service. The IDRM is designed based on three major components which are data and information source; monitoring, analysis detection; followed by the response mechanism. The operation of the model is described and analyzed according to the various components that make up the model.

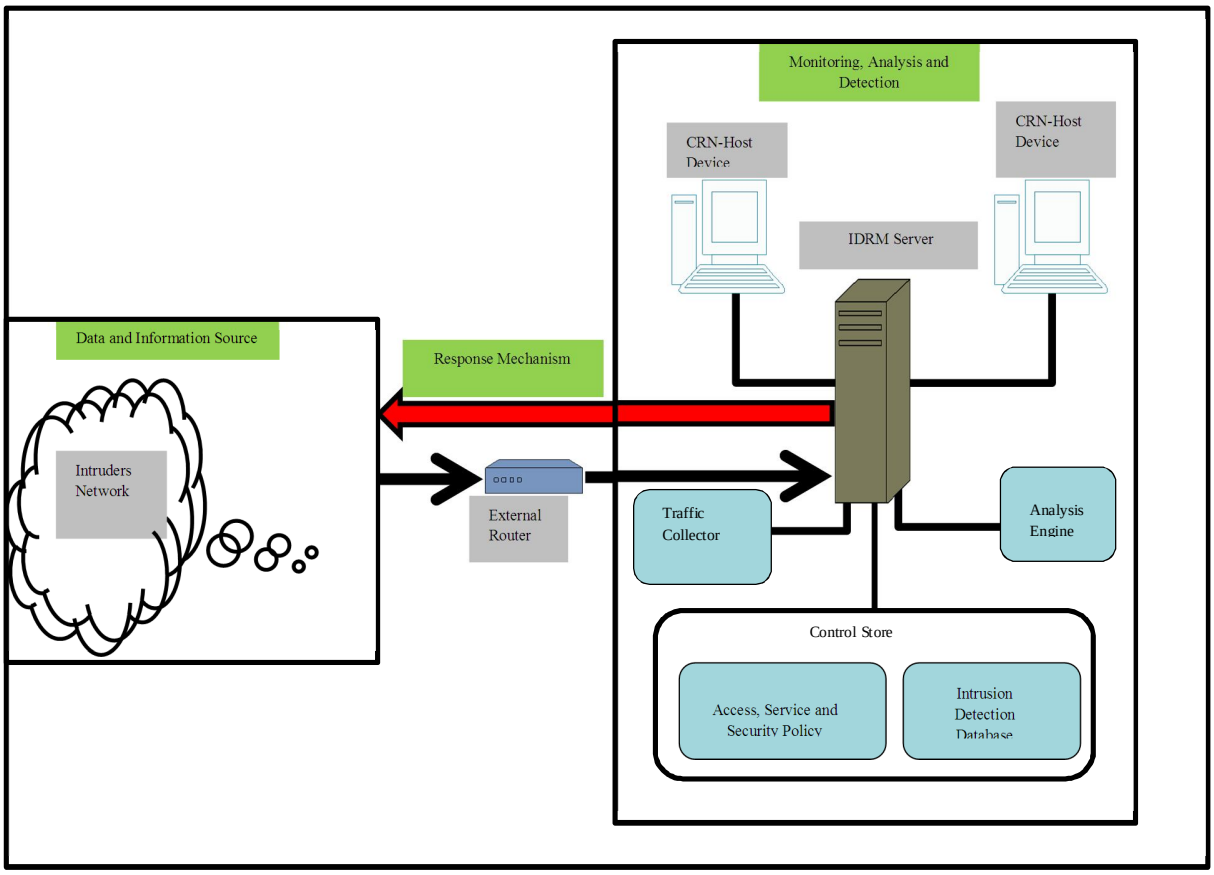


Figure 2: Intrusion Detection and Response Model (IDRM)

(i) Data and Information Source

Data and information source contains the intruders' network, consisting of all wireless and mobile devices (user access) which usually request for access or connection to the CRN network. The information sources are divided into three categories: (i) input data accumulated from individual systems (host based); (ii) input data originated from the network (network based); and (iii) data produced from other sources. However, the IDRM evaluates the operating system audit trail as its significant data source for all network layers (application, network, transport, link, physical).

(ii) Monitoring, Analysis and Detection

Monitoring, analysis and detection contain two major components which are CRN host devices and the IDRM server. Host devices consist of computers and other appliances connected to network, while the IDRM server consists of some logical components such as: traffic collector, analysis engine, access, service and security policy, and intrusion detection database. Generally, the art of monitoring, analysis and detection is the core activity of our IDRM. The IDRM is configured based on the network policy using 'systrace' - a computer security utility which limits an application's access to the system by enforcing networks access policy (service and security policy) for all system calls. By this method we determine whether a particular event is an intrusion or not. This monitoring, analysis and detection component consists of two sub components, which are IDRM Server and CRN Host Devices. The IDRM sever is made up of some logical components such as: (a) the traffic collector and (b) the analysis engine.

(a) Traffic Collector

This component of the IDRM is designed to pull traffic from the network. It collects information or activity for the IDRM to examine. This information or activities could be log files, audit logs, or incoming and outgoing

traffic on a specific system. Because IDRM is network-based, the traffic collector (component) copies traffic outside the network link. This component behaves like a network traffic sniffer where every packet transmitted along its duty path of the network is to be properly examined.

(b) Analysis Engine

This component examines the network traffic that has been collected by the traffic collector. This is done by the analyser from the analysis engine. It is regarded as the most important component of the IDRM due to its responsibility for analysis and detection. It is often referred to as the brain of the IDRM. It consists of the security policy enforcement agent (SPEA) and the security policy decision agent (SPDA). The SPEA ensures connection admission control and handoff by enforcing the respective designed policies on the subjects (network users) while, the result from that component is sent to the SPDP for implementation via SPEA based on the stipulated policy. Then a confirmation message is sent to the client via the security policy retrieval agent (SPRA). This way, the analysis engine decides the activity, communication or transmission that is allowed or disallowed. It is a decision or pattern matching mechanism. It compares the traffic and information supplied to it by the traffic collector against the networks access policy (service and security policy) and known intrusion specification (patterns) stored in the intrusions detection database. If the activity matches any known pattern, or misuse of the security policy is detected, it reacts to it as an intrusion by generating any of the automated responses based on the intrusion and gives a warning. This examination of traffic is done as quickly as possible to enable IDRM react to intrusions in real time and move to the next consecutive channel.

(c) Control Store

The control database store consists of access, service and security policy, and intrusion detection database. The analyser from the analysis engine collects relevant information pertaining to an intrusion besides the main function of identifying intrusion within the network layers. It also collects the supporting evidence and traces of the intrusions and stores them in the intrusion detection database. The networks database stores all user identity details and all normal user behaviour based on the management access, service and security policy. This enables detection of deviations as intrusions. In addition, data packets coming into the system from the five different layers of the network (application layer, transport layer, network layer, link layer and physical layer) are assembled to form complete transmission control protocol (TCP) and protocol data unit (PDU) to be analysed to check intrusions. All intrusions detected or discovered by the IDRM are duly reacted to via the response mechanism.

(iii) Response Mechanism

After analysis is done and the IDRM detects intrusions, it immediately disallows access and reacts to them by sending appropriate automated responses to the intruder's devices, such as drop the packets, shut down the port, coupled with a warning. All information about the intrusions is stored in the intrusion detection database, also referred to as the attack signature database.

Because the spectrum is accessed dynamically, IDRM is designed and configured to influence every protocol layer. The IDRM is a network-based model therefore, it resides on computer or appliance connected to a specified segment of the network. It can also be installed at specific places in the network where it can watch traffic going into and out of particular network segment. It looks for intrusion patterns as well as deviations from service and security policy when analysing or examining packets transmitted over the network.

The primary idea of IDRM is to detect CRN intrusions while allowing genuine and authorized user access. The entire network security is identified with the network traffic by classifying the allowed and disallowed traffic.

In summary, any packet transmitted from an intruder's network to any of the CRN host devices must pass through the IDRM server engine via the external router which serves as the request messenger. This is first analysed by the IDRM that monitors the entire network in order to detect intrusions or attacks that were not handled by other security mechanisms in place (the first line of defence), and also provides quick automated responses without any human intervention. This automated response is generated by the IDRM once an intrusion is detected via the response mechanism which in turn stops it from getting to the CRN host device that is the target of the intrusion. Information that is useful to track new attacks is also provided.

The intrusion detection is based on the network configurations levels such as detection level and response level .The scenario for intrusion detection using the designed IDRM is shown in Figure 4 while the algorithm and the IDRM UML diagram that further describe the IDRM are shown in Figures 1 and 3, respectively. The network IDRM security system is configured and implemented based on specification-based technique (network services and security network

policy) using systrace - a computer security utility which limits or restricts an application's access by enforcing network security and service policy for all system calls**.** Therefore, intrusion detection using IDRM is based on the information recorded in the attack or intrusion specification database. Intrusion detection database consists of the service and security policy specified by the cognitive radio network and also relevant information on the detected intrusions. Thus, attackers are restricted from invading the network for fear of their identity being revealed.

CRN is a distributed intelligent and dynamic network such that its IDRM is also distributed in its configuration to cover a large network area to provide an advanced network monitoring, incident analysis, incident response and instance attack data This enables the network security analysts (NSAs) to have broader view of the occurrences in the entire network per time and identify new intrusion patterns from the record of intrusion detection database. This enables further investigations on the detected intrusions. The implementation of the IDRM is reported in chapter 4.

5. IDRM UML Sequence

The IDRM UML diagram in Figure 3.4 describes the sequence of activities of the IDRM. It shows the operations of its sub components indicating the request and communication (challenge response) protocols.

When the client sends a network or resource request it passes through the air frequency bandwidth because of its wireless nature. The request is delivered to the traffic collector by the network resource broker (NRB) and handed over to the analysis engine which consists of the SPEA (security policy enforcement agent) and SPDA (security policy decision agent).

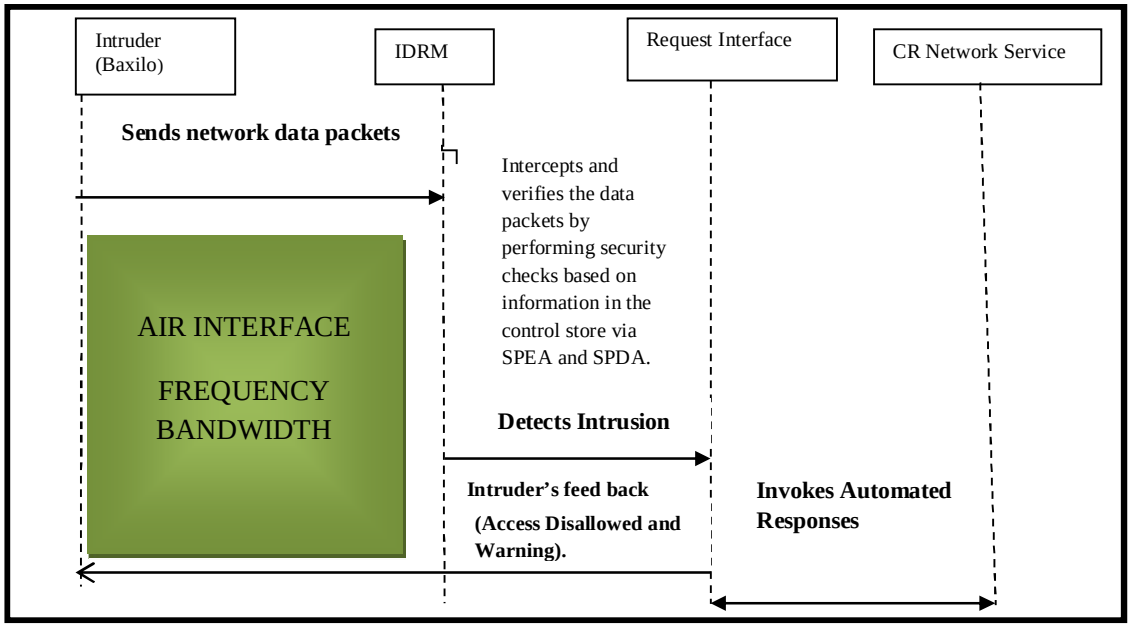


Figure 3: IDRM UML Sequence Diagram

The SPEA component of the IDRM analysis engine performs the verification activities based on the network service and security policy (NSSP). The message is then validated in line with the SPDA decision and the network service is invoked. The client is given feedback via the SPEA. The access is disallowed if an intrusion is detected or allowed if otherwise depending on the verification outcome.

6. Scenario for Intrusions Detection using IDRM

In Figure 4 is a scenario that describes how IDRM reacts to the various intrusions (vulnerabilities and attacks).

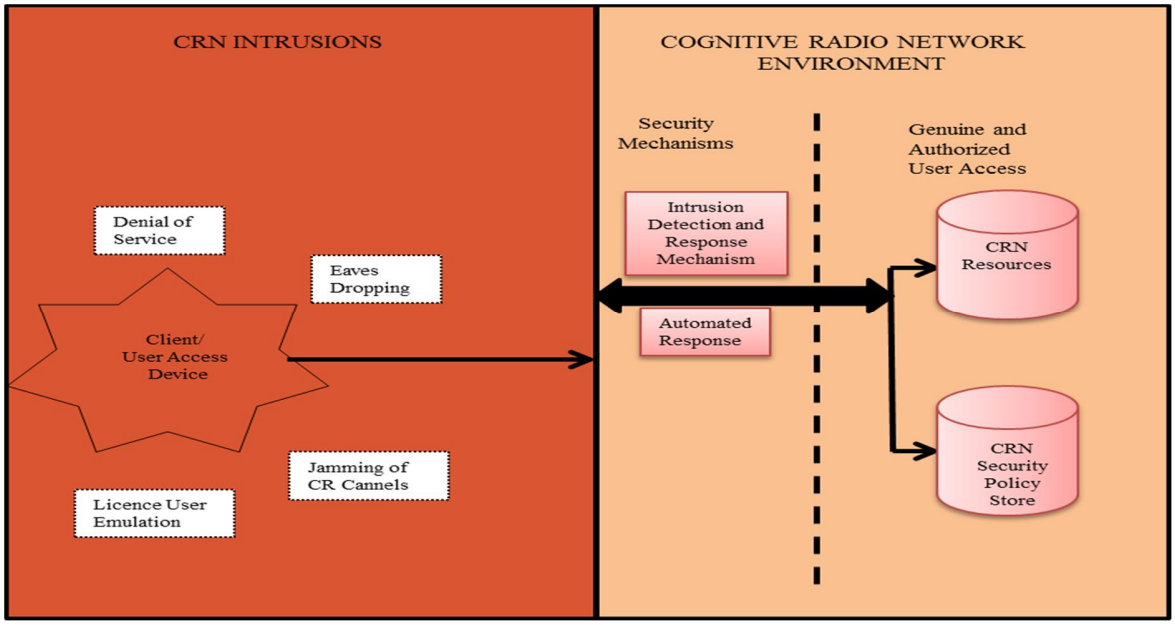


Figure 4 Scenario for Intrusion Detection using IDRM

Any of the various intrusions can be launched via the user access which can be any form of wireless or mobile device. Intrusions have been a constant danger to CRNs and have received increased attention as they can lead to a severe loss of revenue if a site is taken offline for a substantial period of time. The economic dividends provided by the innovation of CRN are not achieved. The target of the intrusion is to establish unauthorized access to the networks services and resources. The IDRM at interception performs proper analysis on the captured data packets, identifies the intrusions and sends quick automated responses to the intruder's device. This means that the IDRM does not allow the intruder to have access to the resources made available by the CRN. Detection of intrusion is done by enforcing the networks specification (service and security policy) on all system calls. Only the genuine and authorized users or clients are allowed access to the CRN resources. However, no specific intrusion or attack is implemented in chapter four as this is not within the scope of this research work. The essence of Figure

9. Conclusion

Cognitive radio offers a promise of intelligent radios that can learn from and adapt to their environment. Much research is currently underway developing various reasoning that allow cognitive radios to operate optimally. However, as with many new technologies, initial research has not focused on security aspects of cognitive radio networks. Typically, security is always "bolted on" after the fact by adding some sort of link authentication and encryption. This typically works well for data traversing a wireless network, but not necessarily for things fundamental to the operation of the wireless link itself. Since cognitive radios can adapt to their environment and change how they communicate, it is crucial that they select optimal and secure means of communications.

Moreover, with the developments of network applications, network attacks are greatly increasing both in number and severity. As a key technique in network security domain, Therefore, reported in this research project is an Intrusion Detection and Response Model (IDRM) to enhance security in cognitive radio networks. It plays the vital role of detecting various kinds of attacks and secures the networks. Intrusion detection is defined as the tools, methods, and resources to help identify, assess, and report unauthorized or unapproved network activity. IDRM is typically one part of an overall protection system that is installed around a system or device. It is used to monitor networks for attacks or intrusions and generate automated responses against the intrusions whenever detected. It also reports these intrusions to the network security administrator in order to take further actions. It is not a stand-alone protection measure. The main purpose of IDRM is to find out intrusions among normal audit data and this can be considered as classification problem. As part of intrusion detection systems, it is an effective security technology, which can

detect, prevent, and react to attacks. It performs monitoring of target sources of activities, such as audit and network traffic data in CRNs, requiring security measures, and employs proper techniques for providing security services.

Apparently, with this tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more important than ever. Intrusion detection is needed in today's computing environment because it is impossible to keep pace with current and potential intruders and vulnerabilities in our computing systems. The CRN domain is constantly evolving and changing field by new technology and the internet. Hence new attacks replicate on daily basis. Therefore intrusion detection systems are used in managing threats and vulnerabilities in the changing environment.

Attacks on computer network systems can be devastating and affect networks and corporate establishments. We need to curb these attacks by installing IDRM to identify the intrusions. Without the use of IDRM to monitor network activities, intrusions which can possibly result in irreparable damage to an organization's network can occur.

References

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications." *IEEE Journal on Selected Areas in Communications*, vol. 23, pp. 201–220, Feb. 2005.

[2] J. Mitola, and G. Q. Maguire. "Cognitive radio: Making Software Radios more Personal." *IEEE Journal on Network Communication*, vol. 6, pp. 13–18, Aug. 1999.

[3] P. Steenkiste, D. Sicker, G. Minder, and R. Dipankar. "Future Directions in Cognitive Radio Network Research," in *Proceedings of NSF Workshop Report*, March 9-10, 2009, pp. 1-3

[4] Y. Zhang, W. Lee, and Y. Huang. "Intrusion Detection Techniques for Mobile Wireless Networks." *Wireless Networks Journal*, vol. 9, pp. 545-556, 1999.

[5] Y. Zhang, and W. Lee. "Intrusion Detection in Wireless Ad-hoc Networks," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking*, 2000, pp. 275- 283.

[6] H. Fuping, S. Wang, and Z. Cheng. "Secure Cooperative Spectrum Sensing for Cognitive Radio Networks," in *Proceedings of IEEE Military Communication Conference*, 2009, pp. 1-7.

[7] V. Sharma, and Y. S Mann. "*Emerging Technologies in Web Intelligence*." *Infosys Technologies Journal*, vol.2, pp. 115-121, May 5, 2010.

A Cyclostationarity-Based Spectrum Sensing Scheme for Cognitive Radios With Dynamic Primary User Signals

Jeongyoon Shim, Youngseok Lee, Youngpo Lee, Jaewoo Lee, and Seokho Yoon†
College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do, Korea
†Corresponding author

Abstract— *This paper addresses a cyclostationarity-based spectrum sensing scheme for cognitive radios in dynamic primary user (PU) traffic environments where the PU might randomly depart or arrive during the sensing period. At first, the spectrum sensing problem in dynamic PU traffic environments is formulated as a binary hypothesis testing problem. Then, a test statistic for spectrum sensing is derived by applying an estimate of spectral coherence function of the PU signal to the generalized likelihood ratio. Numerical results show that the proposed scheme exhibits a better spectrum sensing performance than that of the conventional scheme based on the energy detection in the presence of dynamic PU signals.*

Keywords: Spectrum sensing; cognitive radio; dynamic PU signal; cyclostationarity

1. Introduction

The frequency spectrum is a limited and scarce resource, and thus, the efficient use of the spectrum resource is required. The cognitive radio (CR) is a promising technology to exploit underutilized spectrum in an opportunistic manner and the spectrum sensing technique identifying spectrum opportunities is one of the most important techniques in CR [1], [2].

Conventionally, the spectrum sensing techniques have been developed under static primary user (PU) traffic environments where the spectrum band is assumed to be occupied by the PU or to be vacant during the whole sensing period [3], [4]. Practically, however, the PU signal may depart or arrive during the sensing period, especially when a long sensing period is used to achieve good sensing performance, or when spectrum sensing is performed for a high traffic network, and under such dynamic PU traffic environments, the performances of the conventional spectrum sensing techniques have been found to degrade severely [5]. Although a spectrum sensing technique [6] was proposed based on the energy detection approach for dynamic PU traffic environments, it performs poorly when the signal-to-noise ratio (SNR) is low.

In this paper, a novel spectrum sensing scheme is proposed based on the cyclostationarity in the presence of dynamic PU signals. We first formulate the spectrum sensing problem in dynamic PU traffic environments as a binary hypothesis testing problem and develop the corresponding generalized likelihood ratio (GLR). Obtaining an estimate of spectral autocoherence function (SAF) of the PU signal and applying it to the GLR, then, we propose a test statistic for spectrum sensing in dynamic PU traffic environments. The proposed cyclostationarity-based scheme is expected to perform better than the conventional energy detection-based scheme of [6], since the cyclostationarity approach has an advantage over the energy detection approach in that its detection performance is generally better than that of the energy detection approach, and also, it can distinguish the PU signal from the interference unlike the energy detection approach.

The rest of this paper is organized as follows. In Section 2, we model the spectrum sensing problem in the presence of dynamic PU signals as a binary hypothesis testing problem. In Section 3, we develop a GLR based on the binary hypothesis model, estimate the SAF of the PU signal, and propose a test statistic for spectrum sensing by applying the estimate of the SAF to the GLR. Section 4 compares the spectrum sensing performances of the proposed and conventional schemes in terms of receiver operating characteristic (ROC). Finally, Section 5 concludes this paper with a future work.

2. System Model

We model the spectrum sensing problem in dynamic PU traffic environments where the PU randomly departs or arrives during the sensing period of CR user as a binary hypothesis testing problem: Given the received signal, a decision is to be made between the null hypothesis H_0 and the alternative hypothesis H_1 defined as

$$H_0: \quad y[n] = \begin{cases} x[n] + w[n], & \text{for } n = 1, 2, ..., J_0, \\ w[n], & \text{for } n = J_0 + 1, J_0 + 2, ..., N, \end{cases} \tag{1}$$

and

$$H_1: \quad y[n] = \begin{cases} w[n], & \text{for } n = 1, 2, ..., J_1, \\ x[n] + w[n], & \text{for } n = J_1 + 1, J_1 + 2, ..., N, \end{cases} \tag{2}$$

respectively, where $y[n]$ and $x[n]$ represent the nth sample of the baseband equivalent of the received and PU signals, respectively, $w[n]$ represents the nth sample of an additive white Gaussian noise (AWGN) with mean zero and power spectral density (PSD) $N_0/2$, and N is the number of samples available during the sensing period. Under the hypothesis H_0, the random departure of the PU occurs between the J_0th and (J_0+1)th samples, on the other hand, under the hypothesis H_1, the random arrival of the PU occurs between the J_1th and (J_1+1)th samples. Once a test statistic is obtained for spectrum sensing, the test statistic is compared with a predetermined threshold. If the test statistic exceeds the threshold, the CR user chooses the hypothesis H_1 deciding that the spectrum band is occupied by the PU; otherwise, the CR user chooses the hypothesis H_0 and utilizes the spectrum band.

3. Proposed Scheme

Applying the GLR test to the binary hypothesis model of (1) and (2) gives the following test statistic

$$\sum_{n=J_0+1}^{N} y^2[n] - \sum_{n=1}^{J_1} y^2[n] \underset{H_0}{\overset{H_1}{\gtrless}} \gamma', \tag{3}$$

where γ' is the threshold determined from a given false alarm probability (i.e, $\Pr(H_1|H_0)$). To exploit the cyclostationarity of the received PU signal, in (3), we replace $y[n]$ with the SAF $\rho_y^\alpha(f)$ defined as [7]

$$\rho_y^\alpha(f) = \frac{S_y^\alpha(f)}{[S_y(f+\alpha/2)S_y(f-\alpha/2)]^{1/2}}, \tag{4}$$

where α is a cyclic frequency, $S_y(f)$ is the PSD of $y(t)$, and

$$S_y^\alpha(f) = \int_{-\infty}^{\infty} E\Big[y\left(t+\frac{\tau}{2}\right) y^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi\alpha t}\Big] e^{-j2\pi f\tau} d\tau \tag{5}$$

is the spectral correlation density (SCD) function with $(\cdot)^*$ the conjugation operation. From (4) and (5), we can see that the SAF is the normalized version of the SCD.

Since $\rho_y^\alpha(f)$ is the SAF of a continuous signal $y(t)$, we cannot replace the discrete value $y^2[n]$ of (3) with $(\rho_y^\alpha(f))^2$ directly. Thus, we employ the discrete estimate $|\hat{\rho}_y^\alpha(f)|^2$ of the squared magnitude of the SAF obtained as

$$|\hat{\rho}_y^\alpha(f)|^2 = \frac{\left|\sum_{n=1}^{N} u[n]v^*[n]\right|^2}{\sum_{n=1}^{N}|u[n]|^2 \sum_{n=1}^{N}|v[n]|^2} \tag{6}$$

to replace $y^2[n]$ of (3), where $u[n] = y[n]e^{j\pi(f-\alpha/2)n}$ and $v[n] = y[n]e^{j\pi(f+\alpha/2)n}$ are the frequency-shifted versions of $y[n]$ and its crosscorrelation used in (6) can be obtained as depicted in Figure 1. Now, replacing $y^2[n]$ with (6) yields

$$\sum_{n=J_0+1}^{N} |\hat{\rho}_y^\alpha(f)|^2 - \sum_{n=1}^{J_1} |\hat{\rho}_y^\alpha(f)|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{7}$$

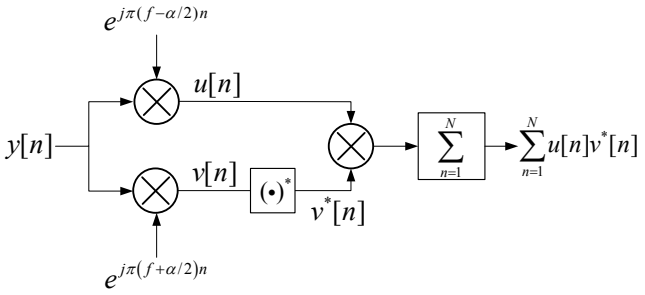


Fig. 1: The crosscorrelation of $u[n]$ and $v[n]$.

where γ is a threshold for the test statistic (7). It should be noted that the values of J_0 and J_1 change randomly depending on the behavior of the PU, and thus, the unconditional test statistics for random departure and arrival are obtained by taking the expectation over (7) with respect to J_0 and J_1, respectively. Generally, the number of events occurring randomly over a period of time is well modeled by a Poisson process [8], and thus, we assume that the departure or arrival of the PU follows a Poisson process and we have

$$\Pr\{J_0\} = \left[1-e^{-\lambda_d T}\right] \cdot \left[e^{-\lambda_d T}\right]^{J_0} \tag{8}$$

and

$$\Pr\{J_1\} = \left[1-e^{-\lambda_a T}\right] \cdot \left[e^{-\lambda_a T}\right]^{J_1}, \tag{9}$$

where λ_d, λ_a, and T represent the departure rate, arrival rate, and sampling interval, respectively. Using (8) and (9), finally, we obtain the unconditional test statistics

$$\begin{aligned} T_d &= \left|\sum_{J_0=0}^{N-1}[1-e^{-\lambda_d T}]\cdot[e^{-\lambda_d T}]^{J_0} \sum_{n=J_0+1}^{N} u[n]v^*[n]\right|^2 \\ &\quad \Big/ \left[\sum_{J_0=0}^{N-1}[1-e^{-\lambda_d T}]\cdot[e^{-\lambda_d T}]^{J_0} \sum_{n=J_0+1}^{N}|u[n]|^2\right] \\ &\quad \Big/ \left[\sum_{J_0=0}^{N-1}[1-e^{-\lambda_d T}]\cdot[e^{-\lambda_d T}]^{J_0} \sum_{n=J_0+1}^{N}|v[n]|^2\right] \\ &= \frac{\left|\sum_{n=1}^{N}[1-e^{-\lambda_d T n}]u[n]v^*[n]\right|^2}{\sum_{n=1}^{N}[1-e^{-\lambda_d T n}]|u[n]|^2 \sum_{n=1}^{N}[1-e^{-\lambda_d T n}]|v[n]|^2}. \end{aligned} \tag{10}$$

for the random departure of the PU, and similarly,

$$T_a = \frac{\left|\sum_{n=1}^{N}[1-e^{-\lambda_a T n}]u[n]v^*[n]\right|^2}{\sum_{n=1}^{N}[1-e^{-\lambda_a T n}]|u[n]|^2 \sum_{n=1}^{N}[1-e^{-\lambda_a T n}]|v[n]|^2}. \tag{11}$$

for the random arrival of the PU. Note that $T_d = T_a$ when $\lambda_d = \lambda_a$.

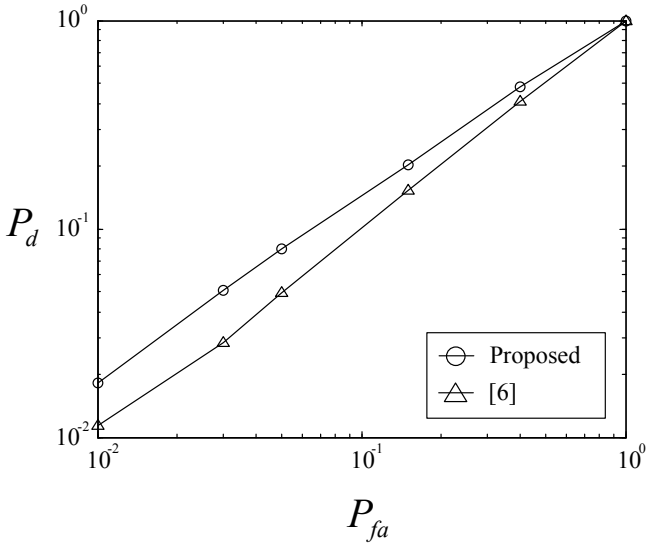


Fig. 2: ROC curves of the proposed and conventional schemes over AWGN channel in dynamic PU traffic environments at SNR $= -15$ dB.

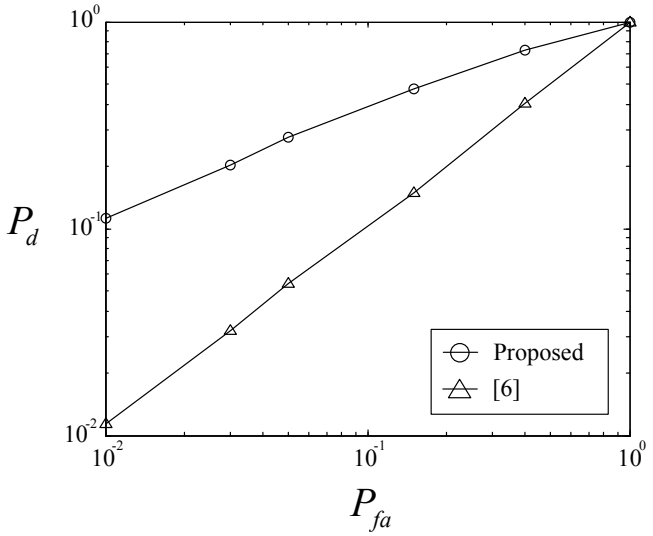


Fig. 3: ROC curves of the proposed and conventional schemes over AWGN channel in dynamic PU traffic environments at SNR $= -10$ dB.

4. Numerical Results

In this section, we compare the spectrum sensing performance of the proposed scheme with that of the conventional scheme of [6] in terms of the ROC. We assume the following parameters: $N = 100$, $\lambda_d T = \lambda_a T = 1$, $\alpha = 2f_c$, $P_{fa} = 0.01, 0.03, 0.05, 0.15, 0.4$, and 1, and a PU signal modulated by the binary phase shift keying with a carrier frequency f_c of 100 Hz. The threshold is determined from the false alarm probabilities given above, and it is assumed that one of the random departure and arrival is chosen randomly with equal probability.

Figures 2-5 show the ROC curves of the proposed and conventional schemes over an AWGN channel in dynamic PU traffic environments with the SNR values of -15 dB, -10 dB, -5 dB, and 0 dB, respectively, where P_d represents the detection probability defined as $\Pr(H_1|H_1)$. From the figures, it is clearly observed that the proposed scheme provides a significant improvement over the conventional scheme, and the improvement becomes more pronounced as the SNR increases. The conventional scheme achieves the worst case of the ROC performances at low SNRs such as -15 dB, -10 dB, and -5 dB since the energy detection approach used in the conventional scheme can scarcely distinguish the signal from the noise in such low SNR environments, whereas the cyclostationarity of the signal is easily distingushable regardless of the SNR value since the AWGN is not a cyclostationary process, and thus, the proposed scheme can generally provide a better ROC performance than the conventional scheme.

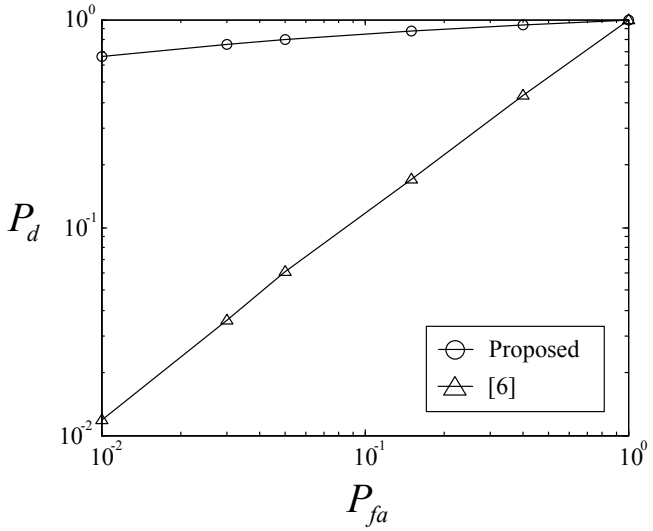


Fig. 4: ROC curves of the proposed and conventional schemes over AWGN channel in dynamic PU traffic environments at SNR $= -5$ dB.

5. Conclusions

In this paper, we have proposed a cyclostationarity-based spectrum sensing scheme for CRs in the presence of dynamic PU signals. We have first modeled the spectrum sensing problem in dynamic PU traffic environments as a binary hypothesis testing problem and developed the corresponding GLR. By applying an estimate of the squared magnitude of the SAF to the GLR, a test statistic for the proposed scheme is derived. Then, the spectrum sensing performance of the proposed scheme has been compared with that of the conventional scheme in several values of SNR. We have

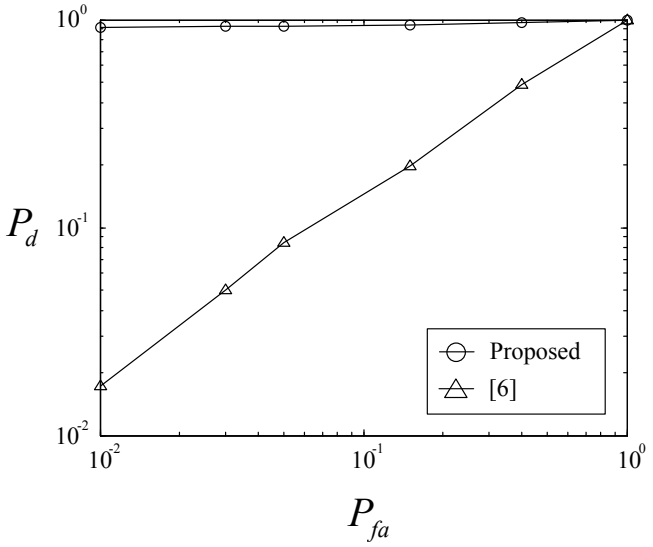


Fig. 5: ROC curves of the proposed and conventional schemes over AWGN channel in dynamic PU traffic environments at SNR $= 0$ dB.

observed that the proposed scheme provides a significant improvement over the conventional scheme in dynamic PU traffic environments.

6. Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2012R1A2A2A01045887 with funding from the Ministry of Science, ICT&Future Planning (MSIP), Korea, by the Information Technology Research Center (ITRC) program of the National IT Industry Promotion Agency under Grant NIPA-2013-H0301-13-1005 with funding from the MSIP, Korea, and by National GNSS Research Center program of Defense Acquisition Program Administration and Agency for Defense Development.

References

[1] J. Mitola, "Cognitive radio: an integrated agent architecture for software defined radio," in *Proc. Doctor of Technology*, Stockholm, Sweden, May 2000.

[2] J. Lunden, S. A. Kassam, and V. Koivunen, "Robust nonparametric cyclic correlation-based spectrum sensing for cognitive radio," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 38-52, Jan. 2010.

[3] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Proc. Asilomar Conf. Signals, Systems and Computers*, pp. 772-776, Pacific Grove, CA, Nov. 2004.

[4] T. S. Shehata and M. El-Tanany, "A novel adaptive structure of the energy detector applied to cognitive radio networks," in *Proc. Canadian Workshop on Information Theory*, pp. 95-98, Ottawa, Canada, May 2009.

[5] T. Wang, Y. Chen, E. L. Hines, and B. Zhao, "Analysis of effect of primary user traffic on spectrum sensing performance," in *Proc. Chinacom*, pp. 1-5, Xian, China, Aug. 2009.

[6] N. C. Beaulieu and Y. Chen, "Improved energy detectors for cognitive radios with randomly arriving or departing primary users," *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 867-870, Oct. 2010.

[7] W. A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Process. Magazine*, vol. 8, no. 2, pp. 14-36, Apr. 1991.

[8] P. Z. Peebles Jr., *Probability, Random Variables and Random Signal Principles*, 4th ed. New York: McGraw-Hill, 2001.

Spectrum Detection and Spectrum Decision in Cognitive Radio Networks Using Intelligent Mobile Agents

Nnenna Christine Eric-Nwonye and Obeten Obi Ekabua
Department of Computer Science
North-West University, Mafikeng Campus,
Private Bag X2046, Mmabatho 2735, South Africa
{23989696,obeten.ekabua}@nwu.ac.za

[ICWNô13]

***Abstract*--Cognitive radios are radios that improve spectrum efficiency and spectrum utilization by operating on unused spectrum channels in their neighborhood. These unused channels are detected through spectrum sensing, which must be performed to ensure the absence of the primary user, before a cognitive radio can utilize the channel. However, spectrum detection is a challenge to radios due to bandwidth constraints imposed on them and also due to degraded channel conditions such as multipath and shadowing. Secondly, these unused channels show different characteristics, and therefore, an appropriate channel needs to be chosen based on the characteristics it exhibits. In this paper, we propose algorithms and implementations for an Intelligent Mobile Agent-based approach for spectrum detection and decision, whereby, mobile agents are injected into the network to perform these two management functionalities, thus enabling the radios to utilize the channel without having to intermittently stop to check for the reappearance of the primary user and ensure non-interference to the primary user.**

Keywords--Cognitive radio; Intelligent mobile agent; Spectrum detection; Spectrum decision

1. INTRODUCTION

Cognitive radio is a technology that enhances efficient spectrum usage, by utilizing temporarily idle spectrum [1]. This technology was introduced because it was found out that most licensed spectrum are not being properly and fully utilized. Spectrum is a finite and costly resource which is managed by government agencies both nationally and internationally, and these agencies use a fixed allocation method in allocating spectrum to particular users and for a particular use [1][2]. Meaning that, once spectrum is allocated to a particular user, that user alone has the right to operate on that band. This is good because it simplifies issues concerning ownership and makes the license owner confident enough to invest in infrastructure, which ultimately leads to better quality of service.

However, studies by the federal communication commission (FCC) and other researchers [1][2][4][5], have shown that these licensed bands lie idle most of the time, leading to wastage of such a valuable resource. Thus, cognitive radios were introduced to utilize these idle channels but it has to be without interfering with the license owners.

To ensure non-interference to primary users (PU), cognitive radios or secondary users (SU) must sense the channel to detect the presence or absence of the primary user, and if absent, they can then make use of the spectrum. They also need to intermittently stop transmission to sense for the reappearance of the primary user. Spectrum detection, performed through spectrum sensing, is the key functionality to ensure efficient spectrum usage by cognitive radios.

It is however difficult for individual radios to reliably sense the occupancy state of the channel due to various channel degradation conditions such as multi-path fading and shadowing, leading to bad estimation of the occupancy state of the channel and false alarm. The dynamic nature of the radio spectrum therefore calls for the development of novel spectrum detection strategies to correctly estimate the occupancy state of a channel [5][6][7]. A feasible solution to this challenge faced by cognitive radio networks is to introduce agent based approaches for spectrum detection and decision.

Agents are programs that perform certain or specified tasks on behalf of the user. Mobile agents perform a userôs task by migrating and executing on several hosts connected to the network. The main difference between an intelligent agent and traditional agent is that the former perform not only actions pre-specified by a user but also those necessitated by later changes in the environment. Mobile agents introduce a new software and communication architecture, allowing a program to travel between machines for remote execution, even in heterogeneous cognitive radio networks. By transporting the agent code to the host machine in a distributed cognitive radio network, there is no need to bring intermediate signals and data across the network and thus a significant amount of network bandwidth use and communication delay can be avoided [8][9][10].

The detection accuracy in spectrum sensing has been considered as the most important factor to determine the performance of cognitive radio networks. Furthermore, idle spectrum bands in a network show different characteristics, so cognitive radios are supposed to select the proper spectrum band according to the application requirement

[5][6][7]. In this paper, we present agent-oriented algorithms and implementations for spectrum detection and decision in cognitive radio networks, to efficiently detect and decide on the availability and appropriateness of a channel.

2. RELATED WORK

This section outlines similar works that has been accomplished in the area of spectrum management in cognitive radio networks using agent technology.

The approach proposed by [11] is on using embedded agent modules, whereby, agents are embedded in the radio devices that coordinate their operations to benefit from network and avoid interference to the primary user. Agents carry a set of module to gather information about the terminal status and the radio environment and act accordingly to the constraints of the user application. The approach is based on agents with common interest who collaborate by sharing their knowledge and expertise to increase to increase their collective and individual gain. Group or coalition formation will identify the environmental information without requiring huge computational effort at the cognitive radio terminal. This helps in conserving the energy resources of autonomous cognitive radio terminal. Also keeping in mind the dynamic state of the radio resource and secondary users, the embedded agents in each cognitive radio device can build their preference models in terms of selecting a device with high level quality of service. These preference models will allow conserving time in dynamic and changing network conditions. In other words, the agents will know in which part of the network it can find information for making a proper decision

A very interesting approach is proposed in [12] where the authors have applied reinforcement learning RL on single-agent (SARL) and Multi-Agent (MARL) to achieve the sensitivity and the intelligence. They show in their results that the SARL and MARL perform a joint action that gives better performance across the network. They finally said reinforcement learning algorithm is adapted too be applied in most application schemas.

In the solution proposed in [13], a learning mechanism as the local MARL is available for each agent. The Local Learning provides a reward for each agent so that it can make the right decision and choose the best action. They modeled each SU node as a learning agent because the transmitter and receiver share a common result of learning or knowledge. The authors presented the LCPP (Locally Confined Payoff Propagation) which is an important function of reinforcement learning in MAS to achieve optimality in the cooperation between agents in a distributed CR network.

A channel selection scheme without negotiation is considered for multi-user and multi- channel in [14]. To avoid collision incurred by non-coordination, each SU learns to select channels based in their experiences. The MARL is applied in the context of Q-learning by considering the SUs as part of environment. In such a scheme, each SU senses channels and then selects a slowed frequency channel to transmit the data, as if no other SU exists. If two SUs choose the same channel for data transmission, they will collide with each other and the data packets cannot be decoded by the receiver. However, the SUs can try to learn how to avoid each other.

3. IMACRN SYSTEM MODULE FOR SPECTRUM RESOURCE MANAGEMENT

Basically, our system, Intelligent Mobile Agent for Cognitive Radio Networks (IMACRN), design is built on five different interlinked parts that form the working of our proposed system to take care of spectrum detection, spectrum decision, spectrum sharing and spectrum mobility. However, this paper deals with spectrum detection and decision only. Spectrum sharing and mobility will be explained in a subsequent paper. These agent parts are explained below:

a) Spectrum Sensing Agent (SSA): the function of SSA is to sense the radio spectrum holes and continuously monitor the primary user signals. SSAs cooperatively sense the channel and measure it against the threshold. Since it is not possible to know what time a spectrum band is occupied or when it is free, the sensing is done by considering a real-time dynamic environment. Factors that are taken into consideration include spectrum traffic, primary userôs signal power and associated noise and sampling time intervals.

b) Spectrum Decision Agent (SDA): SDAs characterizes the spectrum hole and its function is to arrange the idle spectrum information received through the SSAs according to channel capacity and channel information.

c) Secondary Consumer Agent (SCA): SCAs function is to send Spectrum Request (SR) messages to the Agreement Agent, whenever a secondary user indicates that it needs to use a portion of the spectrum. The message sent is of the form: req(s,t), where s represents the size of spectrum needed by the secondary user and depends on its application, for a duration of time t. SCAs also coordinate or share the spectrum amongst secondary users in the network after it has been acquired from the primary user.

d) Agent Memory Module (AMM): AMMs gets the primary userôs signal characterization from SDAs and stores in its database. This database list is regularly maintained and updated, thus it is not a permanent list. This module also serve as database for available spectrum and their characteristics.

e) Agreement Agent (AA): AAs manage the agreement and cooperation between primary and secondary users for spectrum sharing.

4. IMACRN-BASED SPECTRUM DETECTION

Dynamic Spectrum Access (DSA) by cognitive radio-enabled secondary devices is one of the promising approaches to increase utilization of underutilized licensed spectrum bands. However, DSA approach requires that the secondary users should not violate any acceptable interference bounds specified by the primary users. Therefore, the main challenge involved in devising DSA scheme for cognitive radio devices are as follows:

a) The cognitive radio nodes should be able to identify the white spaces in the spectrum and utilize them without interfering with the primary user.

b) The DSA scheme should minimize the channel sensing (and therefore, the energy consumed in the sensing operations) by the secondary node.

c) Cognitive radios cannot transmit while sensing the channel, which undermines the goal of DSA because spectrum is wasting and therefore spectrum efficiency is decreased.

In practical multi-user environments, cognitive radio operation is governed by interference tolerance and sensing limits at the primary and secondary users. The interference limits at the primary and secondary users indicate the amount of protection needed at each primary and secondary user from the multi-user interference to maintain a certain rate. On the other hand, the sensing limits (minimum SNR needed for detection) at the secondary users reflect the amount of protection that each secondary user is individually able to provide to the primary users. In these scenarios the key is to strike a balance between the two conflicting goals: minimizing the interference to the primary users, and maximizing the performance of the entire system.

To overcome these detection challenges, Intelligent Mobile Agents (IMAs) strategy is devised to help in reliably detecting idle channels, thus limiting the number of secondary users sensing the channel. The scheme ensures non interference to the primary user and radios can concentrate on transmission while the agents sense the channel for the reappearance of the primary user.

Spectrum Sensing Agents (SSA) in IMACRN are encoded to sense and detect primary user signals and send the information to the cognitive radio network. SSAs function is to sense the radio spectrum holes and continuously monitor the primary user signals. A predefined set threshold *Y*, is also input in the code for use in measuring the observed signal. This will help the radios to concentrate more on using the available spectrum without having to intermittently check for the presence of the primary user. The agents will individually or collaboratively detect active primary user transmissions over the band, and decide if the sensing results indicate that all the primary user transmitters are inactive at that band.

A. Spectrum Detection Algorithm

To buttress the points made, let us represent spectrum sensing by IMACRN agents using the algorithm below and the flowchart in Fig.1

let;
Y = a set threshold
Where
Y = Energy observed on the primary user signal
let;
s = result obtained from sensing the spectrum
if s > Y
then
H_1 = Primary User is Present
else
if s < Y
then
H_0 = Primary User is Absent.
// The set threshold Y will be used to determine and measure the reliability of the collected results. When the collected signal S*i* exceeds the threshold Y, decision 1 will be made which assumes that the primary user is present; otherwise, decision 0 will be made. The decision D*i* of the agents is then given by:
D*i* = 0; where 0 < S*i* < Y
and
D*i* = 1; where S*i* > Y

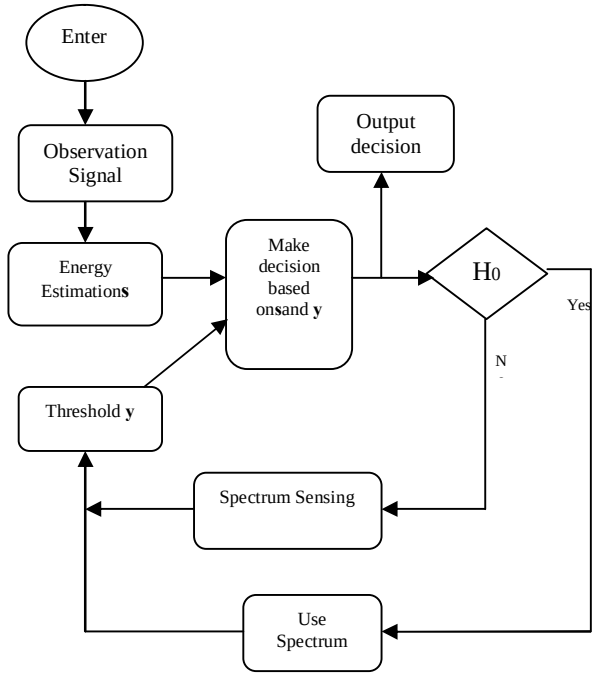


Fig. 1. Agent Spectrum Detection Flowchart

5. IMACRN-BASED SPECTRUM DECISION

The free spectrum bands detected through spectrum detection show different characteristics according to radio environment. Since cognitive radio networks can have multiple available spectrum bands having different channel characteristics, they should be capable of selecting the proper spectrum bands according to the application requirements, this is called Spectrum Decision. Spectrum Decision Agents (SDA) characterizes the available spectrum hole and each spectrum band is characterized based on not only local observations of cognitive radio users but also statistical information of primary networks. Through the local measurement, SDAs can estimate the channel conditions such as capacity, bit error rate (BER), delay and jitter. After the spectrum characterization, the best and appropriate spectrum band is chosen.

A. Spectrum Decision Algorithm

Mobile agents should select the best available channel by characterizing each spectrum hole based on Spectrum Band Information such as (operating frequency; bandwidth; interference level; channel error rate; path loss; link layer delay; wireless link errors; holding time) and Channel Conditions such as (capacity; bit error rate (BER); delay; jitter), and then make decisions D_1 (yes) or D_0(No). Depending on the application and the code parameters, two options are available in IMACRN for spectrum decision:

// **Option 1:** In IMACRN, channel can be characterized based on capacity C. this is calculated using Shannonôs theorem:

$$C = B\log_2(1 + SNR)$$

Where C = channel capacity
B = bandwidth of the channel
S = average signal power over the bandwidth
S/N = Signal-to-Noise Ratio (SNR)
[Secondary user agents characterizes each primary user on the basis of capacity]
For each i{i in PU} do
Evaluate (SNR(i))
[SNR: is the primary userôs signal to noise ratio obtained through SSA]
Evaluate (B(i))
[B: is the bandwidth of the primary user given by SSA]
$C(i) = B(i) \log_2 [1 + SNR(i)]$
[C: is the capacity calculated using Shannon theorem]
If acceptable
then D_1
Else D_0
End For

// **Option 2:** IMACRN characterizes spectrum based on the whole spectrum band information and channel capacity. Here the agent needs to check each parameter separately and then go to the next one. When all parameters have been evaluated and the result is satisfactory depending on the application that needs to use it, decision D_1 will be made, otherwise, decision D_0 will be made.

if s< Y
thenH_0
Check spectrum band information (b)
for b = 1 to 11
if
operating frequency; bandwidth; interference level; channel error rate; path loss; link layer delay; wireless link errors; holding time; bit error rate; delay; jitter
Acceptable
thenD_1
elseD_0

The flowchart depicting the above algorithm is given in Fig. 2:

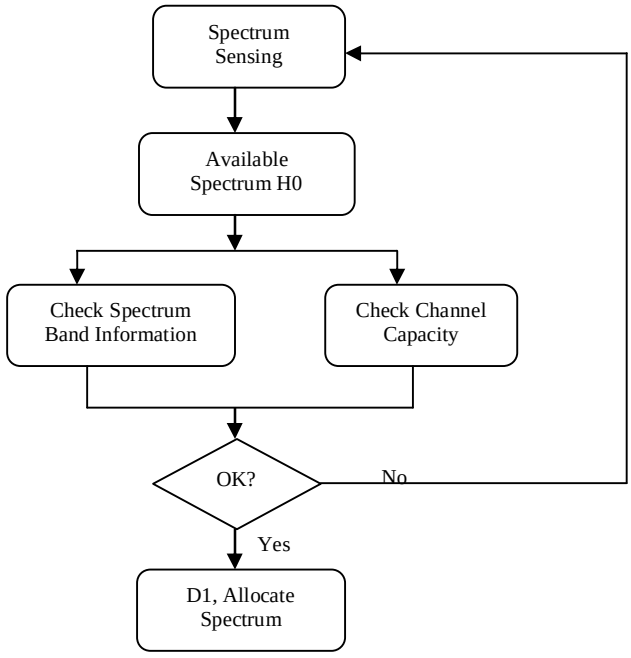


Fig.2. Agent Spectrum Decision

6. IMACRN AGENT CREATION

IMACRN agents are developed in Java Agent Development Environment (JADE), a software environment used for developing agents which comply with Foundation for Intelligent Physical Agents (FIPA) specifications. FIPA is a

non-profit international association that produces specifications and standards for agent technologies. IMACRN agents creation is achieved by defining a class extending the jade.core.agent class and implementing the setup() method. This method was used to create IMACRN agents as can be seen in Fig. 3.

A. IMACRN Agent Identifiers

An Agent Identifier (AID) is used to identify IMACRN agents and this identifier is an instance of the jade.core.AIDclass. Retrieving the agent identifier is made possible through the getAID() method of the agent class. IMACRN AIDs include a name and an address for each agent, and has the form <name>@<platform name>, the platform name being the address. As can be seen in figure 3, the created IMACRN agents are all living in the platform named NnennaC-HP, so as an example, the spectrum sensing agent (SSA) called SSA living in the platform have SSA@NnennaC-HP as its distinguished name.

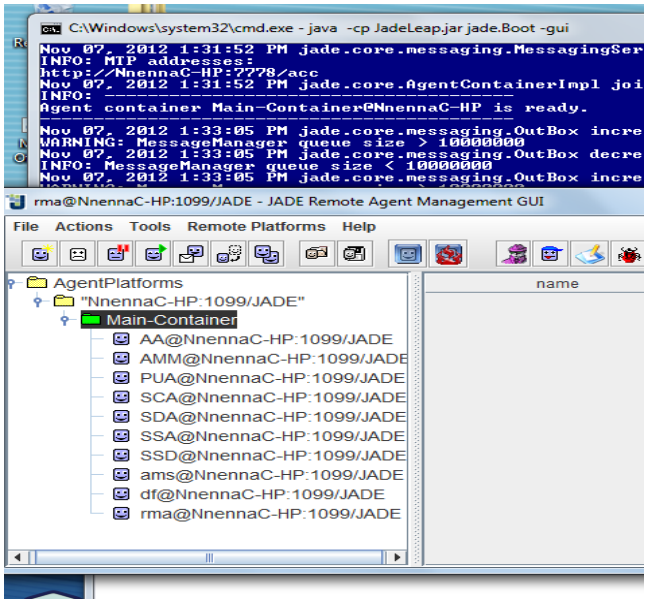


Fig. 3. Created Agents

B. Agent Communication

The communication method used by IMACRN agents is the asynchronous message passing. Using this method, each agent has a mail box where messages sent by other agents are posted by the JADE runtime. For message exchanges, IMACRN agents use Agent Communication Language (ACL) format defined and approved by FIPA for agent interactions [15][16]. The specified ACLs used by IMACRN agents include:

i) identity or name of the sender
ii) identity or name of the receiver(s)
iii) the purpose of the communication (known as performative), showing what the sender wants to achieve. For instance, in IMACRN:
a) if the sender requires the receiver to perform an action, the REQUEST performative is sent
b) if the sender wants to notify the receiver of a fact, the INFORM performative is sent
c) if the sender wants to know the truth of a given condition or statement, QUERY-IF performative is used
d) if the sender wants to initiate negotiation, the CFP (Call for Proposal) perfomative is sent
e) if the sender and receiver are negotiating, the PROPOSE,REJECT_PROPOSAL,ACCEPT_PROPOSAL

7. EXPERIMENTAL RESULTS

In this section, numerical results obtained and used to evaluate the multi-agent approach are presented. The simulation time is set at 120 minutes. All the simulations are conducted in Java Application Development Environment (JADE), over two PCs with 3.40GHz and 2.30GHz processor and 4GB memory. The parameters used are as shown in Table 1.

Table 1. Parameters for Experiment

Parameters	**Value**
Size of spectrum portion	4MHz
Simulation time	120 minutes
Max. number of PUs	30
Max. number of SUs	30
Max. number of each type of agent	5

A. Spectrum Detection

For spectrum detection to take place, two agents must be communicating through messages. The two agents in the experiment are Spectrum Sensing Agent (SSA) in Fig. 4(a) and Secondary Consumer Agent (SCA) in Fig. 4(b). The agents use a set threshold in their estimation for the presence or absence of the primary user. As can be seen in Figure 4(a) to Figure 4(c), the spectrum sensing agent and the secondary consumer agent communicated using messages.

For the purpose of this experiment and for communications to take place between agents, the threshold was set randomly at 3, and on sensing a signal at 1 as shown in message in Fig. 4(b), the SSA sent the information to the SCA indicating that the spectrum is available for use. And on receiving the message, the SCA acknowledged receipt by replying as shown in Fig. 4(c). The experiment was performed with a single and later with a multiple number of agents and the results obtained are reported in Table 2.

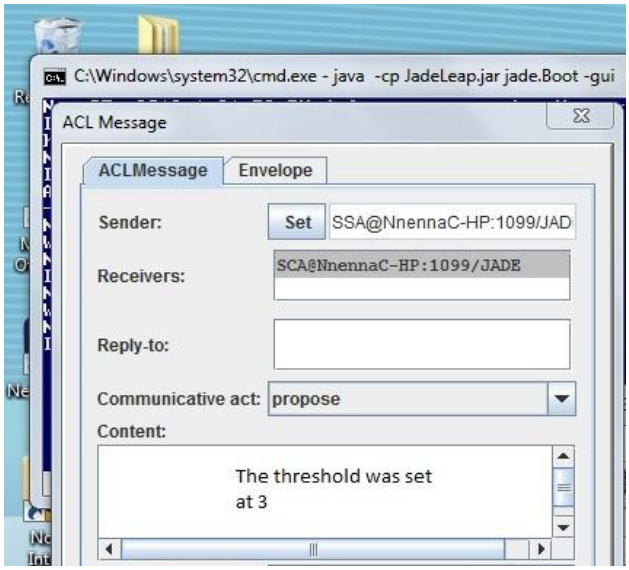


Fig. 4(a). Spectrum Detection Message from SSA

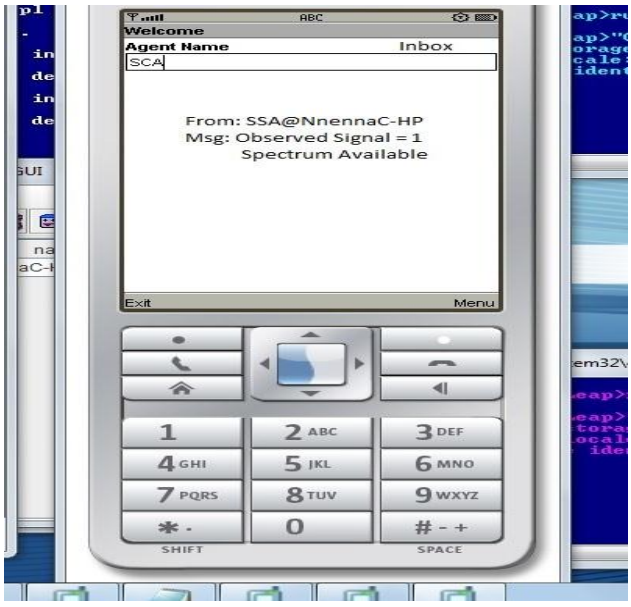


Fig 4(b). Message Received by SCA

Fig 4(c). Acknowledgement Message from SCA

From the results in Table 2, it can be shown that agents were able to correctly detect the occupancy state of the channel, when there are multiple numbers of agents.
With results shown in Table 2, the correlation between the number of agents and their corresponding number of channels correctly detected is graphically represented in Figure 5.

Table 2. Spectrum Detection Data

Number of Channels Sensed	Correct Detection (1 SSA)	Correct Detection (5 SSA)
5	3	5
10	6	8
15	10	13
20	13	17
25	20	23

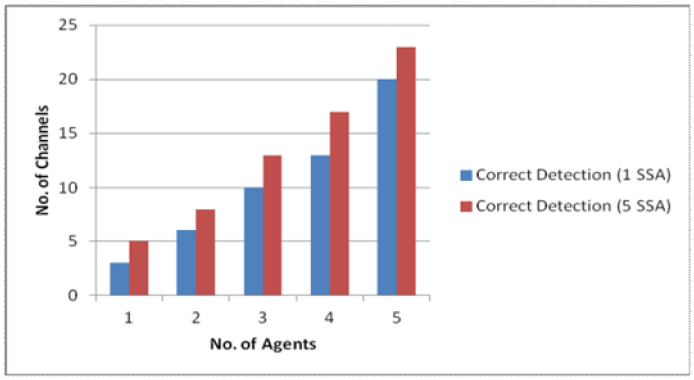


Fig. 5. Spectrum Detection Data

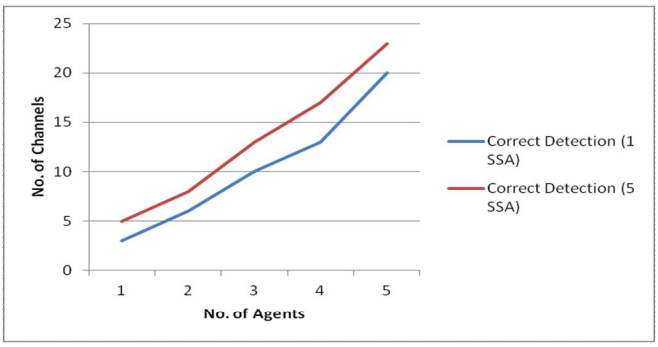


Fig. 5. Spectrum Detection Data

B. Spectrum Decision

The detected spectrum needs to be characterized to ensure that it is in good condition and appropriate for use. The agents were able to correctly decide on the appropriate spectrum band based on the characteristics of the available channels. Here, agents also use messages to pass information to each other. The results obtained based on a single agent decision and a multiple number of agents are as shown in Table 3.The SDA, after characterizing the channel, sent a message to the SCA indicating that the channel is usable, as shown in Figure 6.

Table 3. Spectrum Decision Data

Number of Available Channel	Correct Decision (1 SDA)	Correct Decision (5SDA)
5	3	5
10	6	8
15	10	14
20	13	20
25	20	24

Fig. 6. Spectrum Decision Message

Graphical representations of the correlation between the number of agents and decisions made correctly are shown in Fig. 7. From the results, it can be seen that multi-agents characterized the channels more correctly than a single agent. This is due to the cooperative nature of IMACRN agents in spectrum detection and decision. Each agent shares its detection and decision observation with other agents through messages, these observations are combined and decision made based on collective agreement.

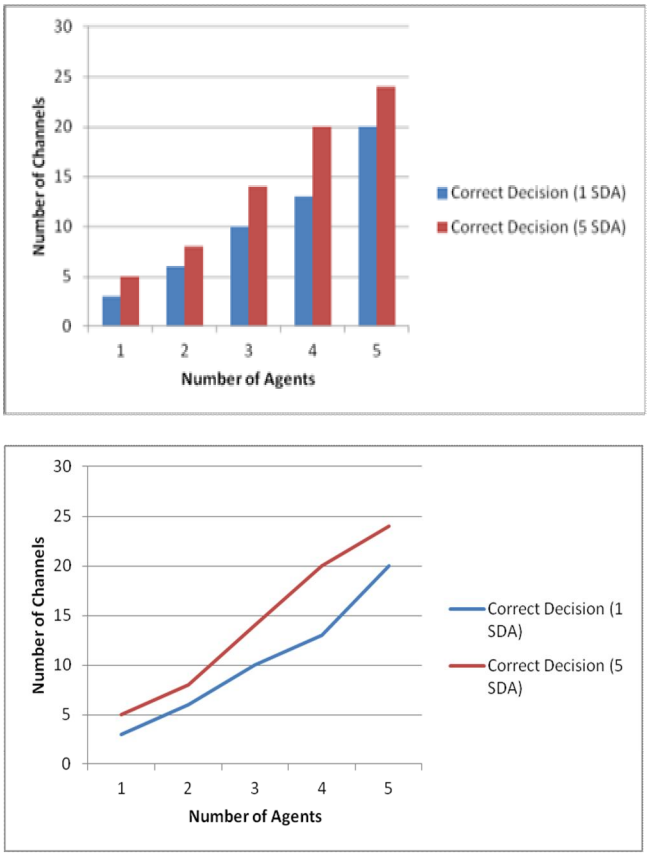


Fig. 7. Spectrum Decision Data

8. CONCLUSION

Spectrum detection and decision are important functionalities to realize dynamic spectrum access principles. Cognitive radios have to correctly estimate the primary userôs signal before utilizing the spectrum as incorrect estimation may lead to collision and false alarm. Presented in this paper are algorithmic approaches and implementation for intelligent mobile agent-based spectrum detection and decision. Our mobile agent system design is made up of five interlinked paths that take care of spectrum detection, spectrum decision, spectrum sharing and spectrum mobility, which are the spectrum management functionalities. The agents are injected into the network to

perform these functionalities for the radios in order to improve cognitive radio network performance. The approaches in this paper, for spectrum detection and decision, are step by step process which the agents use in carrying out their tasks. A set threshold is input into the mobile agent codes for use in detecting an idle or free spectrum and mobile agents evaluate the channel based on spectrum band information and channel capacity. Using the methodologies proposed, the primary userôs signal can be correctly detected and evaluated. A subsequent paper deals with algorithmic approaches and implementation for the two remaining spectrum management functionalities; spectrum sharing and spectrum mobility.

ACKNOWLEDGMENT

This research is sponsored by TELKOM Center of Excellence (CoE) of the department of Computer Science, North West University, South Africa.

REFERENCES

[1] NSF Workshop Report. Future Directions in Cognitive Radio Network (CRN) Research. March 2009.

[2] Federal Communication Commission (FCC). ñSpectrum Policy Task Force Reportò. Report ET Docket No 02ï135, November 2002..

[3] Federal Communication Commission (FCC). ñNotice of Proposed Rule Making and Orderò. ET Docket No 03 ï 322, December 2003.

[4] I.F. Akyildiz, W.Y. Lee, M.C. Vuran and S. Mohanty, ñA Survey on Spectrum Management in Cognitive Radio Networksò, *IEEE Communications Magazine*, Vol. 46, pp. 40-48, April 2008.

[5] I.F. Akyildiz, W.Y. Lee, M.C. Vuran and S. Mohanty, ñNext Generation / Dynamic Spectrum Access / Cognitive Radio Networks: A Surveyò. *Computer Networks Journal*, Vol. 50, pp. 2127-2159, September 2006.

[6] D.Cabric, S.M. Mishara and R.W. Brodersen. ñImplementation Issues in Spectrum Sensingò, *in Proc. 38th Annual Asilomar Conference on Signal, Systems and Computers.* Pacific Grove, CA, USA, November 2004. pp. 772 ï 776.

[7] J. Ma, G. Y Li. ñSignal Processing in Cognitive Radioò, *in proc. of the IEEE*, Volume 97, No 5, May 2009. pp. 805 ï 823.

[8] K. Hosoon., W.R.L. Gottfried and S. Baranitharan ñAn Intelligent Mobile Agent Framework for Distributed Network Managementò Network System Laboratory Telecommunication Research Center, Arizona State University. 2002.

[9] A. Bieszczad, Pagurek, B. Pagurek and T. White. ñMobile Agents for Network Managementò. IEEE Communications Surveys, Fourth Quarter 1998 Vol. 1 No. 1

[10] M. Genesereth, S. Ketckpel, ñSoftware Agentsò, In Communications of ACM, pp.48-53, 1994.

[11] A. Ahmed, M. M. Hassan, O. Sohaib, W. Hussain and M. Q. Khan. ñAn Agent Based Architecture for Cognitive Spectrum Managementò. *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 12, pp. 682-689, 2011.

[12] U. Mir, L. Merghem-Boulahia and D. Gaïti. ñA cooperative multi-agent based spectrum sharing,ò *in Proc. of the 6th Annual Advanced International Conference on Telecommunications*, Barcelona, 2010, pp. 150-165.

[13] G. Ana and L. Giupponi. ñAggregated Interference Control for Cognitive Radio Networks Based on Multi-agent Learningò. In proc. Of the 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communication (CROWNCOM), Hannover, June 22 ï 24, 2009.

[14] C. Wu, K. Chowdhury, M. D. Felice and W. Meleis. ñSpectrum Management of Cognitive Radio Using Multi-Agent Reinforcement Learningò, *in proc. of the 9th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2010)*, Toronto, Canada, May 10-14, 2010, pp. 1705-1712.

[15] FIPA.ñAgent Management Support for Mobility Specificationsò.Foundation for Intelligent Physical Agents. Geneva, Switzerland, pp. 10-18, 2000.

[16] FIPA. ñAbstract Architectural Specificationsò. Foundation for Intelligent Physical Agents. Geneva, Switzerland, pp. 25-38, 2002.

Study and Analysis of Basic Modulation Scheme in Wireless environment using USRP and Matlab (Simulink)

Amith Khandakar, Dr. Amr Mahmoud Salem Mohamed, Dr. Amr El Sherif
Qatar University

***Abstract*— USRP (Universal Software Radio Peripheral) provide comparatively inexpensive hardware platform for software radio and is used by research labs, universities. They connect to a host computer through a USB or Giga Ethernet link which the host-based software (Matlab) uses to control the USRP hardware and transmit/receive data. The main objective of the paper is to make a performance comparison of the basic modulation scheme (Quadrature Phase Shift Keying (QPSK) and Binary Phase Shift Keying (BPSK)) using the USRP .An analysis and study is done on the effect of Channel impairment (varying Frequency offset) and Channel Condition (Eb/No of the AWGN) on the transmission Quality (Bit Error Rate and Output Produced at the Receiver), using Matlab Simulink software interface. Then the AWGN is replaced with the real wireless environment using USRP and further study is done on the effect of varying Transmitter gain on the Transmission Quality for both the modulation schemes. Finally, the effect of an interfering USRP on the transmission quality is also analyzed.**

I. INTRODUCTION

THroughout the years, the demand for mobile connectivity has cause an exponential growth in wireless communications such as data communications, voice communications, video communications, etc. However the hardware –based approach to traditional radio design imposes a set of limitations in terms of hardware basic communication components (Frequency Offset Compensator hardware, Phase Offset Compensator hardware, Modulation Hardware, Demodulator hardware etc). Thus modifying radio devices easily and cost effectively has become business critical. Software defined radio(SDR) technology brings the flexibility, cost efficiency and power to drive communications forward, with wide –reaching benefits realized by service providers and product developers through to end users[1][2]

Software Defined Radio performs the task of transforming hardware aspects to software .System characteristics such as signal modulation scheme; operation frequencies, bandwidth etc are no longer dependent on analog circuits, whereas in SDR they rely on a system that integrates a programmable hardware and software that offers flexibility to modify those characteristics. [3]Therefore this kind of hardware with software interface can give the flexibility to change the parameters for certain operations which otherwise needed hardware replacement.

There has been many work showing the use of SDR in the field of research in Cognitive Radios[4] , OFDM modulation[5] and some work where they have signified the use of SDR in the field of education[1][3]. One thing noticeable in all the work above is that they have used a Software Defined Radio Platform USRP 2[6] and used GNU Radio as the software interface [7]. Another software interface provided by Matlab Simulink can be providing a different approach where inbuilt blocks and ready demos for certain common communication operations are readily available to be used [8]. It could be used and explored and that is what is done in this paper.

The paper provides a comparative study and analysis of the Basic Modulation Schemes BPSK and QPSK in terms of BER versus Channel Condition, Channel Impairment (In Simulink Model) and Transmitter power of the Software Defined Radio. The paper is arranged in the following manner: Section II gives the project description stating the components, problems tackled and the testing algorithm of the experiments, Section III provides the results of the experiments followed by Conclusion and Future Work in Section IV.

II. PROJECT DESCRIPTION

A. Components Used in the Study

The Software Defined Radio platform used in the experiment was USRP 2 which is supplied by Ettus Research [6]. The daughter boards used in the experiment was XCVR 2450 which operates at the frequencies 2.4 GHZ and 5 GHZ.

Matlab was used as the software interface to the USRP 2 and one of the Simulink Demos provided by Matlab for QPSK transmission system with Transmitter and Receiver was used as the starting point. [7]

B. Impairments to be tackled in the study

Any analog signal is represented in the following manner:

$$S(t) = A_t \sin(2\pi f_t t + \varphi_t)$$

Where A_t signifies the Amplitude of the signal ,ft signifies the frequency of the signal and φt signifies the phase of the signal [9]Thus the signal can undergo impairments in terms of Frequency , Phase or Timing during any communication session. The impairments can be termed as offset where it means the difference between the Original and received signal. [10].Modulation is the process of changing the attributes (Frequency, Phase, Amplitude) of the Carrier signal to represent the message signal .Changing the Phase of the carrier signal to represent the Message signal is known as Phase Modulation. Some of the well-known basic Digital Phase Modulation Techniques are Binary Phase Shift Keying (BPSK) and Quadrature Phase Shift keying (QPSK), which are also used in the study in the paper.

In Binary Phase Shift keying the phase of the analog signal is shifted to represent the Digital signal. In BPSK each symbol represents one bit and can be understood from the IQ diagram below:

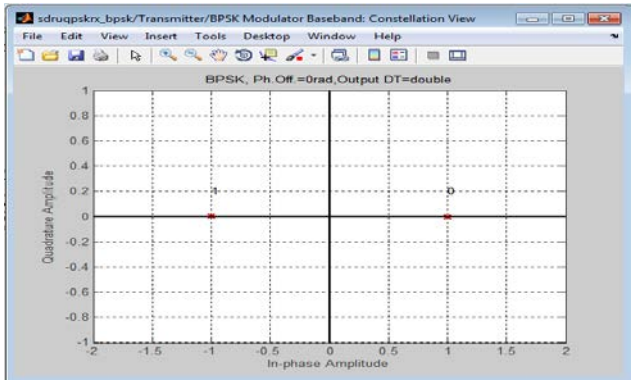


Figure 1: Constellation Diagram of BPSK System

In Quadrature Phase Shift keying the phase of the analog signal is shifted to represent the Digital signal. In QPSK each symbol represents two bit and can be understood from the IQ diagram below(the constellation diagram representation below is known as Gray mapping Constellation diagram where the adjacent representation are obtained by change of a single bit:

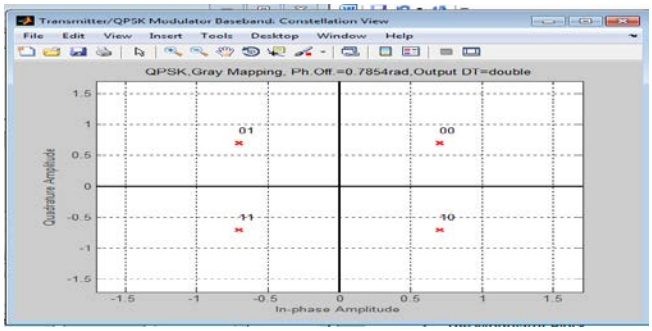


Figure 2: Constellation Diagram of QPSK system

In Phase Modulation there is another impairment which has to be tackled in addition to the 3 general ones discussed above : Phase Ambiguity , as if due to phase shift , and if the phase shift is more than -45 and +45 in QPSK , there phase shift is not detectable and it will lead to an error(refer the constellation diagram in Figure 2 , where if phase shift is – 45 for symbol 00, it shifts and will be detected as 10 and +45 shift will lead it to be detected as 01) similarly in BPSK it is – 90 and +90 phase shift leading to phase ambiguity.

Thus any receiver should have components to compensate for the Frequency offset, Phase Offset, Timing Offset and Phase ambiguity.

The Simulink demo provided by Matlab provides a Receiver model which has sub blocks dealing with all the above mentioned errors, and it is this Simulink model which is used in the study.

C. The Method of the Experiments conducted

There are 4 study (A, B, C and D in Section III) done and this section describes the algorithm and setup of each study.

Setup 1 for Section A and B in Section III

The Simulink Demo model (QPSK Transmitter and Receiver) [8] is used to develop the BPSK Transmitter and receiver for conducting the experiment.

First the Transmitter and Receiver is used in the same Simulink model and the channel parameter i.e. frequency offset and Channel Condition (Eb/No) is varied to see the effect on transmission which is depicted by Bit Error Rate.

As seen in Figure 3, the BPSK transmit data through the simulate channel, which is received, compensated for impairments and the BER (Bit Error Rate) is checked which is objectively checked with the output produced (Text streaming, shown in Figure 6)

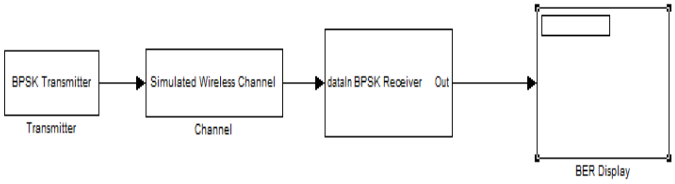


Figure3: Simulink Model for Comparison A and B in Section III, using BPSK Modulation scheme for study

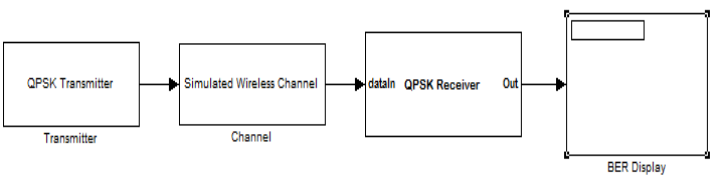


Figure4: Simulink Model Comparison A and B in Section III, using QPSK Modulation scheme for study

The above Simulink models (Figure 3 and 4) are used for the Comparison of A and B in Section III.

The Change in the Frequency offset of the channel is done and also the Channel Condition is changed, the Figure 9 can give a better idea on how the change is conducted for the experiment.

A script is run to change the Frequency Offset and the Channel Gain in the Simulink model shown above(Figure 3 and 4) , and the BER is noted at the end and graph is generated using the information gather which is shown later, refer Figure5

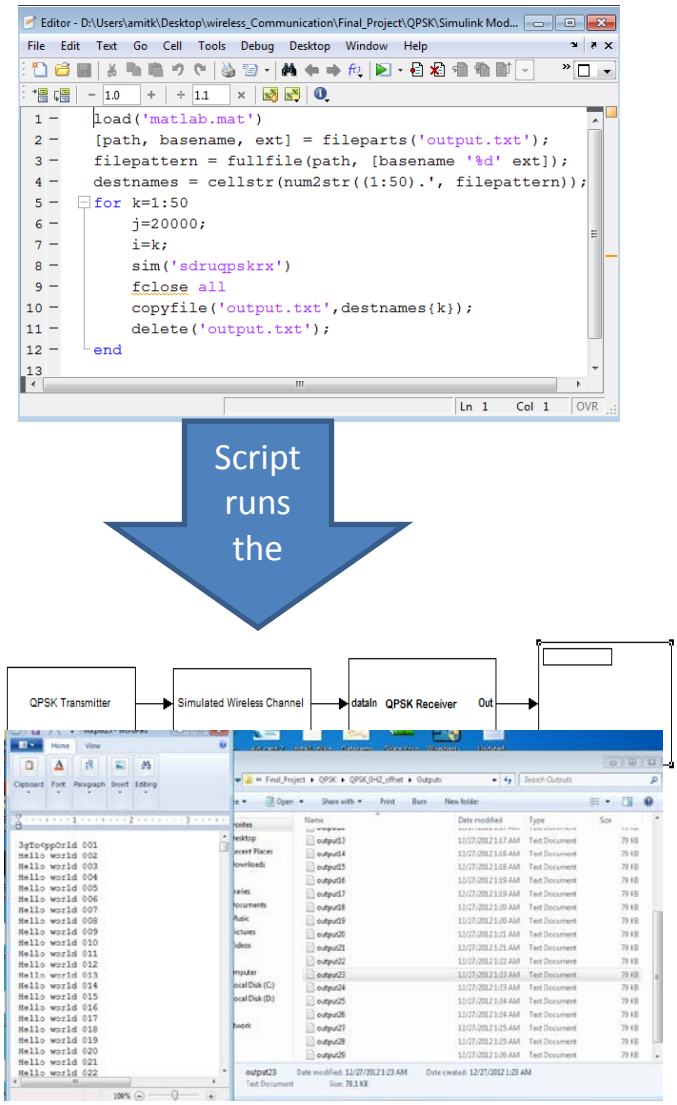


Figure 5 Algorithms for Comparison A and B in Section III

Figure 6: Output (Text Streaming) Generated

The BER information gather is checked with the output produced (i.e. Text streaming, Figure 6) at the receiver to have a subjective (BER) and Objective (Output quality) on how is the Channel condition related with the BER.

Setup 2 for Section C and D in Section III:

The other experiments are done using the USRP instead of the simulated wireless channel, refer figure 10 .The experimental setup is as below:

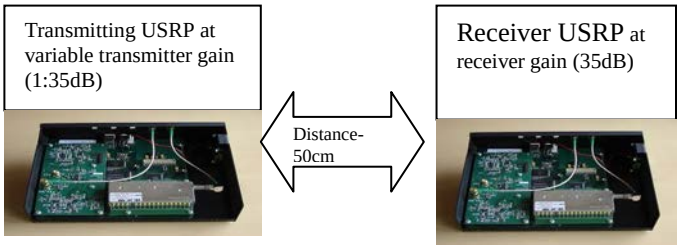


Figure 7: Experimental setup for C in section III

A script is again run to change the parameter such as the transmitter gain and then the effect is seen on the BER at the Receiver, keeping the distance constant between the 2 USRP's, The experiments is done using QPSK modulation scheme and then followed by the BPSK modulation scheme(refer figure 7)

The final experiment was done to check the effect of the interfering USRP on the transmission quality and the experimental setup is shown below in Figure 8.

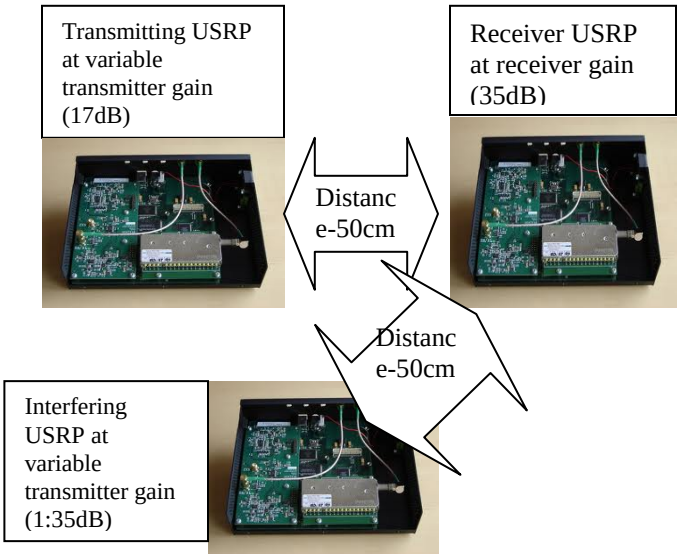


Figure 8: Experimental setup for D in section III

A script is again run to change the parameter such as the gain of the interferer and then the effect is seen on the BER at the Receiver, keeping the distance constant between them The experiments is done using QPSK modulation scheme The experiment setup is as shown in figure 8

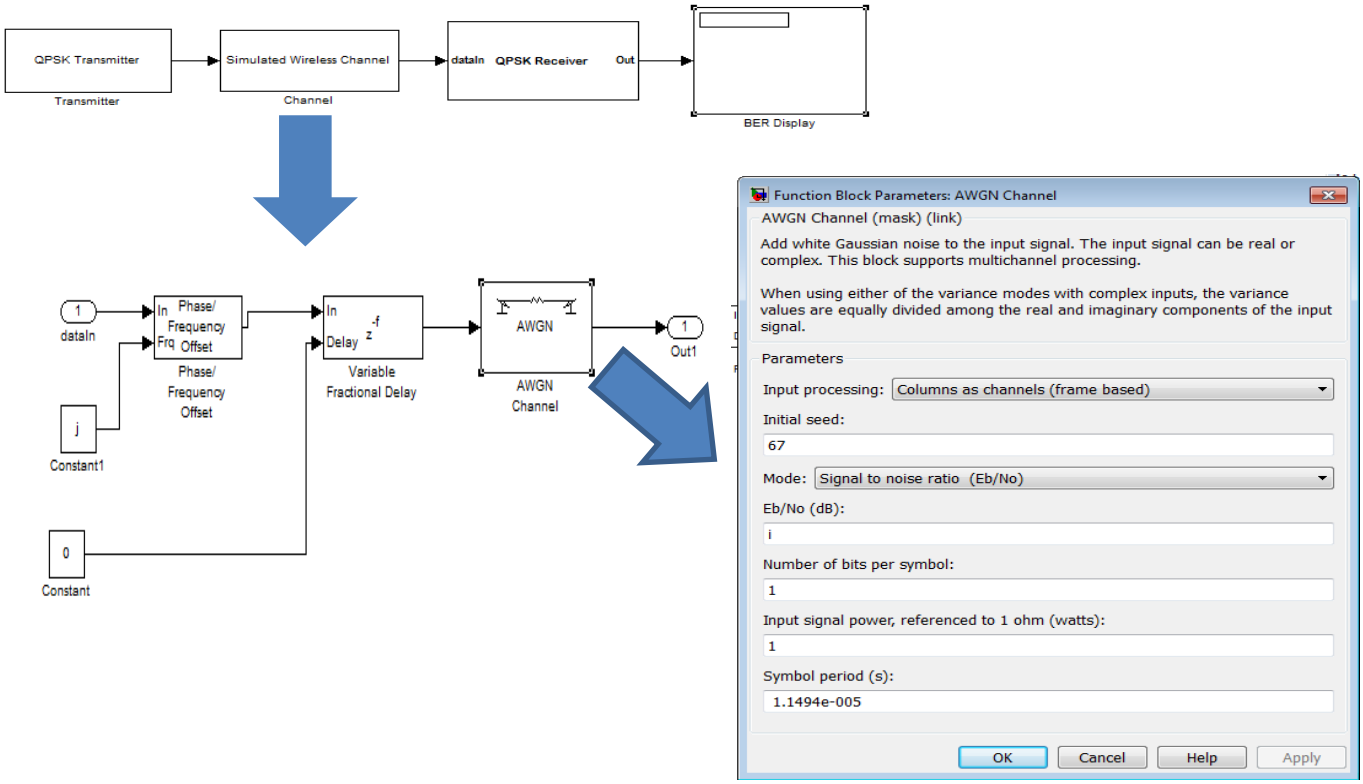


Figure 9: Changing the Phase/Frequency Offset and Channel Gain in the Simulated Wireless Channel

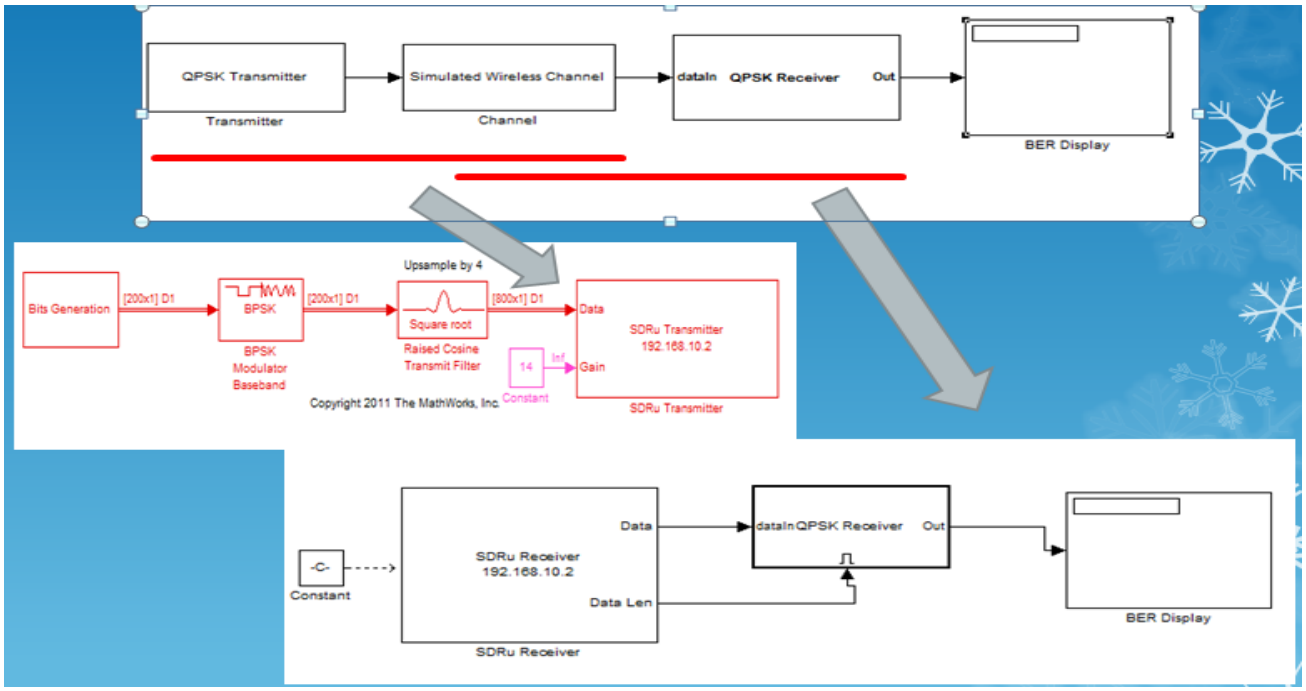


Figure 10: USRP used in place of the Simulated Wireless Channel

III EXPERIMENT ALGORITHM AND RESULTS

Based on the subjective and objective comparison it can be concluded that for the experimental setup 0.44 was a very good BER, which is referred in the later part of the paper. It is also to be noted that the worse BER that a system can have is 0.50 as the probability of error for 1 being 0 is ½ or 0.50

A. Comparison of the BER with Respect to Impairments such as Frequency offset for QPSK and BPSK

Please refer setup 1 in Section II
Bit Error Rate is calculated as follows:
BER=Number of Error Bits/ Total Number of Bits Received

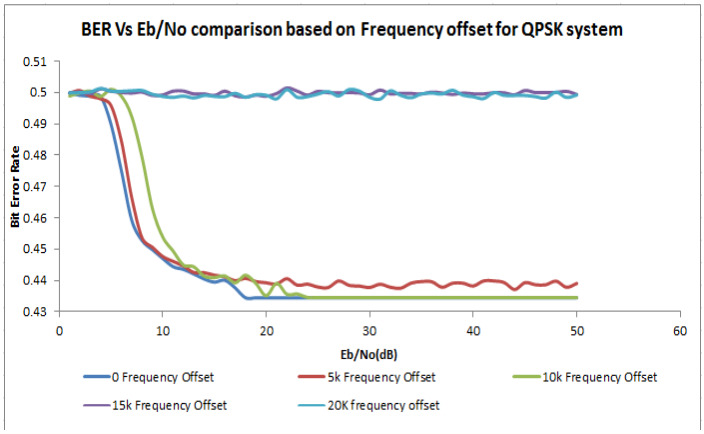


Figure 11: BER vs. Eb/No comparison based on Frequency offset for QPSK system

As it can seen from the above graph the QPSK system breaks down (Worst BER performance 0.50 being prevalent despite the increase of Channel gain) at 15KHZ and 20KHZ. Thus frequency offset more than 15KHZ is not compensated by the Receiver model. As expected the BER reduces with the increase in the Eb/No (dB) of the Channel. Furthermore the BER is comparable at frequency offset up to 15KHZ as the Simulink receiver model compensates for it.

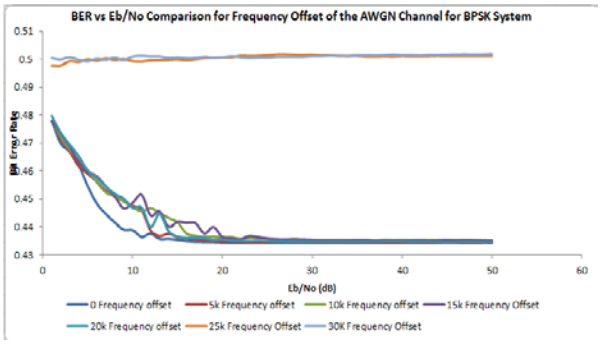


Figure 12: BER vs. Eb/No comparison based on Frequency offset for BPSK system

As it can seen from the above graph the BPSK system breaks down (Worst BER performance 0.50 being prevalent despite the increase of Channel gain) at 25KHZ and 30KHZ. Thus frequency offset more than 25KHZ is not compensated by the Receiver model. As expected the BER reduces with the increase in the Eb/No (dB) of the Channel. Furthermore the BER is comparable at frequency offset up to 25KHZ as the Simulink receiver model compensates for it.

The Break down Frequency Offset is double the QPSK system which is expected as the phase shift caused by the frequency is allowable double in the case of BPSK than QPSK (Refer the Constellation Diagram figure 1 and 2)

B. Study and Analysis of BER with Eb/No for QPSK and BPSK system using a Simulink Model (Without USRP)

Theoretically the BER for QPSK and BPSK system is almost similar and this is proved mathematically below:

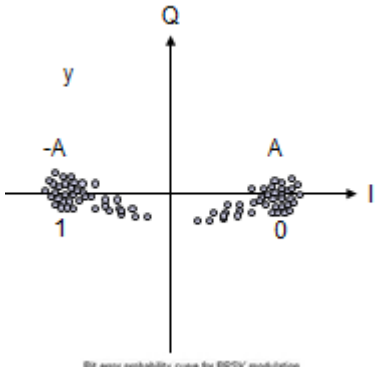


Figure 13: Signal representation of BPSK on IQ diagram

Any Received Signal,

$$y=x+n,\ x \in \{-A, A\}$$

Where,

$$n=\mathbb{N}(0, \sigma^2=N_0/2)$$

$$\mathcal{PDF}_{(n)} = \frac{1}{\sqrt{2\pi N_{0/2}}} e^{-\frac{x^2}{2N_{0/2}}}$$

$$P_e = P(n > A) = \frac{1}{\sqrt{2\pi N_{0/2}}} \int_A^{\infty} e^{-\frac{x^2}{2N_{0/2}}}$$

$$BER = P_e = P(n > A) = \varphi(\sqrt{2E_0/N_0})$$

[9]

Similarly it can be proved that

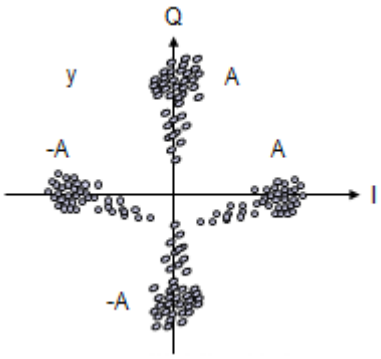


Figure 14: Signal representation of QPSK on IQ diagram

$$BER = 4 * (\frac{1}{4}) P(n > A) = \varphi(\sqrt{2E_0/N_0})$$

Thus theoretically BPSK and QPSK have similar BER

Please refer setup 1 for experimental setup in Section II

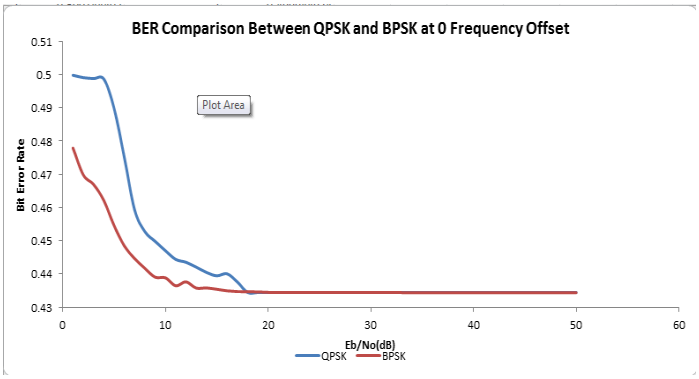


Figure 15: BER Comparison between QPSK and BPSK at 0 Frequency offset

It is also evident from the graph above that BPSK and QPSK have almost comparable BER but BPSK have a throughput half of QPSK .

C. Comparison of BER with respect to the Varying Transmitting power of the USRPs keeping the distance constant (for both QPSK and BPSK system)

Please refer to Setup 2 in Section II

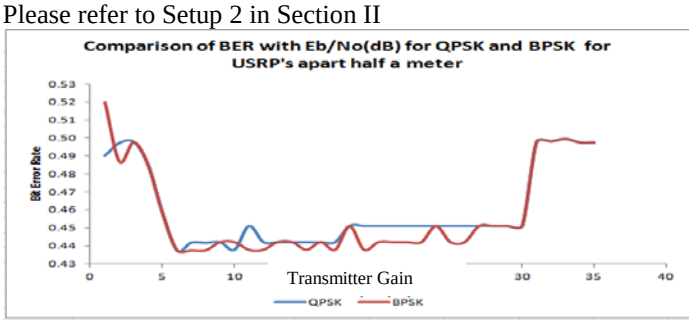


Figure 16: Comparison of BER with Eb/No (dB) for QPSK and BPSK for USRP's apart half a meter

As it is evident from the graph above that BPSK and QPSK have almost similar BER throughout.

As the Transmitter gain is increasing the BER gets better as expected as more the transmitter gain the more proper transmission signal strength is the further it can travel. But as the Transmitter gain reaches a maximum around 30dB the BER degrades and keeps degrading reaching the Maximum. This is due to the fact that the transmitter gain has reached a linear circuitry of the USRP Transmitter where the excess gain leads to increase in temperature of the circuitry and leads to transmitting of wrong transmission instead of the transmission given to it. Thus for a successful transmission an optimal transmitter power level is to be maintained.

The Throughput of the BPSK system was half of the QPSK system (where QPSK system had 3000 frames whereas BPSK system had 1500 frames)

D. Study the effect of interference of another USRP transmitting at various power levels.

Please refer to setup 2 in Section II

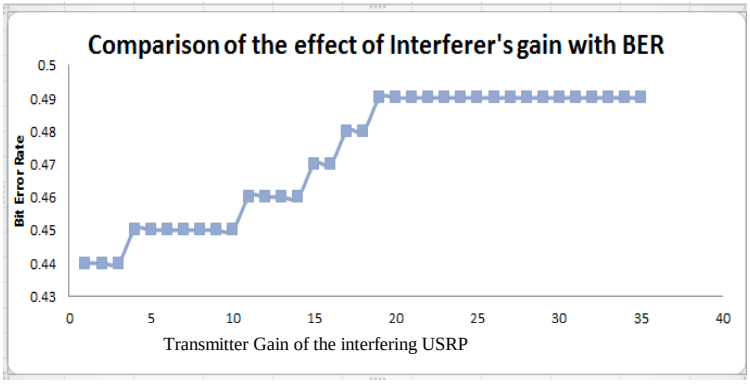


Figure 17: Comparison of the effect of Interferer's gain with BER

As it is seen from the graph above, as the transmitter gain of the interfering USRP increases the BER degrades as it interferes with the Transmission between the other 2 USRP's, but as the gain reaches 17dB and above the BER degrades and becomes worse as both the transmitting usrp's are transmitting at the same power and at the same frequency leading to collision and drop of packets.

IV CONCLUSION AND FUTURE WORK

. The paper have done experiments using USRP and Matlab Simulink as the software interface to do a study and analysis on the effect of transmitting power , interfering transmitting power on the BER of a transmission using BPSK and QPSK modulation scheme. The work has used the demo Simulink Matlab model of QPSK transmitter and Receiver and developed BPSK transmitter and Receiver. It has also developed a Simulink model with simulated Wireless AWGN channel to study the effect of Channel Impairments (i.e.

Frequency Offset) on the BER of a transmission. The work has successfully conducted experiments on real wireless channel and effectively generated expected results and has explained them. It has effectively proved the efficiency of SDR where changes in the modulation scheme were done via software and no hardware modifications were done.

The future work of this paper could be doing use other modulation schemes like GMSK, M-QAM etc. Try to develop a Bidirectional Time Division Duplex Communication system and also use other application (Image Streaming) instead of the Text streaming...

ACKNOWLEDGMENT

The work would be incomplete without paying acknowledgement to Dr. Amr Mahmoud Salem Mohamed from Qatar University and Dr. Amr El-Sherif from Qatar University Research Department, who have introduced the idea to the author and also have provided all the basic knowledge needed to conduct the experiments

REFERENCES

[1] Z.Tong ,M.S.Arifianto and C.F. Liau "Wireless Transmission using universal Software Radio Peripheral" in Proceeding of the 2009 International Conference on Space Science and Communication .

[2] SDR Forum- Software Defined Radio Forum, "Introduction to SDR" Citing Internet sources http://www.sdrforum.org .

[3] Andre L.G.Reis, Andre F.B. Selva , karlo G. Lenzi Silvio E. Barbin and Luis G.P.Meloni "Software Defined Radio on Digital Communications: a New Teaching Tool" Wireless and Microwave Technology Conference (WAMICON), 2012 IEEE 13th Annual

[4] Z. Yan, Z. Ma, H. Cao, G. Li, and W. Wang, "Spectrum sensing, access and coexistence testbed for cognitive radio using USRP", 4th IEEE International Conference on Circuits and Systems for Communications, 2008. ICCSC 2008, 26-28 May 2008.

[5] Marwanto, M. A. Sarijari, N. Fisal, S. K. S. Yusof and R. A. Rashid,"Experimental study of OFDM implementation utilizing GNU Radio and USRP - SDR," Communications (MICC), 2009 IEEE 9th Malaysia International Conference, 15-17 Dec. 2009.

[6] M. Ettus, "Universal Software Radio Peripheral". [Online]. Available: http://www.ettus.com.

[7] GNU Radio, "GNU Radio Companion - usage tips". [Online]. Available:http://gnuradio.org/redmine/projects/gnuradio/wiki/GNURadioCompanion#Usage-Tips

[8] http://www.mathworks.com/discovery/sdr/usrp.html

[9] Lecture Slides of Dr. Amr, in class Wireless Communication CMPT 543, Fall 2012

[10] Digital Communications-A Discrete –Time Approach ,By Michael Rice, Brigham Young University

Livestock Disease Control System using RFID and WSN

Seokkyun Jeong[1], Taewoong Choi[2], Hyun Yoe[*]
[1,*]Dept. of Information and Communication Engineering, Sunchon National University, Suncheon, Jeollanam-do, Republic of Korea
[2]Farm Innovation Research & Development Center, Suncheon, Jeollanam-do, Republic of Korea
sk_jeong@sunchon.ac.kr, turbocab@paran.com, yhyun@sunchon.ac.kr

Abstract - *This thesis collects stock information, cattle shed's facility information and video information through RFID and video treatment devices according to a livestock farmhouse with a RFID/WSN based livestock disease control system in order to prevent spread of stock diseases that are taking sick recently in the country. In addition, as the spread of stock diseases can occur through movement of vehicles, this thesis is tracing movement of feed, milk collection and human waste vehicles by utilizing RFID/GPS and manages the respective information integrated through middleware. And the system was composed to enable quick handling through monitoring and information analysis at a monitoring system for preventing spread of stock diseases in case of generation of stock diseases or abnormality of a stock farmhouse's facility.*

Keywords: Livestock, Diseases, RFID, WSN

1 Introduction

The importance of systematic and smart management of livestock disease prevention system is increasing in recent for livestock and animal disease control and welfare enhancement. Contagious livestock diseases are affecting the safety of not only animals but also people across the world, and they could lead to national crisis situation. This could cause not only financial damages but also the mental damages to livestock and related personnel[1][2].

There are foot-and-mouth disease(FMD) and AI as typical malignant livestock's contagious disease, among them, the FMD having the most serious destructive power takes sick at artiodactyl animals(animals that its hooves are divided into two pieces like cow, pig, sheep, goat and deer, etc.), and it is a disease that livestock gets to be seriously ill or dead because of rapid rise of its body temperature, generation of blisters at the mouth, tongue, hooves or nipple, etc. and decline of appetite[3].

Though the mad cow disease occurred in the country in 2010, the damage amounting to 7 times of 450.3 billion won has occurred due to contagious disease of livestock for 4 years since 2006 by failing in its early treatment, so the amount of the damage estimated by the Government reached a scale of around 3 trillion won. This is primary loss according to loss of livestock, so if considering social/environmental costs, its damage is expected to be larger[4][5]. Accordingly, in order to minimize spreading in case of generation of livestock disease and enable to respond to livestock disease, a system capable of monitoring circumstances of livestock disease spreading nationally is being required[6][7], but a system capable of responding it is currently in the non-existent state. This thesis has proposed a system capable of quickly responding in case of in case of generation of livestock disease by monitoring people (herdsmen, veterinary surgeon, insemination technician, etc.) inside an infected region which is a propagation path of livestock disease, vehicles (feed car, shipment car and milk collection car, etc.), etc. through utilizing IT technology such as RFID, WSN and GPS, etc. and collecting environmental information inside stock shed. The composition of this thesis is as follows.

In the Chapter 2, the detailed function of the system proposed in this thesis is examined, and in the Chapter 3, the service provided by this system is explained. In the Chapter 4, the implementation results of the proposed system are confirmed, and in the last Chapter 5, this thesis aims to complete conclusions of this thesis.

2 Design of the Proposed Livestock Disease Control System

This system automatically records information of a vehicle visiting a farm to prevent spreading of livestock disease by using RFID and gets to enable quick response in case of generation of livestock disease by tracing the location of a vehicle in real time through GPS. In addition, the system was composed to be able to monitor stock shed's environment in real time according to a farm and to be able to execute remote medical treatment in real time, if necessary. This Chapter describes the structure and components of a proposed monitoring system for prevention of livestock disease spreading.

* Corresponding Author

2.1 System Structure

The monitoring structure for prevention of livestock disease spreading is composed of 3 kinds of layers. There are a physical layer collecting information of a vehicle and sensor collecting temperature/humidity of cattle shed, an intermediate layer taking charge of communication between the physical layer and applied layer and an applied layer that provides information and service capable of monitoring circumstances between a vehicle and farmhouse in real time. Figure1 shows a monitoring system diagram for prevention of livestock disease spreading.

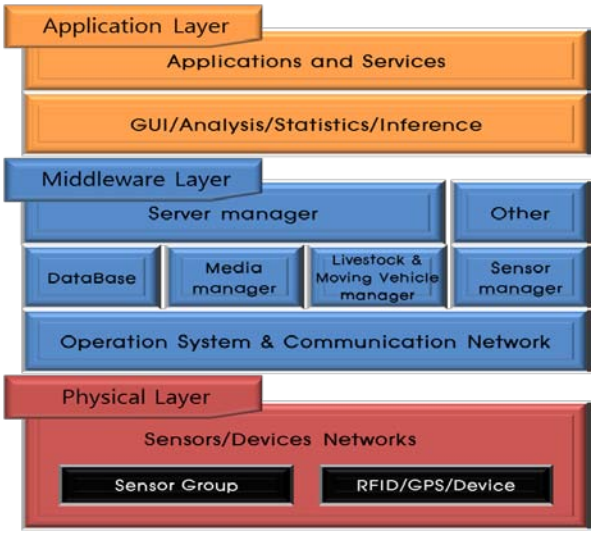


Figure 1. System Structure of the Proposed System

2.2 Service Process

2.2.1 Physics Layer

The physics layer is being consisted of a real-time information collection unit that collects information having an important influence on movement of vehicles and livestock in order to monitor movement of vehicles and livestock in real time, and a sensor unit consisted of sensors collecting environmental information of livestock at cattle shed. The real-time information collection unit attaches RFID tag and GPS at a feed vehicle, milk collection vehicle and night-soil vehicle to monitor movement of a vehicle, confirms movement of the vehicle, and stores information of the vehicle in the integrated information DB system through the Internet, when a vehicle visited a specific farm. The movement information of livestock monitors all processes from a cattle shed for raising livestock by using RFID to consumers, and gets to be stored in the integrated information DB system by generating information of livestock movement.

The senor unit collects information of temperature, humidity and gas of a cattle shed, and each sensor transmits collected data to a sink node in a certain time interval. A camera makes a state of livestock that is difficult to judge by only numerical data collected from a sensor to be confirmed through video by transmitting video information inside a cattle shed to a server via stream data.

2.2.2 Middleware Layer

The middleware layer takes charge of communication between the physical and applied layers, and is consisting of a database, video manager, sensor manager, a manager on livestock and vehicle movement and server manager. First of all, the database is being constructed to enable to store movement information, etc. of vehicles and livestock and is storing environmental information on optimal environment data and livestock disease generation's conditions in growing of livestock. A sensor manager processes raw data collected from a sensor unit and stores it in the database after verifying necessary information. A video manager converts an analog video signal photographed by a camera into a digital video signal, and then stores it in the database. A manager on livestock and vehicle movement treats vehicle location information collected through GPS in real time and stores arrival time and waiting time, etc. of a vehicle in the database. Lastly, a server manager controls all managers inside a server, such as a sensor manager, video manager and database, etc. If a specific service is asked at an applied program, the server manager provides the service by treating data through a corresponding manager.

2.2.3 Application Layer

The application layer is a system providing information on vehicle and livestock movement in real time to allow a user to perform rapid and early treatment in case of generation of livestock disease at any place where the internet is connected. For this, this system supports construction of WEB-based monitoring user interface, provision of various statistics and prediction information through information analysis, provision of livestock disease information and preventive methods, information provision of cattle shed status by region and provision of livestock disease's remote treatment service, etc.

3 Service Process of the Proposed Livestock Disease Control System

3.1 Livestock and Vehicle Movement Record Service

Confirming details about whether any vehicle visited a livestock farmhouse and how long it stayed becomes an important index that measures disease spreading in case of livestock disease generation. This service leaves a record of movement management of vehicle and livestock inside cattle shed through RFID and GPS and records the movement details with a daily record. Through this, in case of livestock disease generation, managers can quickly confirm whether the disease was flown in and spread through any path, and can take measures on it.

3.2 Livestock Farmhouse Environmental Information Monitoring Service

The livestock farmhouse environmental information monitoring service is a monitoring service that can confirm optimal environmental maintenance and disease generation existence in raising livestock by providing environmental data and video data of each livestock farmhouse. Users can confirm temperature, humidity and harmful gas concentration which are environmental elements having the largest influence on livestock growth and development in real time through a terminal. The driving process of this service stores data collected from sensors and camera in the database through each manager. The corresponding information of stored data is provided to a terminal of users though a server manager. This service makes users be able to confirm livestock environment in real time through this process.

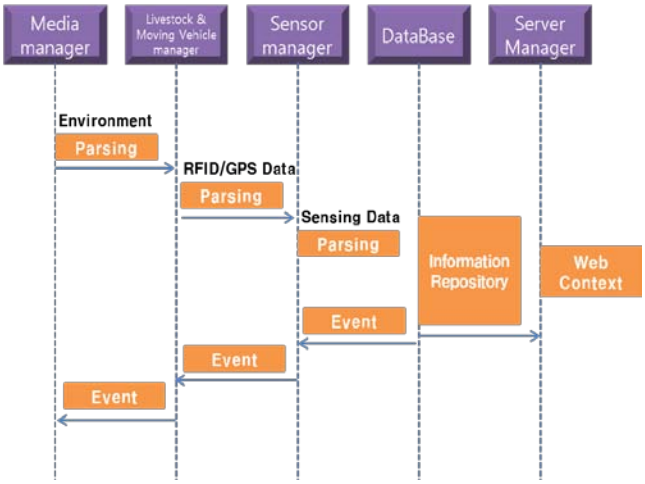


Figure 2. Livestock Farmhouse Environmental Information Monitoring Service Process

3.3 Livestock Management Export Service

The livestock management export service is a service that collects advices for forming optimal growth and development environment through analysis of experts on accumulated information through a monitoring system for prevention of livestock disease spreading and that can diagnose through analysis of experts on the livestock growth and development state based on camera video information even without visiting the field. Figure 3 is a service procedure of the livestock management export service. If users ask advices of an expert, the advice request information is delivered to an expert server located in the outside, and the expert gets to again deliver its content to users by analyzing this. In case of generation of an emergency event, the remote diagnosis and emergency treatment can be executed within fast time by using a camera. If an expert utilizes cattle shed's environment database accumulated through a monitoring system for prevention of livestock disease spreading, it can be used as important data capable of preventing livestock disease spreading.

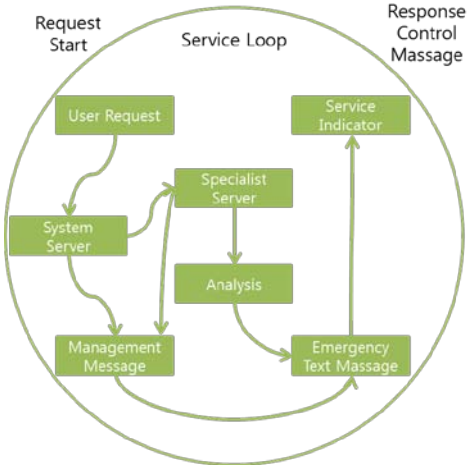


Figure 3. Service Procedure of the Livestock Management Export Service

3.4 Mobile Letter Message Service

The mobile letter message service is a service that sends letter message immediately to farmhouses in terms of events due to absence or emergency situation of farmhouse owners or thief invasion, etc. in case of livestock disease generation. This is a motion procedure of mobile letter message service.

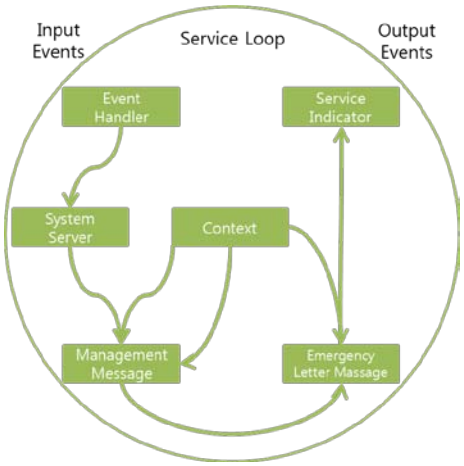


Figure 4. Service procedure of mobile letter message service

4 Implementation of Proposed Livestock Disease Control System

4.1 Implementation Environment

In order to verify performance of this system, a physical environment was composed like Figure 5 in an actual cattle shed.

Figure 5. RFID Reader and GPS

In order to grasp a movement path of a vehicle, Figure 5 GPS and RFID are installed to a system like Figure 6, and the portability of a vehicle was grasped by installing RFID like Figure 6.

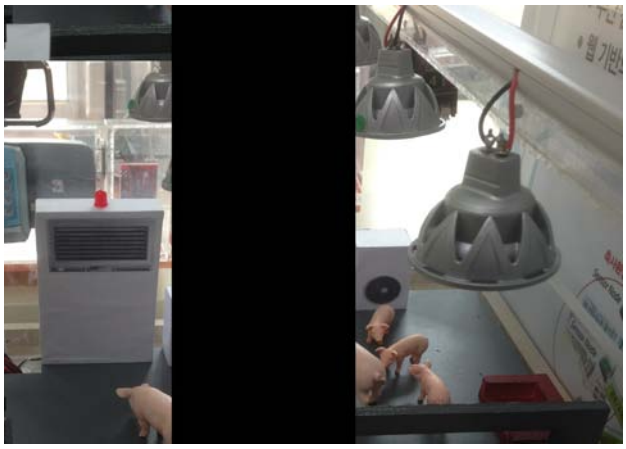


Figure 6. The simple model of livestock for environment information

For cattle shed's environmental data collection, the collection of temperature/humidity data and monitoring service at a remote area were carried out to be executed by installing a sensor at an actual cattle shed like Figure 6, and the information necessary for prevention of livestock disease spreading like environmental information for optimal growth or livestock disease information is stored in the database of the monitoring system for prevention of livestock disease spreading, and makes users be able to confirm livestock and vehicle information in real time by storing sensing information, video information and vehicle & livestock movement information.

4.2 Implementation Results

The performance results of this system were confirmed through GUI. Figure 7 is GUI made to be able to monitor movement of livestock & vehicle and cattle shed's environment. The cattle shed's real-time environment can be monitored according to a region shown in a map like ①, and the livestock disease information that can occur through monitored information from ② was provided. In addition, the expert counseling service is provided through a video camera in an remote area in case of suspecting livestock disease in ③. Lastly, the kind of a moving vehicle and livestock movement that moves from a farm ④ could be confirmed in real time through GUI.

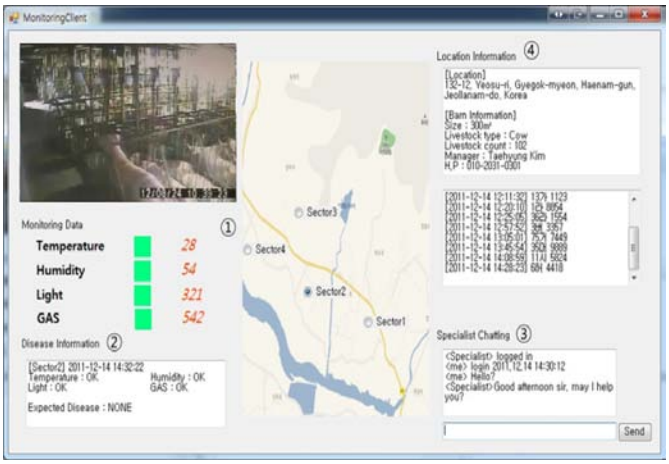


Figure 7. The Proposed Systme GUI

5 Conclusions

This thesis proposed a RFID/WSN based livestock disease control system to reduce a damaged scale of the livestock industry due to malignant livestock's contagious disease with fast propagation speed. For this, this thesis has realized a system capable of confirming disease information within a cattle shed, livestock and vehicle movement situations in real time by making a database on cattle shed's environment information collected according to each farmhouse, and in order to verify performance of the system, a sensor and camera were installed in a cattle shed, and GPS and RFID were installed in a vehicle, and then, the proposed system was applied and driven. Through this research, it is expected that this research will nationally minimize damage of livestock disease spreading by minimizing spreading in case of livestock disease generation and responding livestock disease through quick and early treatment. The future research is about to construct a livestock disease prediction system based on collected data through a RFID/WSN based livestock disease control system spreading and to construct a community capable of easily exchanging know-how that can cope with in case of livestock disease generation.

6 Acknowledgment

This work (Grants No. 2012-0391) was supported by Business for Academic-industrial Cooperative establishments funded Korea Small and Medium Business Administration in 2012 .

7 References

[1] Situngkir, Hoky. Epidemiology through Cellular Automata, Case of Study: Avian Influenza in Indonesia. Department Computational Sociology, Bandung Fe Institute. (2007)

[2] Hong Seung Gil, Lee Nam, Yong." Conceptual Design of Digital Animal Defense System," Soongsil University (2009)

[3] Al Ayubi, S.U., "A framework of spreading disease monitoring system model", Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), 2009 International Conference on. 2009, p1-5 (2009)

[4] Kim Min Gyeong, "Causes foot-and-mouth disease, which could be affected?", GS & J Institute, Special Focus GSnJ, Article No. 94, 2010.2, p1-10 (2010)

[5] United States Government, "Livestock Market Reporting: USDA Has Taken Some Steps to Ensure Quality, But Additional Efforts Are Needed: Report to Congressional Requesters", Books LLC, (2011)

[6] Kim Sun Joong, "South Korea's poultry disease status", Korean Society of Poultry Science, PROCEEDINGS Spring Symposium on Cooperative , (2011)

[7] Jong-Hyeon Park , Su-Mi Kim, Kwang-Nyeong Lee, Young-Joon Ko, Hyang-Sim Lee, In-Soo Cho, "Strategy for Novel Vaccine and Antivirals Against Foot-and-Mouth Disease", JOURNAL OF BACTERIOLOGY AND VIROLOGY, Article No. 40 VOL.1, p1-10 (2010)

A DS-UWB Radar System Using a Short-Length PN Sequence

Youngseok Lee, Jeongyoon Shim, Youngpo Lee, Jaewoo Lee, and Seokho Yoon†
College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do, Korea
†Corresponding author

Abstract— *In this paper, we propose a direct sequence ultra wideband (DS-UWB) radar system using a short-length pseudo noise (PN) sequence. The proposed DS-UWB radar reduces the correlation processing time by averaging out the noise in the correlator outputs, and thus, employs a short-length PN sequence unlike the conventional DS-UWB radar systems using multiple variable-length PN sequences. Numerical results demonstrate that the proposed DS-UWB radar estimates the distance between the radar and an object with a shorter correlation processing time while providing a better estimation performance compared with the conventional DS-UWB radar systems.*

Keywords: DS-UWB; radar; distance measuring; PN sequence

1. Introduction

In direct sequence ultra wideband (DS-UWB) radar systems, a distance between the radar and an object is estimated by transmitting a DS-UWB signal, and then, estimating the time delay until the reflected DS-UWB signal returns to the radar. Due to its high time resolution, the DS-UWB radar has attracted much interest as a distance estimation unit in vehicular parking assistance systems, vehicular pre-crash sensing systems, and security sensors [1]-[3]. As the length of the pseudo noise (PN) sequence used in the DS-UWB radar becomes larger, a higher correlation gain can be achieved, resulting in a more reliable distance estimate. However, a long PN sequence calls for a large amount of correlation processing time [4], and thus, it is not suitable for time-sensitive applications such as vehicular pre-crash sensing systems where a fast distance estimation is required to avoid a collision between a vehicle and an obstacle.

Thus, DS-UWB radar schemes using variable-length PN sequences [5]-[7] have been proposed, where multiple PN sequences with various lengths are employed instead of a single long PN sequence and the PN sequence length is selected based on the standard deviation of the estimated distances, thus reducing the correlation processing time while keeping the estimation performance. Nonetheless, the schemes still have a long correlation processing time at low signal-to-noise ratios (SNRs) since the standard deviation becomes larger at low SNRs, and thus, long-length sequences are required to maintain the performance.

In this paper, we propose a novel DS-UWB radar system using a short PN sequence. Since the fluctuated noise in each correlation is independent, we average out the noise by accumulating the correlator outputs, and thus, obviate the need of long-length sequences at low SNRs, allowing us to employ a single short-length PN sequence in the overall SNR range of practical interest, and eventually, to reduce the overall correlation processing time.

2. Proposed DS-UWB Radar System

Figure 1(a) shows the structure of the DS-UWB radar system for estimating the distance D between the radar and an object. First, the DS-UWB signal

$$s(t) = \sqrt{E_c} \sum_{j=0}^{N-1} p_j g(t - jT_c) \tag{1}$$

is transmitted to the object, where E_c is the chip energy of the PN sequence with a length of N chips, $p_j \in \{-1, +1\}$ is the jth chip, T_c is the chip duration, and

$$g(t) = \left[1 - 4\pi \left(\frac{t - T_c/2}{\gamma}\right)^2\right] \exp\left[-2\pi \left(\frac{t - T_c/2}{\gamma}\right)^2\right] \tag{2}$$

is a UWB pulse with unit energy over $[0, T_c)$, where the time normalization factor γ is set to $\sqrt{4\pi}/7T_c$, making 99.99% of the total waveform energy included within the chip duration [8]. The transmitted signal is reflected by the object and returns to the receiver with a delay τ, and thus, the received signal $r(t)$ is obtained as

$$r(t) = s(t - \tau) + w(t), \tag{3}$$

where $w(t)$ is an additive white Gaussian noise process with mean zero and double-sided power spectral density $N_0/2$. Subsequently, the received signal is correlated with a reference signal with a candidate delay $\tilde{\tau} \in \Delta = \{0,\ T_c,\ 2T_c,\ \cdots,\ (N-1)T_c\}$, yielding the mth correlator output

$$\begin{aligned} R_m(\tau, \tilde{\tau}) &= \int_{(m-1)NT_c}^{mNT_c} r(t)s(t - \tilde{\tau})dt \\ &= E_c S(\tau, \tilde{\tau}) + W_m \end{aligned} \tag{4}$$

for $m = 1, 2, \cdots, M$, where M is the number of correlations,

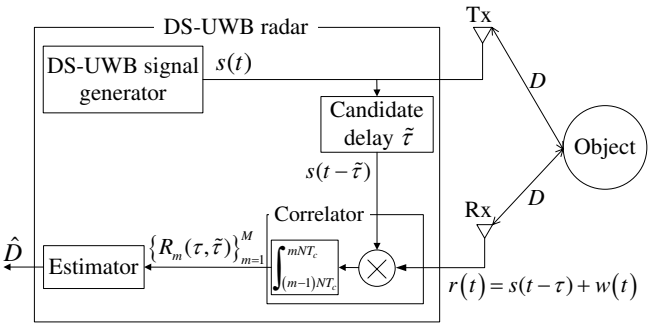
$$S(\tau,\tilde{\tau}) = \begin{cases} N\left[1 - 4\pi\left(\frac{\tilde{\tau}-\tau}{\gamma}\right)^2 + \frac{4\pi^2}{3}\left(\frac{\tilde{\tau}-\tau}{\gamma}\right)^4\right] e^{-\pi(\frac{\tilde{\tau}-\tau}{\gamma})^2} & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } |\tilde{\tau}-\tau| \le T_c, \\ -1 \qquad\qquad\qquad\qquad\qquad\qquad \text{for } |\tilde{\tau}-\tau| > T_c \end{cases} \tag{5}$$

is the autocorrelation function of $g(t)$ [9], and $\{W_m\}_{m=1}^M$ are zero-mean independent identically distributed Gaussian random variables with variance $NT_cN_0/2$. Finally, the estimator yields a distance estimate $\hat{D}$ using $\{R_m(\tau,\tilde{\tau})\}_{m=1}^M$ as shown in Figure 1(b): In the conventional estimators, $\{R_m(\tau,\tilde{\tau})\}_{m=1}^M$ are exploited individually, i.e., M delay estimates $\{\hat{\tau}_m\}_{m=1}^M$ (and consequently, M distance estimates $\{\hat{d}_m\}_{m=1}^M = \{c\hat{\tau}_m/2\}_{m=1}^M$, where c is the speed of light) are individually obtained per correlator output and the final distance estimate $\hat{D}_c$ is obtained by averaging $\{\hat{d}_m\}_{m=1}^M$. In addition, the standard deviation of $\{\hat{d}_m\}_{m=1}^M$ is yielded for selection of the PN sequence length. In the proposed estimator, on the other hand, a single delay estimate $\hat{\tau}$ and the corresponding distance estimate $\hat{D}_p$ are made through $R(\tau,\tilde{\tau}) = \frac{1}{M}\sum_{m=1}^M R_m(\tau,\tilde{\tau})$ obtained by accumulating M correlator outputs. It is easy to see that the noise variance of $R(\tau,\tilde{\tau})$ is reduced by a factor of M than that of $R_m(\tau,\tilde{\tau})$, and thus, it is expected that the proposed DS-UWB radar performs better than the conventional radars when a PN sequence with the same length is employed.

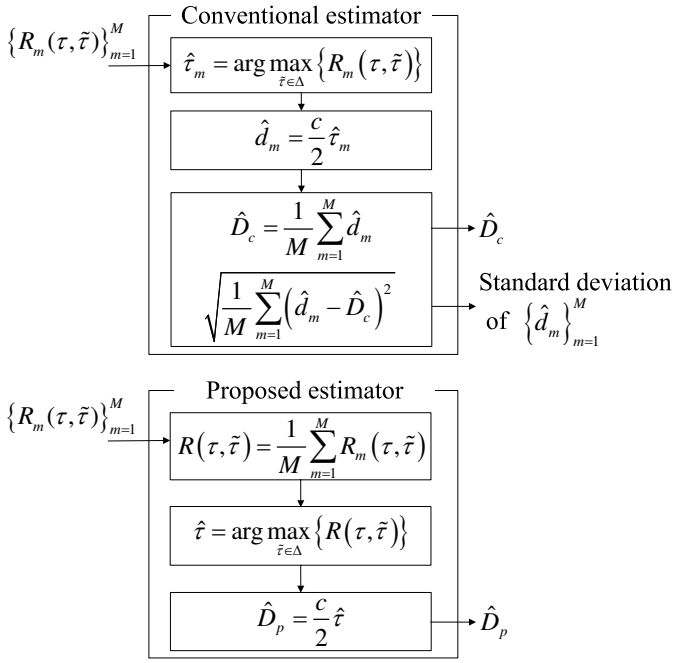
3. Numerical Results

In this section, the proposed and conventional DS-UWB radars are compared in terms of the average correlation processing time (ACPT), which is the time required to obtain an estimate $\hat{D}$ on the average, and root mean square error (RMSE), i.e., $\sqrt{E[(D-\hat{D})^2]}$, where $E[\cdot]$ denotes the statistical expectation. For simulations, we assume the following parameters: $D = 30$ meters (the maximum detectable range for short range radars [10]), $M = 100$, $N = 7$, 15, 31, 63, and 127 chips. For the best performance of the conventional DS-UWB radar, we have first numerically obtained and then employed the optimum value of the standard deviation of $\{\hat{d}_m\}_{m=1}^M$.

Figures 2 and 3 show the ACPT in chips and RMSE of the proposed and conventional DS-UWB radars as a function of E_c/N_0, respectively, where the ACPT of the proposed DS-UWB radar is constant regardless of the SNR values since it uses a single PN sequence unlike the conventional one using variable-length PN sequences. From the figures, it is observed that the proposed radar employing a single PN sequence with a length of 15 provides a shorter ACPT and at the same time a better RMSE performance over the conventional radar using multiple PN sequences including a long PN sequence with a length of 127 chips in the SNR range -20 $\sim$ 0 dB of practical interest. This is because the proposed radar can offer a more reliable distance estimate only with a single short PN sequences by averaging out the noise effect through the accumulation of the correlations, unlike the conventional radar exploiting the correlations individually.



(a) The structure of the DS-UWB radar



(b) The estimator structures of the conventional and proposed DS-UWB radars

Fig. 1: The structure of the distance measuring unit of the conventional and proposed DS-UWB radars

In passing, we would like to stress that the value of M is set to be the same for both the proposed and conventional DS-UWB radars in simulations, which implies that the number of observations is the same for both the radars, and thus, the radar using a longer PN sequence exhibits a longer ACPT.

Table 1 shows the computational complexity in estimating the distance for the proposed and conventional DS-UWB radars, where a flop is defined as a real floating point

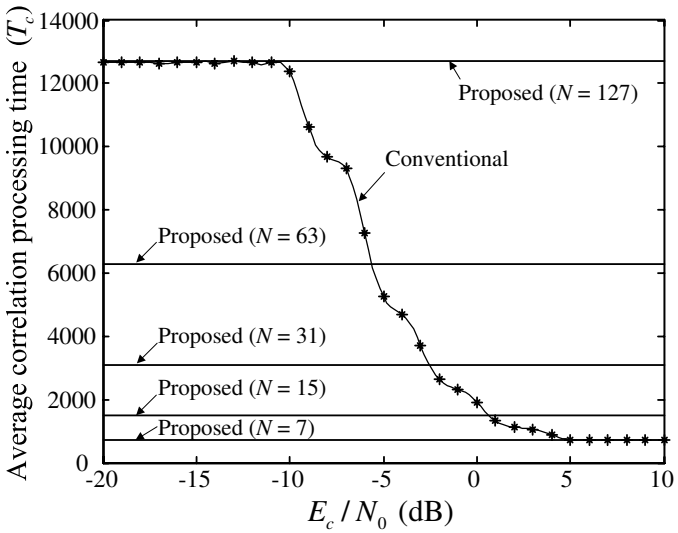


Fig. 2: The average correlation processing time of the proposed and conventional schemes.

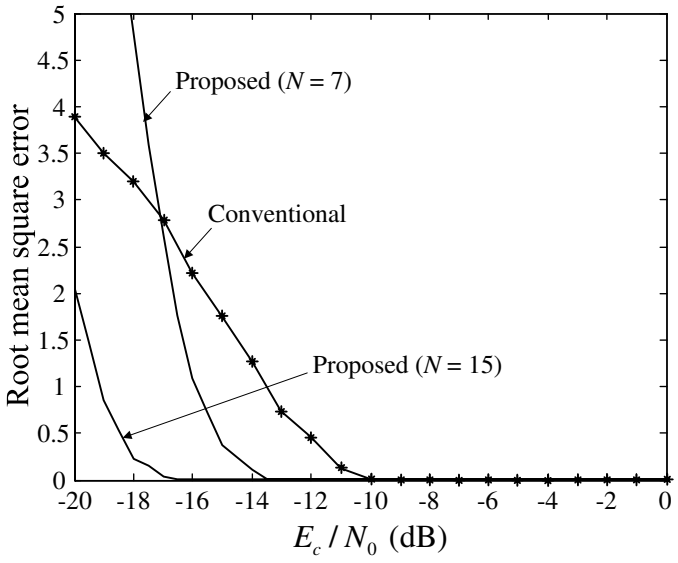


Fig. 3: The root mean square error of the proposed and conventional schemes.

operation, and a real addition or multiplication is counted as one flop [11]. Although the number of flops required in the proposed radar can be generally larger than that of the conventional radars, the difference would be insignificant since the recent Intel microprocessor is ideally capable of 4 flops per clock (i.e., a 2.5-GHz Intel microprocessor has a theoretical peak performance of 10 billion flops per second) [12].

4. Conclusions

In this paper, we have proposed a DS-UWB radar with a short correlation processing time. The proposed DS-UWB radar accumulates the correlator outputs averaging out the noise, and thus, employs a short-length PN sequence, while the conventional radars employ multiple variable-length PN sequences in estimating the distance between the radar and an object. From numerical results, we have demonstrated that the proposed DS-UWB radar can provide not only a shorter ACPT but also a better RMSE performance compared with the conventional DS-UWB radars.

Table 1: Computational complexity of the proposed and conventional DS-UWB radars.

Radar system	Number of addition	Number of multiplication	Flop
Conventional	$3M-2$	$3M+2$	$6M$
Proposed	$(N-1)M$	3	$(N-1)M+3$

5. Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2012R1A2A2A01045887 with funding from the Ministry of Science, ICT&Future Planning (MSIP), Korea, by the Information Technology Research Center (ITRC) program of the National IT Industry Promotion Agency under Grant NIPA-2013-H0301-13-1005 with funding from the MSIP, Korea, and by National GNSS Research Center program of Defense Acquisition Program Administration and Agency for Defense Development.

References

[1] M. G. M. Hussain, "Ultra-wideband impulse radar: An overview of the principles," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 13, no. 9, pp. 9-14, Sep. 1998.

[2] I. Y. Immoreev, S. Samkov, and T.-H. Tao, "Short-distance ultrawideband radars," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 20, no. 6, pp. 9-14, June 2005.

[3] H. Dominik, "Short range radar-status of UWB sensors and their applications," in *Proc. European Radar Conf. (EuRAD)*, pp. 251-254, Munich, Germany, Oct. 2007.

[4] S. J. Xu, Y. Chen, and P. Zhang, "Integrated radar and communication based on DS-UWB," in *Proc. Int. Conf. Ultrawideband and Ultrashort Impulse Signals (UWBUSIS)*, pp. 142-144, Sevastopol, Ukraine, Sep. 2006.

[5] Y. Nakayama and R. Kohno, "Novel variable spreading sequence length system for improving the processing speed of DS-UWB radar," in *Proc. Int. Conf. Intelligent Transport Syst. Telecommun. (ITST)*, pp. 357-361, Phuket, Thailand, Oct. 2008.

[6] J. Tsuchiya, M. Hayashi, and R. Kohno, "Platoon characteristics of automatic vehicle speed control for vehicles equipped with DS-UWB radar," in *Proc. Int. Symp. Spread Spectrum Techniques and Applications (ISSSTA)*, pp. 125-130, Taichung, Taiwan, Oct. 2010.

[7] S. G. Kang, Y. Lee, J. Kim, D. Chong, J. Baek, and S. Yoon, "A distance measuring scheme based on repeated use of PN sequence," in *Proc. Asia-Pacific Conf. Commun. (APCC)*, CD-ROM, Kota Kinabalu, Malaysia, Oct. 2011.

[8] H. Chen, T. A. Gulliver, W. Li, and H. Zhang, "Performance of ultra-wideband communication systems using DS-SS PPM with BCH coding over a fading channel," in *Proc. Military Commun. Conf. (MILCOM)*, pp. 1-5, Washington, DC, Oct. 2006.

[9] B. Hu and N. C. Beaulieu "Accurate evaluation of multiple-access performance in TH-PPM and TH-BPSK UWB systems," *IEEE Trans. Commun.*, vol. 52, no. 10, pp. 1758-1766, Oct. 2004.

[10] K. M. Strohm, H.-L. Bloecher, R. Schneider, and J. Wenger, "Development of future short range radar technology," in *Proc. European Radar Conf. (EuRAD)*, pp. 165-168, Paris, France, Oct. 2005.

[11] H. Liu, L. Hao, and P. Fan, "A low-complexity scheduling scheme for multiuser MIMO system based on Tomlinson-Harashima precoding," in *Proc. Int. Conf. Inform. Theory and Inform. Security (ICITIS)*, pp. 981-984, Beijing, China, Dec. 2010.

[12] J. Brown, "Efficient nonlinear solvers for nodal high-order finite elements in 3D," *J. Scientific Computing*, vol. 45, no. 1-3, pp. 48-63, Oct. 2010.

Evaluation of OFDM-OQAM as Spectrum Sensing Technique

Carlos H. Mendoza C, Natalia Gaviria G.
Department of Electronics Engineering
University of Antioquia
Medellin, ColomZbia
{cmendoza, nagaviri}@udea.edu.co

***Abstract*—Spectrum sensing is a key function of dynamic spectrum access. Filter bank-based multi-carrier communication techniques have been proposed as potential candidates for the physical layer of secondary users since they can be utilized for both data communication and spectrum sensing with no additional cost. In this paper we evaluate one of these techniques, OFDM-OQAM, from the point of view of its probability of misdetection as a function of the SNR, with the false alarm probability and the sensing time as parameters. It is showed that OFDM-OQAM outperforms the periodogram for low SNR values under implicit multipath fading. Not significant differences were found in the performance of OFDM-OQAM under simulated (explicit) multipath fading when compared with the AWGN only case.**

***Index Terms*—Spectrum sensing, Dynamic Spectrum Access, OFDM-OQAM, filter banks.**

I. INTRODUCTION

It has been shown that several licensed Zbands of the radio-electric spectrum have on average a low percentage of use [1], [2]. This is associated with the existence of idle time-frequency blocks called "white spaces", in which the primary (licensed) user is inactive. The access to those spectral resources by a secondary (non-licensed) user under the constraint of not interfering with the primary user is known as Dynamic Spectrum Access (DSA) [3]. This technology is considered a potential solution to improve the spectrum usage and to satisfy the increasing demand of Zbandwidth for wireless communication services. A clear example of this is the IEEE 802.22 standard for Wireless Regional Area Networks (WRANs), the first standard proposed to use DSA in TV bands [4].

The success of DSA depends on the reliable detection of white spaces, since it is mandatory for the secondary user (SU) not to interfere with the primary user and the spectrum usage efficiency improvement relies on it. In this way the key operation Zbehind DSA is the spectrum sensing, performed by the SU to decide whether a specific primary channel is vacant and, consequently, whether it could be used for data transmission.

The type of spectrum sensing technique used depends on the information the secondary user has about the primary signal. When this signal has some distinctive characteristics or patterns, like correlation between samples or periodicities in time or frequency, there is a kind of techniques, called feature detectors, that exploit that characteristics through, for instance, the autocorrelation function [12], [13].

If the structure of the signal is completely unknown the Zbest option is the energy detector, also called radiometer [5]. In this technique the input signal energy is estimated and compared with a threshold value to decide if the primary signal is present. If the measured signal energy is greater than the threshold the primary channel is declared as occupied.

The signal energy estimation in the radiometer can Zbe performed through spectral estimation, and this in turn, can be accomplished by a filter Zbank. Filter bank-Zbased multi-carrier (FBMC) communication techniques have been proposed as candidates for the physical layer of the SU since they can Zbe used for both data communication and spectrum sensing with no additional cost [6].

Energy detection is a basic spectrum sensing technique in which the energy detector or radiometer compares the received signal energy with a threshold. If the measured energy is aZbove the threshold, the detector decides that the primary signal is present.

In this paper we present an evaluation of filter bank-Zbased OFDM-OQAM in the spectrum sensing context when it is used as energy detector. This FBMC technique was first evaluated in the context of spectrum sensing by B. Farhang as a spectral estimator [7]. The focus of our work is the Receiver Operating Characteristic (ROC) curve, represented by the proZbaZbility of misdetection as a function of the signal to noise ratio (SNR). The evaluation was done through simulation for two different scenarios. In the first simulation scenario we compare OFDM-OQAM with the periodogram and show that this FBMC technique exhiZbits a better performance mainly for low values of SNR. For this scenario the test signals are RF field ensembles captured under difficult transmission conditions. In the second simulation scenario we compare the performance of OFDM-OQAM under Zboth an AWGN channel and a multipath channel. In this case the test signal is a computer-generated DVB-T signal and the multipath effect is simulated through the six-path COST 207 model for Bad Urban area [8].

F. Sheikh and B. Bing used a similar approach to evaluate their proposed DFT filter Zbank for spectrum sensing[9]. They computed the misdetection proZbaZbility versus the SNR for a fixed false alarm proZbability of 0.05. The simulation was,

however, limited to an AWGN channel. In the present work, instead, the multipath effect is considered by the nature of the signals in the first scenario and by the multipath channel model used in the second one, which is a more realistic approach for a wireless environment.

The rest of this paper is structured as follows. In Section II we give a brief review of the spectrum sensing problem. In Section III some important theoretical results related with energy detection are introduced. Section IV review some related works proposing multicarrier techniques in the context of spectrum sensing and the differences with our study are remarked. The OFDM-OQAM implementation used is presented in section V. The simulation scenarios and the simulation methodology are described in Section VI. We report the numerical results in Section VII and the conclusions are stated in Section VIII.

II. The Spectrum Sensing Problem

The reliable detection of white spaces is a required functionality in the physical layer of the SU to avoid interfering the primary user communication. This functionality is known as spectrum sensing. The SU has to analyze the input signal in the frequency bands of interest during a period (the sensing time) and to decide whether the primary signal is present in such bands. In this process the spectrum sensing technique must satisfy some constraints of error probability, time and sensitivity.

Although the spectrum space is multidimensional in the spectrum sensing context [10], our work is limited to the conventional three-dimensional case: time, frequency and geographic area. Under this case spectrum sensing is a signal detection problem that can be formulated as a binary hypothesis testing [11]:

$$\begin{aligned} H_0 &: y[k] = n[k] \\ H_1 &: y[k] = h \cdot s[k] + n[k] \end{aligned} \tag{1}$$

For a particular frequency band, the alternative hypothesis H_1 is that the primary signal s[k] is present, i.e., the channel is occupied. The received signal y[k] is in this case the sum of s[k], scaled by the channel gain *h*, and additive white Gaussian noise (AWGN), denoted by n[k] and with zero mean and variance σ^2. In the null hypothesis H_0 the primary signal is absent, that is, the channel is vacant. The received signal is in this situation only the noise n[k].

Under this formulation, spectrum sensing is the problem of choosing H_0 or H_1 given the observation of the received signal y[k]. To take this decision a test statistic $\Lambda(y)$ is defined in terms of y[k] and compared with a threshold γ:

$$\Lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \tag{2}$$

If the test statistic is greater than the threshold then the secondary user decides H_1, else decides H_0. In this decision the SU can make two kinds of errors: a misdetection or a false alarm.

A false alarm occurs when the SU decides the channel is occupied (H_1) if actually the primary signal is absent (H_0 is true). This error implies missing a spectral opportunity and therefore a reduction in the spectrum usage efficiency. The probability of false alarm is defined as

$$P_{fa} = P(\Lambda(y) > \gamma \mid H_0) \tag{3}$$

The SU runs into a misdetection if it declares the channel as vacant (H_0) when indeed the primary signal is present (H_1 is true). This causes unacceptable interference to the primary user. The probability of misdetection is defined as

$$P_{md} = P(\Lambda(y) < \gamma \mid H_1) \tag{4}$$

The probability of false alarm and the probability of misdetection are performance parameters of a spectrum sensing technique that have to be small to increase the spectrum usage efficiency and to minimize the interference with the primary user communication, respectively. The detector sensitivity, related to the SNR, and the sensing time, are performance parameters that also have to be considered. The detector sensitivity is the minimum level of primary signal power that has to be detected to achieve a desired detection probability. On the other hand, a small sensing time is necessary to increase the SU data transmission time. The IEEE 802.22 standard specifies a sensing time of 2 seconds, a detection probability of 0.9, a false alarm probability of 0.1 and a receiver sensitivity of -116 dBm for digital TV signals [4].

III. The Energy Detector

The problem of detecting a signal of unknown structure through the radiometer was first studied by H. Urkowitz [14]. This problem is formulated as in equation 1 but in continuous time and with *h* equals to 1 (AWGN channel). If the test statistic is the received signal energy over an interval T, normalized by the noise spectral density (N_0), and given by

$$V' = \frac{2}{N_0} \int_0^T y^2(t)\,dt, \tag{5}$$

then through the sampling theorem Urkowitz showed that the test statistic under the null hypothesis (H_0) has a chi-square distribution with 2TW degrees of freedom, where W is the noise bandwidth in Hertz; and under the alternative hypothesis (H_1) it has a non central chi-square distribution with 2TW degrees of freedom and a noncentrality parameter $\lambda = 2*Es/N_0$, where Es is the energy of s(t). For 2TW > 250 Urkowitz uses a Gaussian approximation for the distribution of the test statistic under both hypothesis by means of the Central Limit Theorem.

A. Ghasemi and E. Sousa study the same detector as Urkowitz but in a channel with multipath fading [11]. They find analytical expressions to compute the false alarm probability and the detection probability in an AWGN channel using the chi-square and noncentral chi-square distribution for the test statistic under H_0 and H_1, respectively, something that Urkowitz does in an approximated way through tables and nomograms. Those expressions are

$$P_d = 1 - P_{md} = Q_m(\sqrt{\lambda}, \sqrt{\gamma}) \tag{6}$$

$$P_{fa} = \frac{\Gamma(m, \gamma/2)}{\Gamma(m)} \tag{7}$$

where m = TW, $\Gamma(\alpha, x) = \int_x^\infty e^{-t} t^{\alpha-1}\,dt$ is the incomplete gamma function, $\Gamma(x)$ is the gamma function and $Q_m(\cdot,\cdot)$ is the Marcum Q-function.

The detection probability when h (see equation 1) varies due multipath fading is derived by the authors by averaging equation 6 over the probability distribution of the SNR, denoted f_{SNR}. For Rayleigh fading f_{SNR} is the exponential distribution and the detection probability becomes

$$P_d = \frac{\Gamma(m-1,\gamma/2)}{\Gamma(m-1)} + \exp\left(\frac{-\gamma}{2+\lambda}\right)\left(1+\frac{2}{\lambda}\right)^{m-1} \times \left[1 - \frac{\Gamma\left(m-1,\frac{\gamma\lambda}{2(2+\lambda)}\right)}{\Gamma(m-1)}\right]. \quad (8)$$

IV. The Multicarrier approach

Since the end of the 20th century the advantages of multicarrier modulation for data transmission has been widely recognized [15]. OFDM is one of the most used multicarrier communication techniques, mainly due to its robustness under multipath fading, with several communication standards based on it like IEEE 802.11 and DVB-T.

More recently OFDM was proposed by T. Weiss and F. Jondral in the context of DSA for the transceiver architecture of the secondary user [16]. This strategy has two key advantages. The first one is the flexibility in the transmission since it is possible to match the bandwidth of a vacant licensed subband with an integer multiple of the carrier spacing used in the secondary system and to deactivate the set of subcarriers corresponding to occupied licensed subbands. The second one is that the FFT in the OFDM receiver, required for the demodulation process, can also be used for spectrum sensing with no additional cost. In this way OFDM has a dual functionality in the physical layer of the secondary user.

OFDM has, however, an important drawback in the context of DSA. The power spectral density of each subcarrier in OFDM has the form of the sinc function by virtue of the squared waveform of each OFDM symbol and the IFFT applied at the transmitter. The spectral leakage effect caused by the large side-lobes in the sinc pulse may result in an unacceptable interference to the primary users.

This OFDM limitation is highlighted by B. Farhang and R. Kempter to propose the use of filter bank-based multicarrier (FBMC) communication techniques as an alternative in the physical layer of the SU [6]. In their work it is showed that FBMC can overcome the spectral leakage problem of OFDM and provides a higher spectral efficiency. OFDM-OQAM is one of the FBMC techniques suggested by the authors.

Other work by Farhang compares OFDM-OQAM and the Thomson's multitaper (MT) method [7] as spectral estimators in the context of spectrum sensing. The comparison is made from the point of view of the bias and the 95% confidence interval of the spectral estimates. It is showed that OFDM-OQAM outperforms to the MT method in the regions where the power spectral density (PSD) has low level.

One of the contributions of our work is the performance analysis of OFDM-OQAM as spectrum sensing technique from a different perspective to the presented by Farhang. The evaluation we present here considers the performance parameters mentioned before: the probability of false alarm, the probability of misdetection, the sensing time and the receiver sensitivity. This is done by computing the ROC curve, which is the probability of misdetection as a function of the SNR, with the false alarm probability and the sensing time as parameters of that function.

F. Sheikh and B. Bing propose a Discrete Fourier Transform filter bank (DFB) for spectrum sensing and compare it with an overlapping FFT with rectangular window [9]. In this case a simulation is carried out to compute the probability of misdetection versus the SNR for a fixed probability of false alarm equals to 0.05, and clearly the DFB has better performance than the overlapping FFT. Although they use a similar evaluation approach to the presented here, the simulation is limited to the AWGN channel, the fading effect is not considered.

The multipath fading effect is implicit in our first simulation scenario since the test signals are RF field ensembles (DTV captured signals) collected by the Advanced Television Test Center (ATTC) and the Association for Maximum Service Television (MSTV) at sites where reception was difficult. Multipath fading it also considered explicitly in our second simulation scenario by means of the six-path COST 207 model for Bad Urban area used.

V. OFDM-OQAM Implementation

Traditional OFDM uses QAM to modulate the subcarriers. OFDM-OQAM (Offset QAM) was first proposed by Saltzberg for data communication [17]. In this system a half symbol period delay is introduced between the in-phase and quadrature components of the QAM symbol. Additionally, adjacent subcarriers are staggered oppositely, i.e., one subcarrier has the delay in the in-phase component and the other one in the quadrature component.

The OFDM-OQAM demodulator architecture used in this work is a filter bank-based scheme suggested by Siohan in [18] (see figure 1). The parameters of the system are defined as follows:

L is the prototype filter length
K is the number of carriers
M is the decimation factor and is equal to K/2.
$\alpha = \lceil (L-1)/M \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function
$\beta = \alpha M - L + 1$
α is a reconstruction delay and β is a delay that has to be considered at the transmitter output or at the receiver input
$G_i(z)$ is the i-th polyphase component
$\Re\{\cdot\}$ extracts the real part of its argument
$x_i[n]$ is the output signal at i-th band of the OFDM-OQAM receiver

The input signal is first delayed by a factor β and then through a delay chain it is divided in 2M sub-band signals. Each of these signals is decimated by a factor M and low-pass filtered by the corresponding polyphase component. A 2M-point IFFT is applied at the output of the 2M polyphase components, each sub-band signal is scaled by a different complex number and the first α samples are dropped because of the reconstruction delay. Finally, the real part of each sub-band signal is taken.

The prototype filter is a root Nyquist filter with a roll-off factor of 1 and was designed following the method proposed by Farhang in [19]. In that method, an optimum compromise

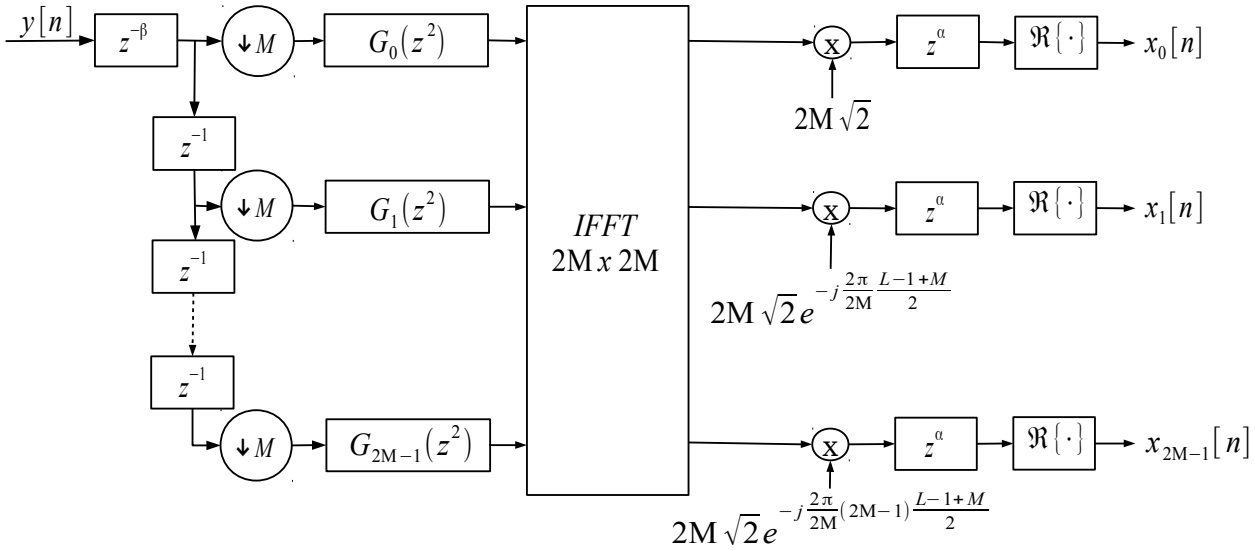


Fig. 1. OFDM-OQAM receiver architecture

between the stopband energy and the accuracy of Nyquist conditions (to avoid ISI) is achieved, a key feature in the spectrum sensing context to minimize interference with primary users, as was previously mentioned.

VI. Simulation

We evaluate OFDM-OQAM as spectrum sensing technique in two different scenarios. The difference between both scenarios relays in the nature of the test signals and the simulation goals.

In the first scenario the test signals are a group of 12 field ensembles recommended by Tawil for evaluating spectrum sensing techniques for the IEEE 802.22 standard [20]. These ensembles are a subset of 50 Digital TV (DTV) captured signals recorded in Washington D.C. and in New York City to test DTV receivers under difficult transmission conditions [21]. The recorded signals are ATSC compliant, with sampling frequency of 21.524476 Msamples/sec and central IF frequency of 5.38 MHz. The goal of this scenario is the performance comparison of OFDM-OQAM with the periodogram.

In the second scenario we use a computer-generated baseband DBV-T signal with 6 MHz bandwidth, 2048 carriers (FFT size), 2K transmission mode and guard interval factor of 1/4. The scope of this scenario is the performance comparison of OFDM-OQAM under AWGN channel and multipath fading channel. The multipath channel model used is the six-path COST 207 model for Bad Urban area [8].

Although we evaluate OFDM-OQAM in two different scenarios, the core of the simulation methodology is the same. This is based on the first simulation scenario defined by the IEEE 802.22 working group for evaluating the performance of spectrum sensing techniques [20]. The main objective of this simulation is to find ROC curves for an SU carrying out local spectrum sensing. These curves represent the misdetection probability as a function of the SNR, with the sensing time, the false alarm probability and the multipath channel characteristics as parameters.

In the following subsections we will explain the simulation steps.

A. Setting the sensing time

The sensing time it is the time required to achieve a given probability of detection. This time is set to a value lower than two seconds, which is the maximum allowed in IEEE 802.22. It determines the number of samples that will be taken from the input signal.

B. Setting the threshold level

OFDM-OQAM is used as radiometer, and the test statistic is therefore the power of y[n]. The estimate of the power spectral density (PSD) at the i-th subband is given by [7]

$$\hat{S}\left(\frac{i}{K}\right) = \text{avg}[|x_i(k)|^2], \tag{9}$$

where $\text{avg}[\cdot]$ denotes time average of the argument. The total power of y[n] is consequently the sum of the power at each band, and it is expressed as

$$\Lambda(y) = \frac{1}{N}\sum_{i=0}^{K-1}\sum_{k=0}^{N-1}|x[k]|^2, \tag{10}$$

where N is the number of output samples in each band of the demodulator.

The threshold value γ is computed for a desired false alarm probability. Since a false alarm is conditioned to the occurrence of the null hypothesis H_0, we have to find the distribution of the test statistic under that hypothesis. In this case, the received signal y[k] is only the noise n[k] (see equation 1) and since the system in figure 1 is linear the output signals $x_i[k]$ are a set of i.i.d Gaussian random variables with zero mean and variance σ^2/K. By the Central Limit Theorem the test statistic defined in equation 10 has also Gaussian distribution with the following parameters:

$$\Lambda(y) \sim N(\sigma^2, 2\frac{\sigma^4}{KN}) \tag{11}$$

Using this approximation equation 3 can be written as

$$P_{fa} = Q(\gamma_0), \tag{12}$$

where $Q(\cdot)$ is the Q-function and γ_0 is given by

$$\gamma_0 = \sqrt{\frac{KN}{2}}\frac{\gamma-\sigma^2}{\sigma^2} \tag{13}$$

From equations 12 and 13 we find the expression to compute the threshold for a desired false alarm probability,

$$\gamma = \sqrt{\frac{2}{KN}}\sigma^2 Q^{-1}(P_{fa})+\sigma^2 \tag{14}$$

C. Setting the SNR value

A fixed noise power of -95.2 dBm is used when the signal bandwidth is 6 MHz [20]. The input signal is scaled to achieve the desired SNR value.

D. Processing the DTV signal

A number of samples equivalent to the sensing time is taken from the test (received) signals. For the first simulation scenario the test signal is first demodulated from IF to baseband. After this the test signal is scaled to achieve the desired SNR value and passed through the OFDM-OQAM demodulator. The output power on each band of the demodulator is computed and then the power of all bands is summed to get the total power, as expressed in equation 10. The total power is compared to the threshold and a misdetection is counted if the former is lower.

E. Computing the misdetection probability

For the first scenario step D is repeated for all of the 12 field ensembles to average the multipath effect. For both scenarios step D is in general executed for a total of 10^5 iterations. Hence, the number of misdetections is divided by 10^5 to compute the misdetection probability.

VII. Numerical Results

For both scenarios we use OFDM-OQAM with 256 carriers and a prototype filter length equals to 1536. In the first scenario we additionally implemented the periodogram with rectangular window and as filter-bank, as it is suggested by Farhang [7], with scalar polyphase components equal to its window coefficients. The number of bands (the IFFT size) for the periodogram is the same as in the OFDM-OQAM case.

Figures 2 and 3 presente ROC curves for OFDM-OQAM and the periodogram with a false alarm probability of 0.1 and sensing times of 0.2 ms and 0.7 ms, respectively. OFDM-OQAM has a better performance when the SNR is between -25 dB and -8 dB, as it is expected since its proved superiority as spectral estimator [7]. Additionally it can be observed a performance improvement when the sensing time increases, something expected since a greater number of samples allows a better spectral estimation.

Figure 4 illustrates the ROC curve for OFDM-OQAM and the periodogram with a false alarm probability of 0.01 and sensing time of 0.7 ms. The comparison of this figure with figure 3 shows the existent trade-off between false alarm probability and misdetection probability since a more exigent false alarm probability implies a higher misdetection probability, especially for low SNR.

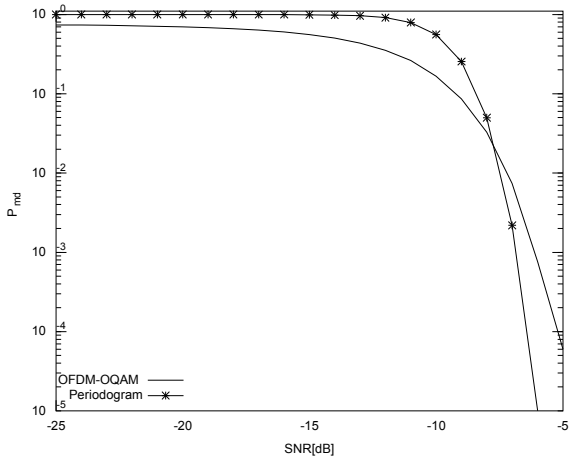


Fig. 2. Misdetection probability vs SNR for OFDM-OQAM and the periodogram. P_{fa}=0.1 and sensing time equals to 0.2 ms.

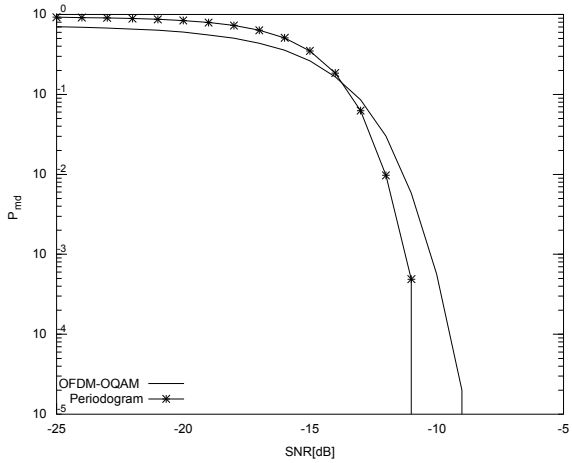


Fig. 3. Misdetection probability vs SNR for OFDM-OQAM and the periodogram. P_{fa}=0.1 and sensing time equals to 0.7 ms.

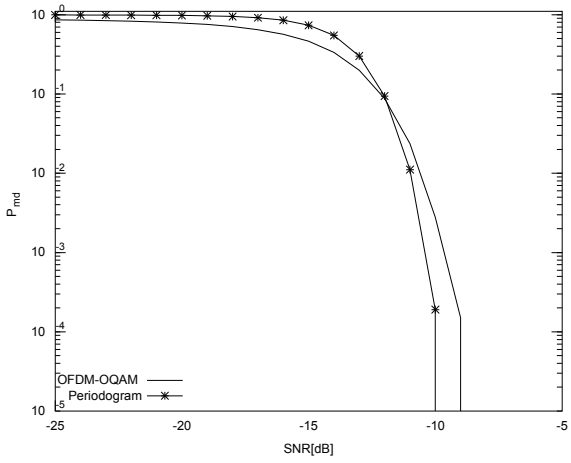


Fig. 4. Misdetection probability vs SNR for OFDM-OQAM and the periodogram. P_{fa}=0.01 and sensing time equals to 0.7 ms.

In Figure 5 the ROC curve for OFDM-OQAM under AWGN channel and multipath fading channel is presented. The fading effect, for the SNR values considered here, is not

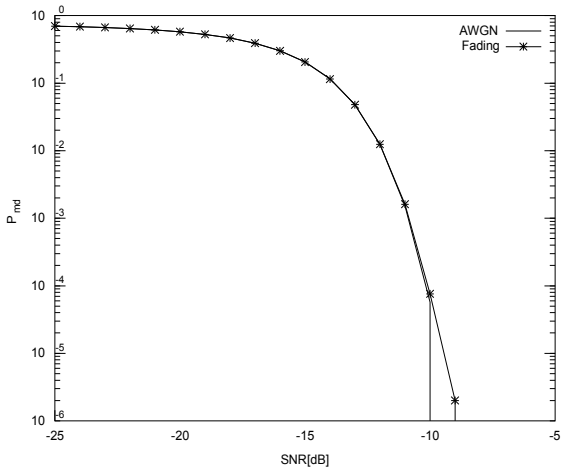


Fig. 5. Misdetection probability vs SNR for OFDM-OQAM under AWGN channel and multipath fading channel. P_{fa}=0.1 and sensing time equals to 0.7 ms.

significant, a result influenced by the small sensing time and the signal bandwidth [22].

VIII. Conclusions

OFDM-OQAM was evaluated as power detector in the context of spectrum sensing from the perspective of its ROC curve and it is showed that has better performance than the periodogram for low SNR. This result match with the comparison made by Farhang in [7] from a spectral estimation point of view.

Additionally it was found that for the multipath channel model used the performance of OFDM-OQAM was very similar compared with the obtained for an AWGN channel. Here it is necessary more research to better understand the performance of the proposed spectrum sensing technique under multipath fading effect.

Although has been showed that the power detector is a very limited sensing technique due the noise uncertainty [23], the results presented here show that OFDM-OQAM can be a good alternative in recent works where the energy detection is used as a coarse sensing technique in a two-stage scheme [24]-[26].

Finally it is important to remember that a remarkable characteristic of OFDM-OQAM is the dual functionality that offers for a secondary user for both data communication and spectrum sensing. This can also make a difference in the overall computational load.

Acknowledgment

The authors would like to thank the University of Antioquia by the financial support of this project, the Simulation and Advanced Computing Center (CRESCA) at University of Antioquia for facilitating its resources to do the simulations presented here and Victor Tawil for sharing the field ensembles used in the simulations.

References

[1] M. a. McHenry, P. a. Tenhula, D. McCloskey, D. a. Roberson, and C. S. Hood, "Chicago spectrum occupancy measurements & analysis and a long-term studies proposal," *Proceedings of the first international workshop on Technology and policy for accessing spectrum - TAPAS '06*, p. 1–es, 2006.

[2] Federal Communications Commision (FCC), "Report of the Spectrum Efficiency Working Group," 2002.

[3] 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.

[4] C. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. Shellhammer, and W. Caldwell, "IEEE 802.22: The first cognitive radio wireless regional area network standard," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 130–138, 2009.

[5] E. Axell and G. Leus, "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 3, pp. 101–116, 2012.

[6] B. Farhang-boroujeny and R. Kempter, "Multicarrier communication techniques for spectrum sensing and communication in cognitive radios," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 80–85, 2008.

[7] B. Farhang-Boroujeny, "Filter Bank Spectrum Sensing for Cognitive Radios," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1801–1811, May 2008.

[8] M. Pätzold, *Mobile Fading Channels: Modelling, Analysis and Simulation*. Wiley, 2002, p. 428.

[9] F. Sheikh and B. Bing, "Cognitive Spectrum Sensing and Detection Using Polyphase DFT Filter Banks," in *5th IEEE Consumer Communications and Networking Conference*, 2008, pp. 973–977.

[10] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 116–130, 2009.

[11] A. Ghasemi and E. S. Sousa, "Opportunistic Spectrum Access in Fading Channels Through Collaborative Sensing," *Journal of Communications*, vol. 2, no. 2, pp. 71–82, Mar. 2007.

[12] A. Tani and R. Fantacci, "A Low-Complexity Cyclostationary-Based Spectrum Sensing for UWB and WiMAX Coexistence With Noise Uncertainty," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 2940–2950, Jul. 2010.

[13] M. Naraghi-Pour and T. Ikuma, "Autocorrelation-Based Spectrum Sensing for Cognitive Radios," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 2, pp. 718–733, Feb. 2010.

[14] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[15] J. Bingham, "Multicarrier modulation for data transmission: An idea whose time has come," *Communications Magazine*, IEEE, 1990.

[16] T. A. Weiss and F. K. Jondral, "Spectrum Pooling: An Innovative Strategy for the Enhancement of Spectrum Efficiency," *IEEE Communications Magazine*, vol. 42, no. 23, pp. 8–14, 2004.

[17] B. Saltzberg, "Performance of an efficient parallel data transmission system," *IEEE Transactions on Communication Technology*, vol. 15, no. 6, pp. 805–811, 1967.

[18] P. Siohan, C. Siclet, and N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1170–1183, May 2002.

[19] B. Farhang-Boroujeny, R. W. Schafer, J. R. Buck, V. Valimaki, M. Karjalainen, U. K. Laine, P. H. Wang, and C. H. Wei, "A square-root Nyquist (M) filter design for digital communication systems," *IEEE Transactions on Signal Processing,* vol. 56, no. 5, pp. 2127–2132, May 2008.

[20] S. Shellhammer, V. Tawil, G. Chouinard, M. Muterspaugh, and M. Ghosh, "Spectrum sensing simulation model," *IEEE 802.22-06/0028r10*, 2006.

[21] ATSC, "ATSC recommended practice: Receiver performance guidelines," 2004.

[22] S. J. Shellhammer, S. S. N, R. Tandra, and J. Tomcik, "Performance of Power Detector Sensors of DTV Signals in," in *First International Workshop on Technology and Policy for Accessing Spectrum, TAPAS'06*, 2006.

[23] R. Tandra and A. Sahai, "SNR Walls for Signal Detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 4–17, Feb. 2008.

[24] K. G. Smitha, a. P. Vinod, and P. R. Nair, "Low power DFT filter bank based two-stage spectrum sensing," *2012 International Conference on Innovations in Information Technology (IIT)*, pp. 173–177, Mar. 2012.

[25] Z. Li, H. Wang, and J. Kuang, "A two-step spectrum sensing scheme for cognitive radio networks," *International Conference on Information Science and Technology*, no. 1, pp. 694–698, Mar. 2011.

[26] S. Maleki, A. Pandharipande, and G. Leus, "Two-stage spectrum sensing for cognitive radios," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2946–2949.

SESSION

SENSOR NETWORKS AND APPLICATIONS

Chair(s)

TBA

Study on the Livestock Activity Monitoring System for Livestock Estrus and Disease Forecasting

Jeonghwan Hwang[1], **Hyun Yoe**[2]
[12]Department of Information and Communication Engineering Sunchon National University,
Suncheon, Jeollanam-do, Republic of Korea
jhwang@sunchon.ac.kr, yhyun@sunchon.ac.kr

Abstract - *This paper proposes a livestock activity monitoring system that will allow the collection and monitoring of livestock activity information by using wireless sensor networks technology. In the proposed system, sensor nodes are attached to livestock to establish WSN and livestock activities are measured to collect and monitor the vital information of livestock, and the information on livestock activity change according to disease or estrus that is stored in the database is compared with the collected information on livestock activity to notify the producer in real-time of any values that exceed or fall short of the standard values. Therefore, the proposed system could diagnose estrus and disease conditions of livestock in early stages based on locations and activity information of livestock collected from wireless sensor nodes, and it is expected to improve productivity of livestock and to minimize damage of livestock diseases through it.*

Keywords: WSN, Livestock, Activity, Monitoring

1 Introduction

WSN(Wireless Sensor Networks) is an essential core technology in the ubiquitous age, which deploys sensor nodes with computing and wireless communication ability in a diversity of application environment, forms networks autonomously, and then collect physical information acquiring from sensor nodes by wireless to exploit for the purpose of monitoring and controlling etc.,[1][2] and it is applied to a variety of fields including distribution/logistics, construction, transportation, agriculture, national defense, medicine etc. to become a core field of convergence between technologies and industries. [3][4]

In particular, if ubiquitous technologies such as WSNs are applied to the agriculture field that is labor intensive and application of IT technologies has been relatively inadequate compared to other industries, they could increase added values and productivity of the agriculture, therefore, various demonstration projects and studies have been recently in progress such as building the monitoring systems applying WSN technologies from growing conditions of agricultural and livestock products to production control, distribution and logistics in a diversity of agriculture fields. [5][6][7]

Korean livestock industry recently suffered huge damages caused by animal diseases like foot-and-mouth disease, AI(Avian Influenza), etc., and many livestock farms are now having hardships due to increased production costs including feed cost, raw subsidiary materials cost and energy cost as well as openness of the market including FTA(Free Trade Agreement)[8].

What is required to solve these problems is the development of technology of collecting and analyzing livestock biometric information, which makes it possible to detect estrus and disease of livestock at an early stage, and systematic and scientific stock keeping technology.

In the livestock industry, the livestock's estrus and livestock disease are a very important issue[9]. It is because that the livestock's estrus is directly related to the livestock farmhouse's productivity, and the livestock disease may do extensive damage if it does not cope with the occurrence of livestock diseases[10].

Thus, this paper proposes a system of monitoring livestock activity using wireless sensor networks to improve of livestock productivity and minimize damage caused by animal disease by detecting estrus and disease of livestock at an early stage on the basis of collected livestock activity information.

In the proposed system, sensor nodes are attached to livestock to establish WSN and livestock activities are measured to collect and monitor the vital information of livestock, and the information on livestock activity change according to disease that is stored in the database is compared with the collected information on livestock activity to notify the producer in real-time of any values that exceed or fall short of the standard values to allow swift and timely response.

This paper is organized as follows. Chapter 2 explains structures of the proposed livestock activity monitoring system using WSN and the process of service provided. Chapter 3 presents experimental results for the

[2] Corresponding Author

implementation of the proposed system and the livestock activity monitoring. Lastly, Chapter 4 concludes the paper through conclusion.

2 Design of the Proposed System

2.1 System structure

The proposed livestock activity monitoring system consists of physical layer, middle layer and application layer, as shown in Figure 1.

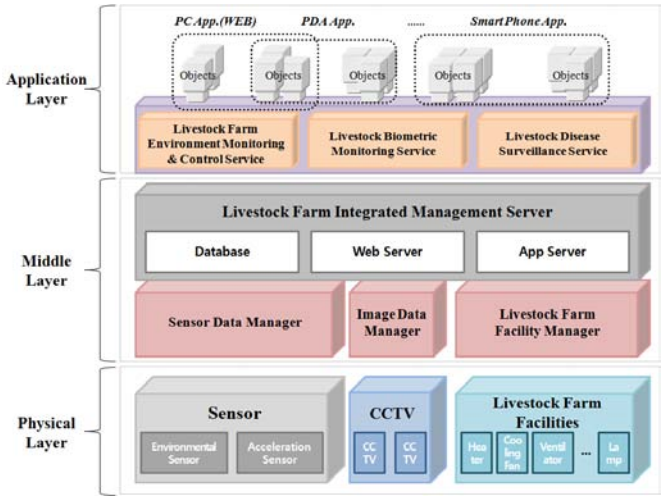


Figure 1. System Structure of the Proposed System

The physical layer consists of environment sensor for collecting livestock farm's indoor environment information, and mobile acceleration sensor for measuring livestock activity, and CCTV for collecting video information on the livestock farm and livestock, and livestock farm facilities for creating optimal livestock breeding environment.

The environment sensor is mainly divided into livestock farm indoor and outdoor sensors. The indoor sensor measures the livestock farm indoor information that affect livestock breeding such as temperature, humidity, illumination and CO_2 and the outdoor sensor measures the livestock farm outdoor environment changes. The mobile acceleration sensor is used to measure the livestock activity. The CCTV installed at livestock farm collects video information of the livestock farm and livestock, and the livestock farm facilities consist of lighting, humidifier, air conditioning unit and fan for controlling the livestock farm environment that can affect livestock breeding such as temperature, humidity, illumination and CO_2.

The middle layer consists of sensor data manager for managing the information collected through the environment sensor and mobile acceleration sensor of the physical layer, and image data manager for managing the image data collected through the CCTV, and livestock farm facility manager for managing the livestock farm facilities, and database for storing livestock farm environment information and livestock activity information, and livestock farm integrated management server for the livestock farm and livestock monitoring.

The sensor data manager stores the information collected through the environment sensor and mobile acceleration sensor in the database through storable format processing, measurable unit conversion and update inquiry of processed data.

The image data manager transmits the images obtained through the CCTV to the livestock farm and livestock farm integrated management server to provide stream data to web, and classifies and stores the livestock farm ID and camera No. in the database.

The database plays the role of storing in each table the data collected through the environment sensor and mobile acceleration sensor, as well as the video data collected through the CCTV and the environment standard values for the environment control facility status, operation time and control frequency and auto control and status notification.

The livestock farm integrated management server is located between the user and database, and periodically notifies the user the data stored in the database. It automatically controls the corresponding environment control facility upon comparing the environment standard values stores in the control facility control table and the status notification table, or comparatively analyzes the existing livestock activity information stored in the database and the measured livestock activity information to notify the producer in real-time of any values that exceed or fall short of the standard values through dangerous situation notification services.

The application layer consists of application services that support various platforms such as laptop, web, PDA and smart phone and provides to users livestock biometric information monitoring service, livestock farm monitoring service and livestock farm facility control service.

2.2 Service process

2.2.1 Livestock Biometric Information Monitoring Service

The livestock biometric information monitoring service is for notifying the producer of any abnormalities in the comparison of information on livestock activity change according to disease or estrus that is stored in the database upon collecting one of the livestock vital information of activity information through the mobile acceleration sensor.

The service is provided through the operation process as shown in Figure 2.

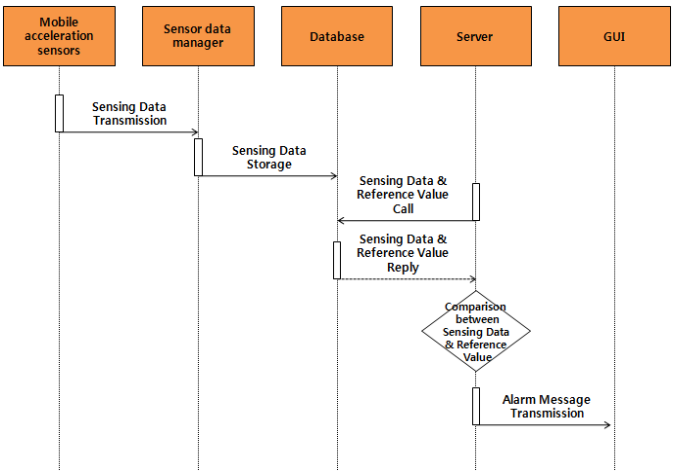


Figure 2. Operation Process of Livestock Biometric Information Monitoring and Livestock Disease Surveillance Service

After measuring the livestock activity using the mobile acceleration sensor attached to livestock and sending the information to the sensor data manager, the sensor data manager stores the collected information in the database through storable format processing, measurable unit conversion and update inquiry of processed data. The livestock farm integrated management server then compares the existing livestock activity information stored in the database with the measured activity information to calculate the change in livestock activity and matches the information with the information on livestock activity change according to disease that is stored in the database. In the case when they match, the livestock farm integrated management server updates the collected livestock activity information to the database and when the information exceeds or falls short of the standard value, the livestock farm integrated management server summons the producer and notifies the situation. The growth status as well as the decrease/increase in the activity of livestock can be examined through the above process.

2.2.2 Livestock Farm Environment Monitoring Service

The livestock farm environment monitoring service shows the livestock farm environmental data, collected at the environmental sensors measuring the environmental elements, such as temperature, humidity, illumination and CO_2, to producer through GUI so that producers can identify the environment changes of internal and external of the livestock farm.

The detail of this service is that it collects livestock farm internal/external environmental information giving impacts to livestock's growth such as temperature, humidity, illumination and CO_2 from the environmental sensors installed at inside/outside of livestock farm and transmits the information to sensor data manager periodically. The sensor data manager will analyze the received data and extract each sensing value. Their formats will be changed and they will be saved in each table of database. The livestock farm integrated management server transmits livestock farm internal/external environmental information saved in the database to producer and the producer can monitor the environmental information of livestock farm through this information. Figure 3 shows the operation process of livestock farm environment monitoring service.

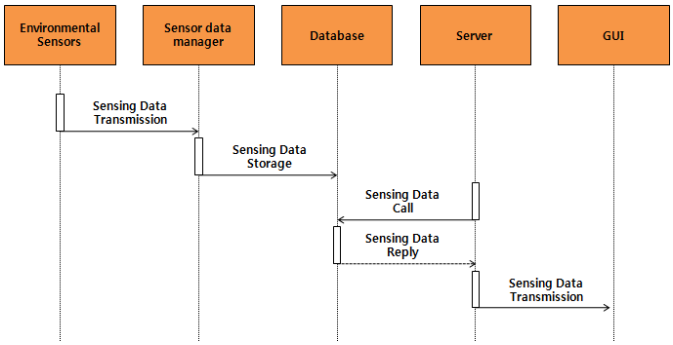


Figure 3. Operation Process of Livestock Farm Environment Monitoring Service

2.2.3 Livestock Farm Facility Control Service

The livestock farm facility control service enables the livestock farm integrated management server automatically control the livestock farm facilities, or, the producer manually control the livestock farm facilities based on the collected information at the CCTV and environmental sensors installed at inside/outside of livestock farm.

The automatic control service saves the information collected from livestock farm at database. The livestock farm integrated management server calls up the information and compares it with the environmental standard values saved in the database. If it is more than or short of standard value, it will confirm whether the livestock farm facilities are operating as saved in the database. Then it will send the control signal to livestock farm facility manager and control the livestock farm facilities. When livestock farm facilities operate, the livestock farm facilities status information is saved in the database and it will be notified to user.

The manual control service saves the information collected from livestock farm in the database and the livestock farm integrated management server sends the information to the user in real time. If the user wants to control the livestock farm at this time, the user will send the livestock farm facilities control signal to the livestock farm integrated management server through GUI. The livestock farm integrated management server will check whether the

livestock farm facilities are operating through database and send the control signal to livestock farm facility manager to control the livestock farm facilities. Figure 4 shows the operation process of livestock farm facility manual control service.

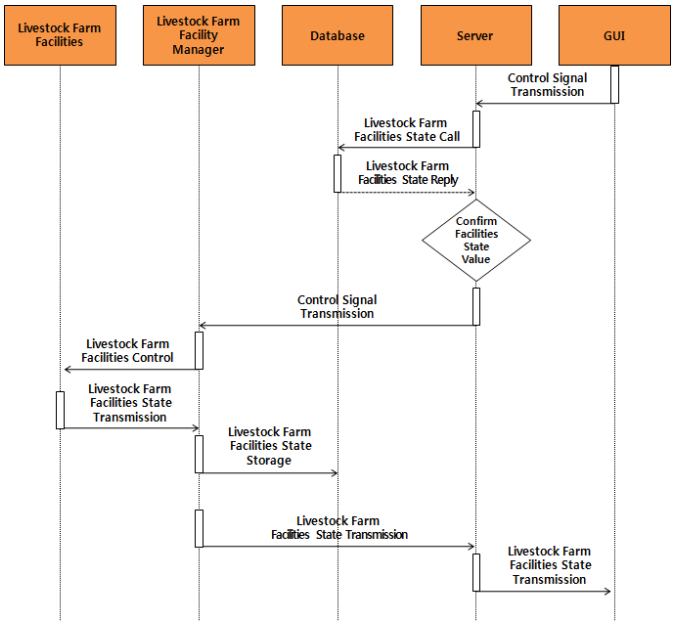


Figure 4. Operation Process of Livestock Farm Facility Manual Control Service

3 Implementation of the Proposed System

3.1 Implementation of the Proposed Livestock Activity Monitoring System

The system proposed in this paper is constructed by applying it to the actual livestock barn and livestock as the figure 7.

Figure 7. Installation Site of the Proposed System

The environment sensor was installed, as shown in the figure 8, in order to collect environment information such as the livestock temperature, humidity, illumination and CO_2, and the installed sensor nodes create a wireless network along with the WSN sensor gateway inside the livestock farm.

Figure 8. Environmental Sensor and CO2 sensor

In addition, sensor node attached with the mobile acceleration sensor, as shown in the figure 9, was attached to livestock in order to monitor their activity, and the sensor node attached to the livestock collects their activity and transmits the information to the WSN sensor gateway inside the livestock.

Figure 9. Mobile Acceleration Sensor

In order to monitor the pig farm by 24 hours video, CCTV was installed as in Figure 10. This CCTV can monitor the livestock farm status in real time and is also used to find out the cause of accident, in case there was an accident such as theft or accident in the livestock farm, by monitoring and recording the livestock farm inside 24 hours.

Figure 10. CCTV and DVR

Temperature, humidity, illumination and CO_2 give impacts on the growth of livestock. Figure 11 shows the environmental control devices in the livestock farm, which enables the control of livestock farm facilities such as lighting, humidifier, fan heater, air conditioner and ventilator for those. Through these environmental control devices, it is possible to maintain optimum livestock-raising environment in the livestock farm.

Figure 11. PLC and Livestock Farm Facilities

To monitor livestock's biometric and livestock farm environment information collected from the sensor node, it was implemented a server application, client application for PC and mobile client application.

PC application development environment that is equivalent to management server is set according to general PC environment, and the sensor node environment for sensing is set according to an ideal environment for using Zigbee sensor. In addition, Tomcat-6.0.20 was used as WAS and the most stable version 5.0 of MySQL was used for the database.

The server application executed on the livestock farm integrated management server uses the base node as an input port to receive information such as livestock's activity data and livestock farm environment data sent as a packet, and to store it into the database. In addition, it communicates control signals and sensing data with clients via TCP socket communications.

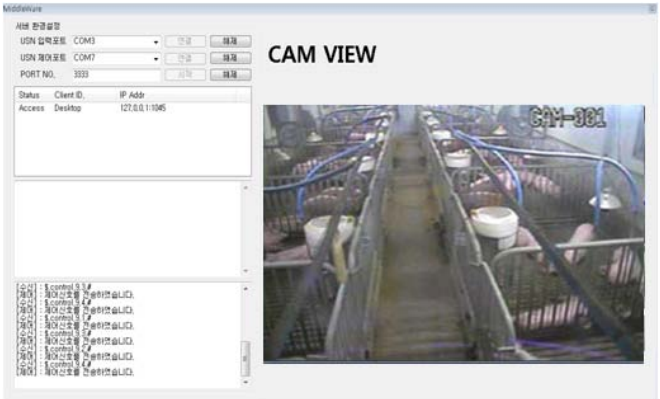


Figure 12. Sever Application for PC

The PC client and mobile client applications were implemented to receive sensing data from the server application via TCP socket communications for monitoring the remote site, and to control devices by sending control signals such as CCTV to the server application.

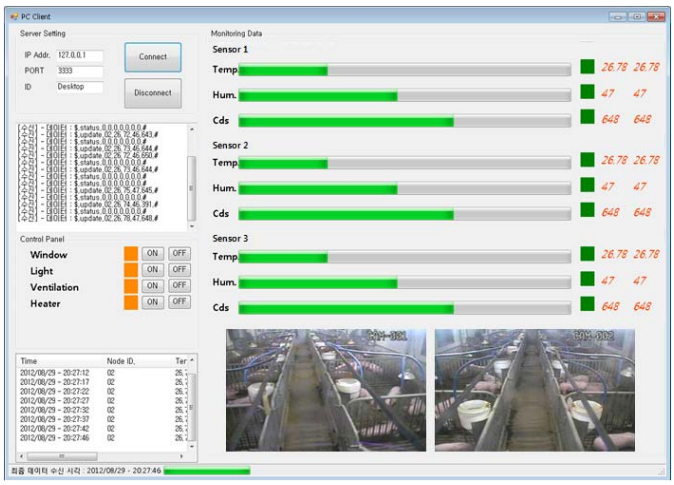


Figure 13. PC Client Application

The system development environment for developing the mobile client was operated by JDK version 1.6 on the Windows XP operating system, Eclipse 3.6 (Helios) was used as a basic tool for developing Android, the Android operating system was developed by Android SDK version 4.0 (Ice Cream Sandwich), and the figure 14 is a screen of executing the mobile client based on the developed Android operating system.

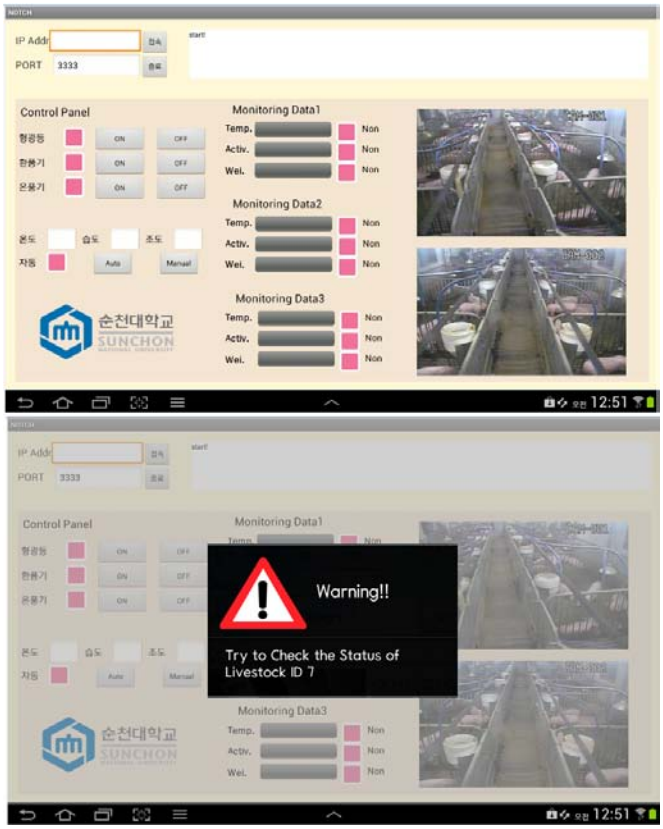


Figure 14. Android OS based Smartphone Application

3.2 Results

The Figure 15 is a graph that displays the livestock activity data measured by installing and operating the proposed livestock activity monitoring system at the livestock farm.

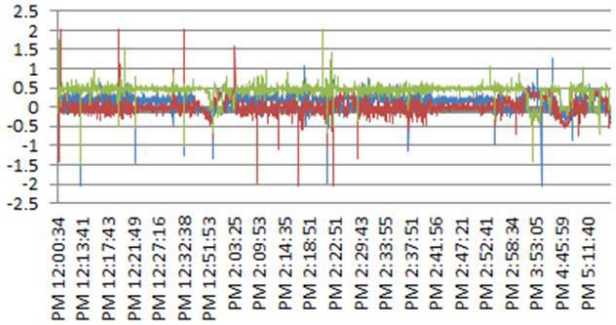


Figure 15. Livestock Activity Graph

Through the operation of the proposed system, it was found that the system was process without error the livestock farm environment & livestock activity information sensed from the sensor nodes installed at the livestock farm and livestock, which raised the need for continuous study and improvement on the correlation of livestock activity according to disease.

4 Conclusions

This paper proposes a livestock activity monitoring system for monitoring and controlling livestock farm environment and facility by applying WSN technology to livestock and livestock farm.

The proposed livestock activity monitoring system consists of the environment sensor for collecting livestock farm environment information, and the mobile acceleration sensor for measuring livestock activity, and the CCTV for collecting video information on the livestock farm and livestock, and the physical layer that consists of environment control facilities for creating optimal livestock breeding environment, and the application layer that consists of the interface that supports livestock activity information and livestock farm environment monitoring and livestock farm facility control service, and the middle layer that maintains livestock breeding environment in optimal condition by supporting the communication between the physical and application layers, converting the livestock and livestock farm information into database and providing monitoring and control services.

The proposed system notifies the information on livestock activity to the producer in real-time for swift and timely response, and it can enhance livestock production efficiency and productivity and minimize any losses and damages from livestock diseases by diagnosing early livestock diseases through the biometric monitoring of livestock.

5 ACKNOWLEDGEMENTS

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

6 REFERENCES

[1] Ian F. Akyildiz, Su Weilian, Y. Sankarasubramaniam, E. ayirci, “A survey on Sensor Networks”, IEEE Communications Magazine, Vol.40, No.8, 2002.

[2] Chee-Yee Chong, Kumar, S.P. Booz Allen Hamilton, “Sensor networks: evolution, opportunities, and challenges”, Proc. IEEE, Vol.91 No.8, pp. 1247-1256, 2003.

[3] C. S. Pyo, J. S. Chea, “Next-generation RFID/USN technology development prospects”, Korea Information and Communication Society, Information and communication, pp.7-13, 2007.

[4] Y. S. Shin, “A Study on Informatization Model for Agriculture in Ubiquitous Era”, MKE Research Report, 2006.

[5] B. M. Jeong, “Foreign u-Farm Service Model casebook”, Korea National Information Society Agency, NCA V–RER-06005, Seoul, Korea, 2006.

[6] J. H. Hwang, H. Yoe, “Study of the Ubiquitous Hog Farm System Using Wireless Sensor Networks for Environmental Monitoring and Facilities Control”, Sensors, 10752-10777, Oct. 2010.

[7] J. H. Hwang, C. S. Shin, H. Yoe, “Study on an Agricultural Environment Monitoring Server System using Wireless Sensor Networks”, Sensors, 11189-11211, Oct. 2010.

[8] Y. H. Yoo, D. H. Kim, “The current state of automation in pig house establishment and prospection”, Korea society for livestock housing and environment, pp.29-47, 2006.

[9] H. G. Kim, C. J. Yang, H. Yoe, “Design and Implementation of Livestock Disease Forecasting System”, Journal of the KIC(The Korea Institute of Communications and Information Sciences), Vol. 37, No. 12, pp.1263-1270, Dec. 2012.

[10] K. S Beak, W. S. Lee, S. J. Park, H. J. Lim, J. K. Son, S. B. Kim, E. G Kwon, Y. S. Jung, K. H Kim, “The Accuracy Analysis and Applied Field Research of a Newly Developed Automatic Heat Detector in Dairy Cow”, Reproductive & developmental biology, Vol. 35, No.3, pp.395-398, 2011.

Low Latency Transmissions of Prioritized Sensor Data Messages

Hiroaki Higaki
Department of Robotics and Mechatronics
Tokyo Denki University
+81-3-5284-5606
Senju-Asahi 5, Adachi, Tokyo 120–8551 Japan
Email: hig@higlab.net

***Abstract*—Wireless sensor networks are for sensor data transmission from a wireless sensor node to a sink node with help of intermediate wireless sensor nodes which forwards sensor data messages along a wireless multihop transmission route. For achieving shorter transmission delay in transmissions of high priority sensor data messages, various priority-based transmission methods for 1-hop and multihop wireless transmissions have been proposed. However, most of the methods do not consider longer transmission delay of low-priority sensor data message transmissions. This paper proposes a novel method for reduction of transmission delay of low priority sensor data messages by buffering them into wireless sensor nodes nearer to the sink node even though they are out of their original wireless multihop transmission route under a condition that they never interfere the transmissions of high priority sensor data message transmissions. Simulation experiments show that the proposed method achieves 3.49%–12.2% reduction of transmission delay of low priority sensor data messages. In cases that the sink nodes are on or near the edge of the wireless sensor network, higher performance improvement is achieved. Thus, the proposed method is better to be applied to such sensor network.**

I. Introduction

In a wireless sensor network, sensor data messages containing sensor data achieved by sensor nodes with wireless communication devices are carried to a sink node. Generally, there are no continuous battery supplies to the sensor nodes and they work only by using limited capacity. Thus, reduction of battery consumption is critical to realize long-life sensor networks. Especially, battery consumption in communication modules is required to be reduced and various methods for intermittent communication such as X-MAC [1], B-MAC [6] and S-MAC [9] have been proposed. In addition, in order to carry the sensor data messages to the sink node even with low transmission signal power in wireless sensor nodes, wireless multihop transmissions in which the sensor nodes contribute to configure a wireless multihop transmission route as intermediate nodes have been introduced. Since wireless communication is intrinsically based on broadcast transmissions, a wireless signal transmitted for a data message to one of the neighbor nodes also reaches to all the neighbor nodes. Thus, due to collisions of wireless signals caused by the exposed and hidden node problems, reachability of data messages gets lower and the transmission delay of data messages gets longer by retransmissions. Thus, in most of widely available wireless LAN protocols such as IEEE 802.11, Bluetooth and ZigBee, CSMA/CA and RTS/CTS controls are introduced for collision avoidance and improve reachability and reduce the transmission delay.

In wireless sensor networks, the following two kinds of sensor data messages are transmitted from each source sensor nodes to the sink node:

- periodical, regular and constant sensor data messages
- temporary, unexpected and urgent sensor data messages

The former is for transmission of sensor data achieved by usual and periodical observation in sensor nodes. The amount of sensor data is usually small and there are no strict deadlines for the transmissions. Most of such sensor data are stored into databases after receipts by the sink node. For transmissions of these messages, the sensor network is designed to have enough capacity. On the other hand, the latter is for transmission of sensor data achieved in case of an urgent event and a sequence of sensor data messages are required to be transmitted in short period. Such sensor data are required to be used soon after being receipt by the sink node. Thus, the transmission deadline is strict and shorter delay transmission is mandatory. Therefore, such event-driven sensor data messages are treated as high priority ones. Needless to say, collisions of the event-driven sensor data messages should be avoided and transmissions of the other low prioritized sensor data messages should be suspended temporarily. This causes longer transmission delay of low priority sensor data messages. In addition, more low priority sensor data messages are stored in communication buffers in intermediate sensor nodes for longer period which causes reduction of transmission performance of the wireless sensor network. This paper proposes a novel transmission method of both high and low priority sensor data messages to solve these problems.

II. Related Works

There are some methods to achieve shorter transmission delay, higher throughput and higher reliability of high priority data messages by introduction of priority-based protocols in wireless networks. In IEEE 802.11e [7], [8], EDCA (Enhanced Distributed Channel Access) which introduces AIFS (Arbitration Inter Frame Space) and control of range of contention windows based on priority of transmitted data messages and HCCA (Hybrid Coordination Function Control Channel Access) which introduces TXOP (Transmission Opportunity) while a dedicated wireless node monopolize to transmit data messages for a determined period [5].

The above methods are for 1-hop wireless transmissions of prioritized data messages. MACA/PR protocol [4] is for

prioritized wireless multihop transmissions of data messages. For transmissions of a sequence of high priority data messages, while a head data message is transmitted from a source sensor node to a destination sink node, required reservation windows in intermediate nodes and 1- and 2-hop neighbor nodes of them are reserved. Thus, the required wireless communication media is monopolized and the sequence of the high priority data messages are exclusively transmitted without collisions and contentions with the other low priority data messages. The low priority data messages are only transmitted by using the other free reservation windows. In [3], another exclusive transmission method with an extension of TXOP to wireless multihop transmissions is proposed.

As in these methods, collisions and contentions between transmissions of high prioroty data messages and low priority ones are avoided by excusive usage of wireless transmission media around a wireless multihop transmission route by a sequence of high priority data messages. This results in shorter transmission delay and high reaciabliity of high priority data messages. However, these methods do not consider about the degradation of performance of low priority data message transmissions. They are inevitable both should be decreased.

In most wireless sensor networks, both periodical and event-driven sensor data messages are transmitted to the sink node. Since the event-driven sensor data messages usually carry urgent sensor data and need to be transmitted within a certain deadline, these are treated as the high priority messages and it is reasonable to apply such a priority-based wireless multihop transmission protocol to occupy the wireless transmission media. On the other hand, the periodical sensor data messages are not required to be transmitted with a strict deadline, too long transmission delay is not suitable since these low priority sensor data messages fill communication buffers in some intermediate sensor nodes which causes degradation of the wireless network especially for urgent transmissions of high priority event-driven sensor data messages. Thus, even for the low priority sensor data messages, transmission delay should be lower even with the monopolized transmissions of the high priority sensor data messages.

III. Proposal

This section proposes transmission protocols for high priority sensor data messages which occupy wireless transmission media for exclusive transmissions and for low priority sensor data messages which detour dynamically for parallel transmissions with the transmissions of the high priority sensor data without collisions and contentions.

A. Protocol for High Priority Messages

Suppose that a sequence of event-driven burst high priority sensor data messages are transmitted from a source sensor node N^s to a sink node N^d along a wireless multihop transmission route $R := ||N_0(= N^s) \ldots N_n(= N^d)\rangle\rangle$ as in Figure 1. Here, 1-hop neighbor nodes of R which is in a wireless signal transmission range of at least one of the intermediate sensor node N_i in R should suspend transmissions of both data messages and control messages for avoidance of collisions and contentions with the transmissions of the high priority sensor data messages. Thus, low priority sensor data messages stored in communication buffers in the intermediate sensor nodes and 1-hop neighbor sensor nodes should wait for termination of the sequence of the high priority sensor data messages as in Figure 2. In most widely available wireless LAN protocols, hop-by-hop acknowledgement is applied for data message transmissions for required reliability. That is, for data message transmission from N_i to N_{i+1}, an ack control message is transmitted for confirmation and combination of timeout and retransmissions provides the required reliablity (Figure 3). In addition, in spite of the predetermined number of retransmissions of a sensor data message, if N_{i+1} does not receive the sensor data message, i.e., N_i does not receive an ack control message, N_i discards the sensor data message. Since 1-hop neighbor sensor nodes of R are prohibited to transmit not only data messages but also control messages such as ack control messages, 2-hop neighbor nodes of R whose next-hop node is a 1-hop neighbor node of R do not receive an ack control message and discard the transmitting sensor data message as in Figure 5.

In the TCP/IP internet, data messages are retransmitted in an end-to-end manner as in TCP, it is difficult or impossible to apply the end-to-end retransmissions in a wireless sensor network, since the end-to-end retransmission requires end-to-end confirmation of receipt by an ack control message and buffering of data messages is required for retransmissions even though wireless sensor nodes are usually lack of communication buffers [2].

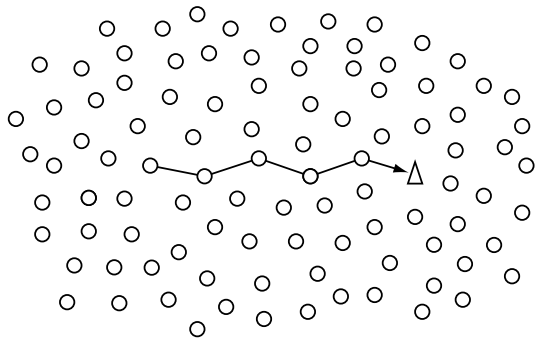


Fig. 1. High Priority Transmission Route R.

Thus, in order to exclusive transmissions of a sequence of high priority sensor data messages along R, control messages to request to suspend any data and control messages are transmitted to 1-hop neighbor nodes of R and other control messages to suspend data message transmissions to the 1-hop neighbor nodes of R are also transmitted before beginning of a sequence of sensor data messages along R. Therefore, the high priority sensor data messages are transmitted without collisions and contentions with low priority sensor data messages and low priority sensor data messages are never lost due to waste transmissions for which no ack control message returns are expected.

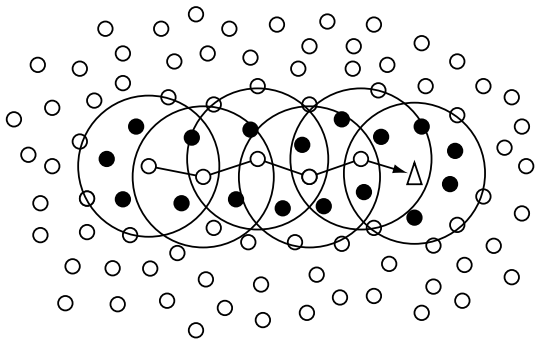


Fig. 2. 1-Hop Neighbor Nodes of R.

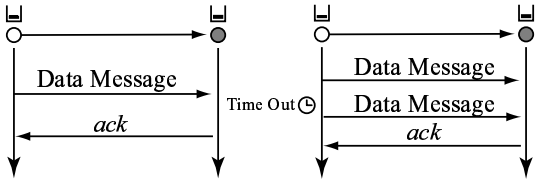


Fig. 3. Retransmissions with ack and Timeout.

B. Protocol for Low Priority Messages

According to the proposed protocol in the previous subsection, 2-hop neighbor sensor nodes suspend to transit its current holding sensor data messages to their next-hop nodes if they are 1-hop neighbor nodes of R. However, the suspension time is relatively long period for frequency of low priority data messages due to a long sequence of the event-driven high priority sensor data messages, too many sensor data messages are required to be stored temporarily in the communication buffer in the 2-hop neighbor node of R. However, the capacity of the communication buffer is not always enough in a wireless sensor node and the communication buffer may overflow with the low priority sensor data messages and some messages might be discarded as shown in Figure 6.

.

In order to solve this problem, for a sensor data message transmission from N_i to N_{i+1}, N_{i+1} explicitly sends back a $nack$ (negative acknowledgement) control message to N_i if its communication buffer is full as shown in Figure 8. In case that

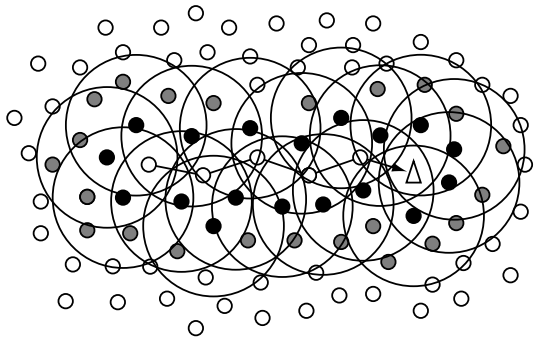


Fig. 4. 2-Hop Neighbor Nodes of R.

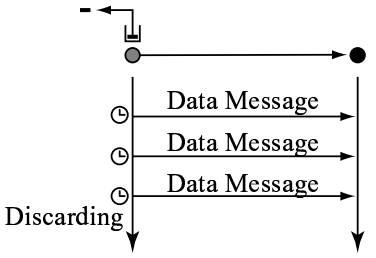


Fig. 5. Message Discarding by No ack Returns.

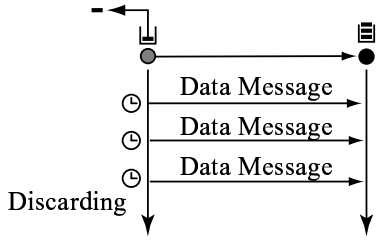


Fig. 6. Message Discarding due to Buffer Overflow.

N_i transmits a low priority sensor data message to N_{i+1}, N_i stores or keeps the sensor data message into its communication buffer until N_i receives an ack control message from N_{i+1}. If N_i receives a $nack$ control message from N_{i+1} notifying a full communication buffer in N_{i+1}, N_i suspends its sensor data message transmissions to N_{i+1}. This avoids buffer overflow in N_{i+1}; however, a sequence of filled communication buffer is configured as shown in Figure 7.

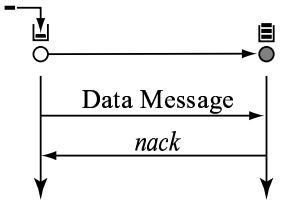


Fig. 7. Avoidance of Discarding Messages by $nack$.

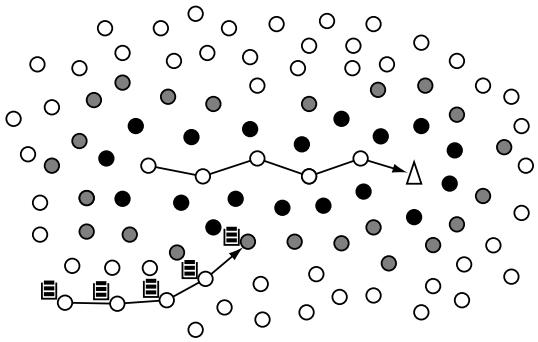


Fig. 8. A Sequence of Filled Communication Buffers.

Even though the sequence of high priority sensor data messages have been transmitted and suspended transmissions of low priority sensor data messages restart, it requires longer

transmission delay for buffered low priority sensor data messages to reach the destination sink node. The two main reasons are as follows:

- Low priority sensor data messages are not possible to be forwarded to the next-hop nodes only when there are free communication buffers in the next-hop nodes.
- For each intermediate sensor node, its previous-hop and next-hop nodes are both its exposed nodes and its 2-hop previous and 2-hop next nodes are both hidden nodes. Thus, it is impossible to transmit sensor data messages simultaneously with these nodes.

All the intermediate sensor nodes with filled communication buffer try to forward their buffered sensor data messages, they tends to cause collisions and contentions which makes transmission delay longer. Especially, contentions with its 1- and 2-hop previous nodes and the above request for free communication buffer in its next-hop nodes are contradict one another.

In order to solve this longer transmission delay problem, low priority sensor data messages are transmitted along a detour wireless multihop transmission route whose intermediate nodes are never suspended their transmissions due to the exclusive transmissions of high priority sensor data messages concurrently. This makes low priority sensor data messages carried to intermediate sensor nodes nearer to the sink node than the naive buffering along the default route which makes a sequence of filled communication buffers. By using flooding based routing protocols to the dedicated destination sink node such as TORA [] ad-hoc routing protocol, all the sensor nodes achieves its hop counts to the sink node. By comparison of the hop counts with the neighbor sensor nodes, each sensor node determines its candidate of its next-hop nodes. If it cannot transmit a low priority sensor data message to a part of them which is detected by receipt of a $nack$ control message notifying a filled communication buffer, it forwards the message to one of the other available neighbor nodes as shown in Figure 9 and 10.

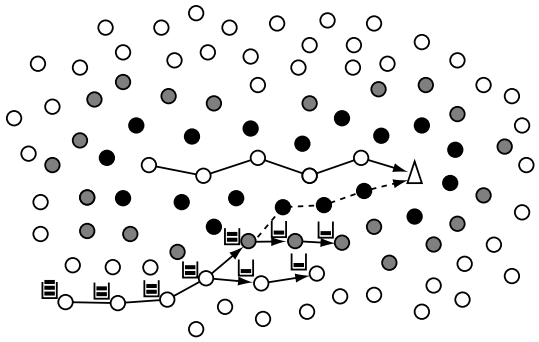


Fig. 9. Detour of Low Priority Messages.

By applying this method, the low priority sensor data messages never configure a sequence of filled communication buffers and are transmitted to other sensor nodes nearer to the sink node and stored into their communication buffers. After termination of the transmission of high priority sensor data

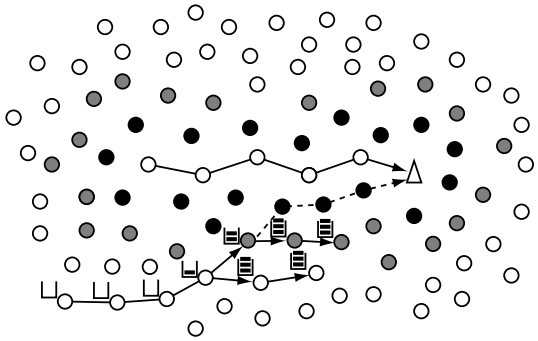


Fig. 10. Buffering of Detour Low Priority Messages.

messages, buffered low priority sensor data messages reach the destination sink node with shorter transmission delay. This is because the low priority sensor data messages are transmitted nearer to the sink node and do not configure a sequence of filled communication buffers since they are stored into buffers located on a concentric circle as in Figure 11.

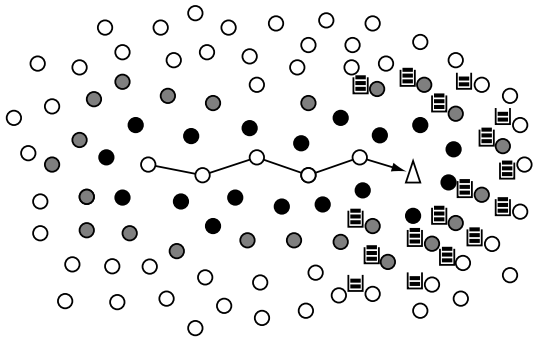


Fig. 11. Filled Communication Buffers near Sink Node.

IV. Evaluation

This section evaluate the performance of transmissions of low priority sensor data messages in our proposed protocol in simulation experiments. Here, 500 wireless sensor nodes with 100m wireless signal transmission ranges are randomly distributed in a 2,000m × 500m area. To make clear the difference of performance improvement caused by the location of a sink node, pairs of locations of the source sensor node of the high priority event-driven sensor node and the sink node are ((-1,000m, 0m), (0m, 0m)), ((-900m, 0m), (100m, 0m)), …, ((0m, 0m), (1,000m, 0m)) in the coordinate shown in Figure 12. The high priority event-driven sensor data messages are transmitted for 5 second whose transmission rate is 20 packets per second and the low priority sensor data messages are transmitted with an interval of 15 seconds by randomly selected 50–100 sensor nodes. For evaluation, the transmission delay is compared with the naive transmissions in which all the sensor data messages are transmitted along their default route without detour and only $nack$ transmissions are introduce to avoid loss of sensor data messages caused by communication buffer overflow.

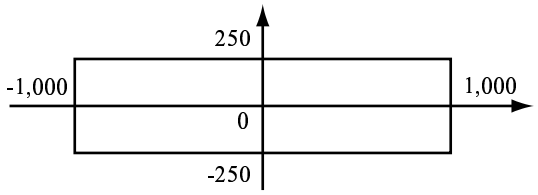


Fig. 12. Simulation Field.

Figure 13 shows the simulation result.

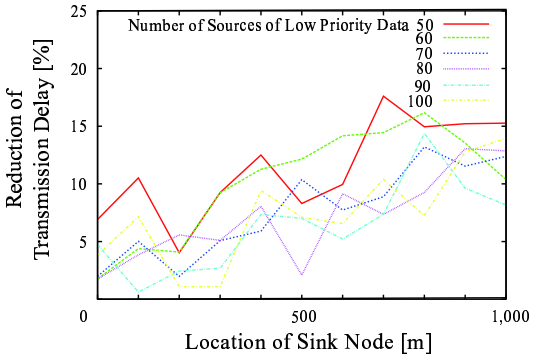


Fig. 13. Transmission Delay of Low Priority Messages.

In any simulation settings, the proposed protocol achieves shoter transmission delay than the conventional protocol. The proposed protocol carries buffered low priority sensor data messages in shorter time to the sink node and less sensor data messages remains in communication buffers after termination of high priority sensor data message transmissions. The simulation result shows that the performance improvement depends on the locations of the sink node. If the sink node is far from the center of the simulation field, i.e., if it locates near the edge of the simulation field, the transmission delay becomes shorter. This is because the detour concurrent transmissions of the low priority sensor messages effects more if they are transmitted from the directions same as that of the high priority sensor data messages. Thus, our proposed protocol works better in wireless sensor networks where sink nodes are located on the edge of the network.

V. Conclusion

For better transmissions of prioritized sensor data messages, high prioritized event-driven urgent sensor data messages are exclusively transmitted by suspension of transmissions of 1- and 2-hop neighbor sensor nodes. To reduce the performance degradation of transmissions of the low priority sensor data messages, they are transmitted along a detour route where no collisions and contentions are caused. This avoids configuration of filled communication buffers and carries low priority sensor data messages nearer to the sink node. The simulation experiments show it achieves 3.49–12.2% shorter transmission of low priority sensor data message after terminations of high priority sensor data messages.

References

[1] Buettner, M. Yee, G.V., Anderson, E. and Han, R., "X-MAC: A Short Preamble MAC Protocol for Duty-Cycled Wireless Sensor Networks," Proceedings of the ACM SenSys, pp. 307–320 (2006).

[2] Kaneko, Y. and Higaki, H., "Ad-Hoc Buffering in Neighbor Nodes for Burst Data Transmissions in Wireless Sensor Networks," Proceedings of the 3rd International Workshop on Future Information System Technologies and Applications, pp. 14–19 (2012).

[3] Kieu, P.K., Miyamoto, S., "A Study on QoS Guaranteed Transmission Protocol for Multi-hop Wireless LAN System," IEICE Technical Report, RCS2006-15, pp. 85–90 (2006).

[4] Lin, C.R. and Gerla, M., "Real-Time Support in Multihop Wireless Networks," Wireless Networks, Vol. 5, pp. 125–135 (1999).

[5] Mangold, S., Choi, S., Klein, O., Hiertz, G. and Stibor, L., "IEEE 802.11e Wireless LAN for Quality of Service," Proceedings of the European Wireless, Vol. 1, pp. 32–39 (2002).

[6] Polastre, J., Hill, J. and Culler, D., "Versatille Low Power Media Access for Wireless Sensor Networks," Proceedings of the ACM SenSys, pp. 95–107 (2004).

[7] Tran, S.N.H. and Lee, W., "QoS Provisioning in IEEE 802.11 Wireless LANs," Proceedings of the 1st International Conference on Communications and Electronics, pp. 23–28 (2006).

[8] Xiao, Y., "Enhanced DCF of IEEE 802.11e to Support QoS," Proceedings of the IEEE Wireless Communications and Networking Conference, Vol. 2, pp. 1291–1296 (2003).

[9] Ye, W., Heidemann, J. and Estrin, D., "An Energy-Effi cient MAC Protocol for Wireless Sensor Networks," Proceedings of the IEEE INFOCOM, Vol. 3, pp. 1567–1576 (2002).

Security in Wireless Sensor Networks Using Coalition Games with Transferable Payoff

Mehran Asadi[1], Afrand Agah[2], and Christopher Zimmerman[2]
(Corresponding author: Afrand Agah)
Center of Excellence in Business and Entrepreneurial Studies , Lincoln University[1]
Lincoln University, PA 19352
Department of Computer Science, West Chester University[2]
West Chester, PA 19383
(Email: aagah@wcupa.edu)

Abstract

In this paper, we investigate the impacts of using game theory in recognizing malicious nodes and voltage loss in a wireless sensor network. All nodes in the wireless sensor network are divided into several clusters, where in each cluster one node acts as the cluster head. All cluster heads are in charge of receiving and forwarding packets to each other and to the base station. Here, we study the model of a coalition game. One primitive of this model is the collection of sets of joint actions that each group of players (clusters) can take independently of the remaining players. Our goal is maximizing the correct percentage of malicious nodes while minimizing the voltage loss.

1 Introduction

Game theory attempts to model decision making which has been used in various fields such as economics, politics and biology [21]. Game theory has previously applied to wireless sensor networks, but within the context of modeling multiple nodes in the network attempting to share a shared medium: their radio communication channels [10].

We use game theory to help the sensors optimize their decision making process about whether or not to forward any data packets they may receive [1, 2, 3]. On one hand, if a node decides to never forward any packets, it conserves its battery power, but no data flows through the network. However, if a node forwards every packet that it receives, that node demonstrates its reliability and traffic flows through the network but the node will run out of battery power much faster than if the node were to not forward any packets. By using game theory, we attempt to find an optimum configuration that will extend a node's battery life while still allowing the node to forward an acceptable amount of packets through the network [4].

In this work, we assume that the network consists of groups of sensors called clusters, where the sensors in each cluster report to a sensor in a cluster that is designated as the cluster head of that cluster in the network. All non cluster heads within a cluster, known as members of a cluster, only communicate directly with their respective cluster heads. Cluster heads transfer data to the base station where the data is to be collected and stored. In our simulations, the process of determining which cluster head a sensor reports to is based on the sensor's battery level. If a member receives a packet broadcast by a cluster head within range, the member will forward that packet directly to its cluster head.

2 Related Work

Defining a suitable cost and profit to routing and forwarding incoming packets and keeping a history of experiences with non-cooperating nodes drives malicious nodes out of the wireless sensor network. Reputation systems are being used in many systems to provide a means of obtaining a quality rating of participants of transactions by having all parties give each other feedback on how their activities were perceived and evaluated [18, 19]. In order to avoid centralized rating, local lists are maintained at each node and nodes can look up senders in their blacklist containing any node with a bad rating before forwarding anything to them [14].

There is a trade-off between good cooperation and resource consumption; therefore nodes have to economize on their resources. At the same time, however, if they

do not forward messages, others might not forward either, thereby denying service. Total non-cooperation with other nodes and only exploiting their readiness to cooperate is one of several boycotting behavior patterns. Therefore, there has to be an incentive for a node to forward messages that are not destined to itself [7, 8, 9, 13].

The performance of the network can reach an undesirable state due to the selfish behavior of individual wireless nodes. Therefore, incentives are proposed to steer nodes towards desirable operational equilibrium of the network behavior.

Often node decisions at a particular layer are made with the objective of optimizing performance at some other layer, therefore game theory can provide insight into approaches for optimization. It allows us to investigate the existence, uniqueness and convergence to a steady state point when network nodes perform independent adaptations. It helps us to design incentive schemes that lead to independent, self-interested participants towards outcomes that are desirable from a system-wide point of view [15].

We use a reputation system for incentivizing. Each node gains reputation by providing services (forwarding incoming packets) to others [16, 17]. Each node builds a positive reputation for itself by cooperating with others and is tagged as selfish or malicious otherwise. Reputation is maintained as a probabilistic distribution, enabling the node to have full freedom and not get constrained by some discrete levels of reputation as used in eBay, Yahoo auctions [22]. Note that reputation is not a physical quantity but it is a belief; it can only be used to statistically predict the future behavior of other nodes and cannot define deterministically the actual action performed by them [11].

Recent technological advances within the field of wireless sensor networks have made it possible to support long-lasting operating lifetimes and large amounts of data transmission in wireless sensor networks. A major challenge is to maximize the lifetime of these battery powered sensors to support such transmissions. Battery powered sensors might waste a huge amount of energy if we do not carefully schedule and budget their discharging [6, 20].

3 Game Formulation

In our previous work [4, 5], we studied the sets of possible actions and their preferences over the possible outcomes for each node in a wireless sensor network, where an outcome is a profile of actions.

In this paper we study the model of a coalition game. Here we focus on what groups of players can achieve rather than what individual players can do. In this work we will not discuss how coalitions forms. Each of a group of individuals, which is a cluster of wireless sensor nodes owns some inputs (data that is sensed and thus generated locally) and has access to resources (a cluster head) for producing a valuable single output, which in this case is packets that they need to forward to the next cluster head or the base station. A coalition game starts with the sets of payoff vectors that each group of individuals can jointly achieve. The main goal is a solution for this game that requires that no set of players be able to break away and take a joint action that makes all of them better off.

We assume that each group of players or a cluster is associated with a single number, which is the payoff that is available to the group, there is no restrictions on how this payoff may be divided among the members of each cluster.

Our goal is to study the coalition of joint actions that group of players (wireless sensor networks) can take independently. A solution concept for coalition games assigns a set of outcomes to each game. Each solution is the consequences of a line of reasoning for the players in a game. Thus the outcome must be immune of any deviations.

A coalition game consists of a finite set of N (the set of wireless sensor nodes), a function u that associates with every nonempty subset S of N (a coalition or a cluster) a real number $u(S)$. For each coalition S the number $u(S)$ is the total payoff that is available for division among the members of S. Here we are also assuming that actions of the players who are not part of S do not influence $u(S)$.

As with any Nash equilibrium of a non cooperative game, an outcome is stable if no deviation is profitable [21]. An outcome is stable if no coalition can deviate and obtain an outcome better for all its members. Therefore no coalition can obtain a payoff that exceeds the sum of its members' current payoffs. Thus, the solution is the set of feasible payoff profiles $(x_i)_{i\in N}$ for which $u(S) <= x(S)$ for every coalition S, where $(x_i)_{i\in N}$ are real numbers and for any coalition S we let $x(S) = \Sigma_{i\in S} x_i$.

In each unit of time, each node in the wireless sensor network might be active (forwarding or sensing data) so that coalition will have a payoff. A game is played by the cluster heads, where in each cluster a node will be selected to serve as a cluster head. Let game G be played several times and let us award each cluster head node a payoff which is the sum of the payoffs it received from the nodes in it's cluster in each period, by playing G.

At time t, each cluster head node calculates the utility to be gained for each of the two actions available, forwarding or not forwarding. For forwarding a packet, the utility is calculated as:

$$u^t = \omega * r_i^{t+1} - \beta * (c_s + c_r),$$

where r_i^{t+1} is the predicted gain of node i's reputation, c_s is the voltage cost to send a packet and c_r is the voltage

cost to receive a packet. β is the weight parameter for cost, and represents the importance of being conservative about sending packets when a node has a low battery level. At a node's highest battery level, β will be 1. As the node's battery level crosses designated thresholds by decreasing, β will increase.

ω is the weight parameter for the gain component of the equation and represents the number of units of time since node i has last forwarded a packet. ω starts at 1 for each node i and increments every time any node i decides to not forward a packet. When a node sends a packet, ω is reset back to 1. If a node has recently sent a packet, it may not be important to send another packet right away, which is why ω starts at a low value. But as time passes without forwarding any packets, it is important that a node sends data through the network, which leads ω to increase.

The utility for not forwarding a packet is calculated as:

$$u^t = \omega * 0 - \beta * c_s.$$

Since there is no gain in reputation when not sending a packet, the gain is 0. However, receiving a packet from another cluster head node still costs energy.

After calculating the utility for each of these actions, the node will perform the action that yields the greater utility. In order to compute the values of a node's gain, we turn our attention to the work proposed in [16]. In this work the authors proposed the concept of subjective reputation, which reflects the reputation calculated directly from the subject's observation. In order to compute each node's reputation at time t, we use the following formula:

$$r_i^t = \sum_{k=1}^{t-1} \rho_i(k)$$

where $\rho_i(k)$ represents the ratings that the base station has given to node i, and $\rho_i \in [-1, 1]$. If the number of observations collected since time t is not sufficient, the final value of the subjective reputation takes the value 0. The base station increments the ratings of nodes on all actively used paths at periodic intervals. An actively used path is one on which the node has sent a packet within the previous rate increment interval. Recall that reputation is the perception that a person has of another's intentions. When facing uncertainty, individuals tend to trust those who have a reputation for being trustworthy. Since reputation is not a physical quantity and only a belief, it can be used to statistically predict the future behavior of other nodes and can not define deterministically the actual action performed by them.

3.1 Coalition Game with Transferable Payoff

In the wireless sensor network at each unit of time, we have two sets. The sets of cluster head nodes that have data to transfer, lets denote them by L_1 and the set of cluster head nodes that are willing to transfer the data L_2. Each node that has data to transfer holds one unit of the data and has a reservation price of 1; each node that is willing to transfer one unit of data has the reservation price of 0. Here one unit of data is one packet of data.

We can model this network as a coalition game as $N = L_1 \cup L_2$ and $u(S) = min\{|S \cap L_1|, |S \cap L_2|\}$ for each coalition S. If $|L_1 < L_2|$ then the solution consists of the single payoff profile in which every node who is willing to transfer receives 1 and every node that has data to transfer receives 0. Suppose that payoff profile x is a solution for the game. Let l_1 be a node that is willing to transfer data whose payoff is minimal among the payoffs of all those who are willing to transfer and let l_2 be the node that has data to send whose payoff is minimal among the payoffs of all those who have data to transfer. Since x is the solution we have

$$x_{l_1} + x_{l_2} >= u(\{l_1, l_2\}) = 1$$

and

$$L_2 = u(N) = x(N) >= |L_1|_{x_{l_1}} + |L_2|_{x_{l_2}},$$

$$L_2 = u(N) >= (|L_1| - |L_2|)x_{l_1} + |L_2|.$$

This implies that $x_{l_1} = 0$ and $x_{l_2} >= 1$ and hence (using $u(N) = |N|$ and the fact that l_2 is the worst-off node to have data to transfer) $x_i = 1$ for every node i that has data to transfer.

4 Performance Evaluation

In our simulations, the process of determining which cluster head a sensor reports to is based on the sensor's battery level. If a member receives a packet broadcasted by a cluster head within range, the member will forward that packet directly to its cluster head. For each configuration, simulations are run both with and without implementing game theory. By doing so, we can compare average network throughput, as well as voltage loss, and see under which scenario using one would be favorable over the other. The sensors are programmed identically for the game theory and non game theory configurations. The networks are either entirely comprised of nodes that utilize game theory or entirely of nodes that do not utilize game theory. A node that acts maliciously is one that randomly drops packets in order to conserve its energy. For

malicious node that are not utilizing game theory, before forwarding a packet the node randomly decides whether or not it wants to not forward the packet. For malicious nodes that are utilizing game theory, the node randomly decides whether or not it wants to not forward the packet before the strategy is applied. All tests are run with a network size of 30 nodes.

We introduce malicious nodes into the network to see how they affect the network and if there is a way to detect and neutralize such nodes. Malicious nodes randomly drop packets, reducing the throughput of the network. Malicious nodes also consume additional power when randomly deciding whether or not to drop packets. The base station keeps track of the reputation of each node in the network. Periodically, the base station will decide whether or not a node is acting malicious based on its throughput. The base station takes the current reputation of each node in the network and calculates the average, as well as the standard deviation. If a node's reputation is lower than the average minus the standard deviation, that node is deemed malicious. The base station sends a packet to that node ordering the node to turn its radio off and shut down.

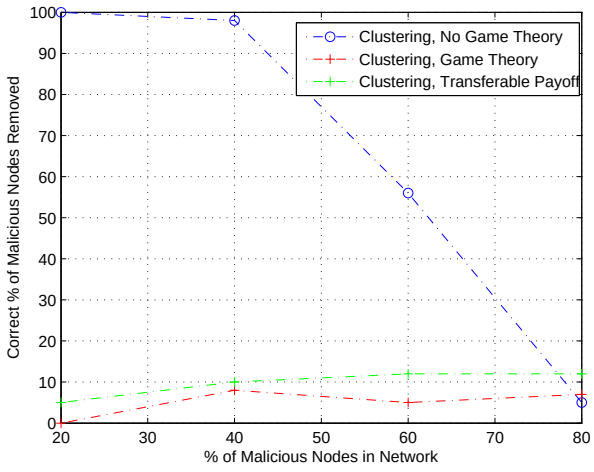


Figure 1: Average percentage of malicious nodes correctly removed from network.

In this work we have used MICAZ sensors [12], which run on TinyOs [23]. A packet is broadcast once every 200 milliseconds for 300 seconds, then a final voltage reading is sent to the base station. An initial voltage reading is sent from each sensor. Next, cluster membership is established for each non cluster head node in the network. After that, a packet is broadcast once every 200 milliseconds for 300 seconds. Afterward, a final voltage reading is sent to the base station. During the simulation, if a node receives a packet it will forward it or apply the game theory strategy, depending on the scenario. After sending the packets, each node turns its radio off for 10 seconds to get rid of the traffic in the network. Then, every node turns their radio on and sends one final voltage packet to the base station. This gave us a clear start and end voltage for calculating voltage loss. For detecting malicious nodes, the base station checks to see if any of the nodes are malicious after 60 seconds into the simulation, and then once every 30 seconds after that. Any nodes that are deemed as malicious are turned off via radio.

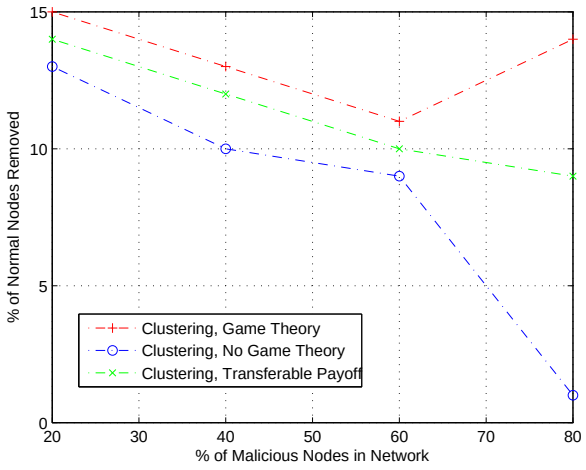


Figure 2: Average percentage of normal nodes correctly removed from network.

As indicated by figure 1, our procedure for detecting malicious nodes works best for networks containing small amounts of malicious nodes. Since malicious nodes usually have a lower reputation, if there is a small number of them present in the network, it is easier to detect them. However, if normal nodes in a network typically have low reputations, or if many nodes in the network have lowered reputations, it is difficult to detect malicious nodes because they don't stand out. This is true whether we utilize game theory or not.

Figure 2 shows that our procedure for detecting malicious nodes raises a low percentage of false positives, as a small amount of false positives are raised. Since there are less normal nodes in the network as the number of malicious nodes increases, the percentage of false positives detected increases. By not utilizing game theory, the percentage of false positive decreases, because reputations of nodes that use game theory decreases quicker than nodes that are not utilizing game theory and therefore base station marks them malicious more frequently. Also if we use game theory with transferable payoffs, then we have less false positives because as long as nodes have battery available they will forward incoming packets.

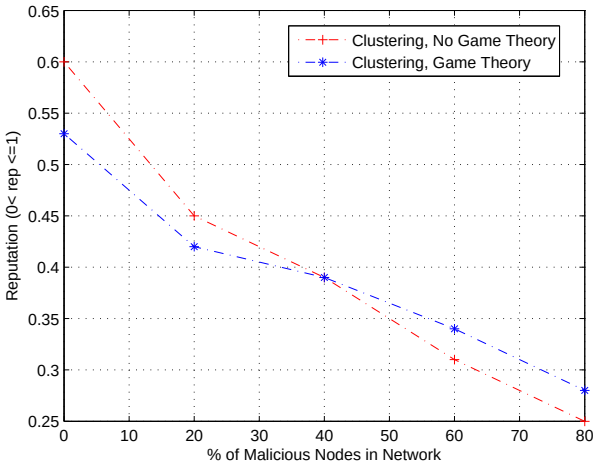


Figure 3: Average reputation in the network consisting of malicious nodes.

As seen in Figure 3, average reputations for cases using game theory are higher than non-game theory cases, as reputation is higher by using game theory in networks with a larger percentage of malicious nodes. As seen in Figure 4, in most cases, voltage loss is lower with game theory implemented than if not, even with the presence of malicious nodes.

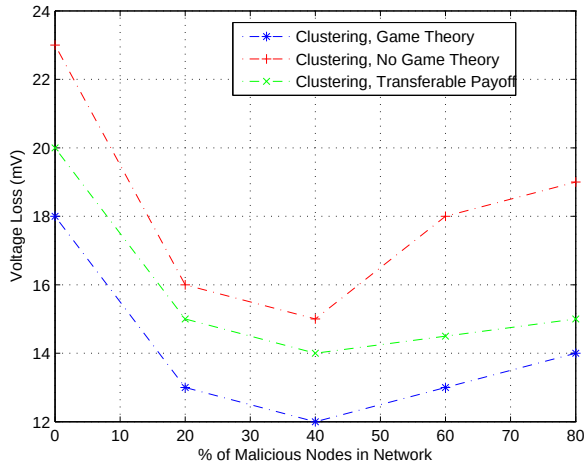


Figure 4: Average network voltage loss for malicious nodes.

5 Conclusion and Future Work

A coalition model is distinguished from a non cooperative model by its focus on what groups of players can achieve rather than individual players can do. We wish to model the possibility of coalition formation in a non cooperative game and see how members of coalitions choose joint actions.

Acknowledgements

This work is supported by the National Science Foundation under grant number 1054492.

References

[1] A. Agah, M. Asadi, and C. Zimmerman, "Maximizing battery life: Applying game theory to wireless sensor networks," *The WCU Research Consortium*, 2011.

[2] A. Agah, S. K. Das, and K. Basu, "Enforcing security for prevention of dos attack in wireless sensor networks using economical modeling," *Proceedings of the 2nd IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)*, 2005.

[3] A. Agah, S. K. Das, and K. Basu, "Preventing dos attack in sensor and actor networks: A game theoretic approach," *IEEE International Conference on Communications (ICC)*, 2005.

[4] M. Asadi, C. Zimmerman, and A. Agah, "A quest for security in wireless sensor networks: A game theoretic model," *The International Conference on Wireless Networks (ICWN)*, 2012.

[5] M. Asadi, C. Zimmerman, and A. Agah, "A game-theoretic approach to security and power conservation in wireless sensor networks," *The International Journal of Network Security (IJNS)*, vol. 15, no. 1, pp. 50–58, 2013.

[6] M. Chi and Y. Yang, "Battery-aware routing for streaming data transmissions in wireless sensor networks," *Springer Science and Business Media, LLC*, 2006.

[7] S. Eidenbenz, L. Anderegg, and R. Wattenhofer, "Incentive-compatible, energy-optimal, and efficient ad hoc networking in a selfish milieu," *Proceedings of the 40th Hawaii International conference on system Sciences (HICSS)*, 2007.

[8] S. Eidenbenz, V. S. Kumar, and S. Zust, "Topology control games for ad hoc networks," *ACM Mobile Networks and Application*, vol. 11, no. 2, 2006.

[9] S. Eidenbenz, G. Resta, and P. Santi, "The commit protocol for truthful and cost-efficient routing in ad

hoc networks with selfish nodes," *IEEE transactions on mobile computing*, vol. 7, no. 1, 2008.

[10] M. Felegyhazi and J. P. Hubaux, "Game theory in wireless networks: A tutorial," *EPFL - Switzerland, LCA-REPORT*, 2007.

[11] S. Ganeriwl and M. B. Srivastava, "Reputation-based framework for high integrity sensor networks," *ACM workshop on Security of Ad Hoc and Sensor Networks*, 2004.

[12] MEMSIC Inc., "Micaz wireless measurement system," *http://www.memsic.com*.

[13] D. Levin, "Punishment in selfish wireless networks: A game theoretic analysis," *Workshop on the Economics of Networked Systems*, 2006.

[14] Y. Liu and Y. R. Yang, "Reputation propagation and agreement in mobile ad-hoc networks," *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, 2003.

[15] R. Machado and S. Tekinay, "A survey of game-theoretic approaches in wireless sensor networks," *Elsevier Computer Networks Journal*, vol. 52, 2008.

[16] P. Michiardi and R. Molva, "Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks," *Communications and Multimedia Security Conference*, 2002.

[17] P. Michiardi and R. Molva, "Game theoretic analysis of security in mobile ad hoc networks," *Research Report, Institute Eurecom*, 2002.

[18] P. Michiardi and R. Molva, "Prevention of denial of service attack and selfishness in mobile ad hoc networks," *Research Report RR-02-063, (Institute Eurecom)*, 2002.

[19] P. Michiardi and R. Molva, "Simulation-based analysis of security exposures in mobile ad hoc networks," *European Wireless 2002: Next Generation Wireless Networks: Technologies, Protocols, Services and Applications*, 2002.

[20] P. Nurmi, "Modeling energy constrained routing in selfish ad hoc networks," *International conference on Game Theory for networks(GameNets)*, 2006.

[21] M. Osborne and A. Rubinstein, *A Course In Game Theory*. The MIT Press, 1994.

[22] P. Resnick, K. Kuwabarra, R. Zeckauser, and E. Friedman, "Reputation systems: Facilitating trust in e-commerce systems," *Communications of the ACM*, vol. 43, no. 12, 2000.

[23] TinyOS, "Tinyos documenatation," *http://docs.tinyos.net*.

An Energy Efficient Hierarchical Clustering Protocol for Wireless Sensor Networks

Tong Duan, Ken Ferens, and Witold Kinsner
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada
{duant@cc.umanitoba.ca, Ken.Ferens@ad.umanitoba.ca, Witold.Kinsner@ad.umanitoba.ca}

Abstract— This paper presents an innovative hierarchical clustering protocol for wireless sensor networks. In networks that mainly apply multi-hop communications, the huge amount energy consumed by relay tasks of nodes near the sink node cause premature network death, and, so, the lifetime of these nodes needs to be improved efficiently in order to prolong the duration of network service. The aim of the proposed hierarchical clustering design is to minimize energy dissipation difference among these nodes. Furthermore, the hierarchical clustering mechanism reduces transmission delay. The energy efficiency of the proposed algorithm is verified through simulation and it demonstrates that the network lifetime has been significantly extended by 45% compared with LEACH.

Keywords—cluster-based; cross layer; energy efficient; wireless sensor networks; TDMA; multi-hop

I. INTRODUCTION

Wireless sensor networks have gained much interest in the routing protocol research field for more than one decade. It is an emerging technology that benefits from the ongoing developments of sensor techniques, low energy consumption electronics, and low-power radio frequency design [HeCh00]. In recent years, reliable and inexpensive sensors have been extensively utilized in different applications ranging from civil purpose to military applications [RaSc02] [ChCh01]. Though such sensors are not as accurate as expensive sensors, they are still popular because of their high accessibility. Furthermore, in many applications, trading low budget devices for an increased number of devices is more important. But, what is even more important is the Quality of Service (QoS).

Quality of service (QoS) of wireless sensor networks may be evaluated from two points of view. On the one hand, reliability of sensor data can be used to measure quality of service, but this implies the use of accurate and expensive sensors. The use of such high quality sensors may not be a good choice in certain applications, such as forest fire tracking and battle field monitoring, since the sensors may not be retrievable in those scenarios. In these cases, practicality demands the use of inexpensive sensors, and a compromise between QoS and expense must be made. On the other hand, maximum network lifetime is a significant goal of wireless sensor networks. If the nodes in the network prematurely cease to function, then the quality and reliability of sensor readings become irrelevant, since the nodes no longer participate in the network. Extending the lifetime of a WSN has attracted much interest in the research field of energy consumption and distribution of loads.

Since wireless sensor networks are usually deployed in fields where power supply is not available, the consumption of energy becomes the most important issue in order to maximize network lifetime. Some algorithms have been proposed to minimize energy consumption of each sensor node. Ye *et al.* [YeHe04] proposed S-MAC protocol to reduce unnecessary energy consumption by putting nodes into sleep mode when not working. However, this scheme results in delay when one node is trying to communicate with another node while it is in sleep mode, called sleep delay. Dam *et al.* [DaLa03] present a T-MAC protocol to further reduce unnecessary energy consumption. When two nodes are communicating, all neighbor nodes and their surrounding nodes enter sleep mode. Nevertheless, the sleep delay problem is still exhibited here, and it even may be potentially worse than S-MAC, since more nodes are in sleep mode at the same time, which means higher possibility of sleep delay. Furthermore, multi-hop transmission type is applied in wireless sensor networks due to its lower energy cost than single-hop transmission. Kim *et al.* [KiLe09] proposed a cross layer design (ECLP) and energy efficient listening window schedule to minimize energy consumption.

In general, any wireless sensor network that has one sink and applies multi-hop transmission faces the same problem called “hot spot”; this is the problem whereby the nodes near the sink node, especially those which can communicate with the sink node with one hop, may consume much more

energy (*i.e.,* more data relay operations, called one-hop nodes) than nodes that are farther away. This problem exists in many WSN routing protocols such as ECLP and T-MAC. Common sense dictates that the energy consumption distribution must be optimized. Unequal distribution may result in critical disconnections, which means major connection failure (alive nodes cannot reach the sink node) in the wireless sensor networks. Li *et al.* [LiYe05] proposed an energy-efficient unequal clustering mechanism (EEUC). This algorithm was designed to relieve energy cost pressure of nodes near the sink node by assigning distinct size of clusters. Heinzelman *et al.* [HeCh00] proposed low-energy adaptive clustering hierarchy (LEACH); it was intend to relieve this problem by combining clustering technique and single-hop transmission. Both EEUC and LEACH apply single-hop transmission, which may consume more energy in larger scale of wireless sensor networks. This paper proposes a hierarchical clusters approach that is designed to distribute energy consumption of nodes near the sink node as equal as possible, to reduce the likelihood of critical disconnection. Moreover, the proposed protocol aims to balance quality of service and energy consumption; a new approach of how to utilize single-hop transmission is introduced.

The remaining parts of this paper are organized as follows. The next section describes an innovative hierarchical clustering design and a few important elements. This is followed by a detailed description of different network stages. The simulation and analysis results are given in Section IV. Finally the conclusions and recommendations for future works are given.

II. HIERARCHICAL CLUSTERING

The algorithm creates hierarchical layers of clusters of nodes. Each cluster is assigned a distinct level ID, which represents the cluster-hop count between sink node and the particular cluster. Member nodes of a cluster are assigned the same level ID as the cluster in which they belong. During the data communications stage, cluster heads from higher layer clusters relay data to lower layer cluster heads until the messages are received by the sink, which is located at the lowest layer cluster. In order to improve transmission efficiency, the transmission scheme is a combination of multi-hop and single-hop routing.

The hierarchical clustering algorithm is explained in four sections. The first part describes the motivation and overview architecture, which is followed by the details of packets used for synchronization. The third section presents the procedure of hierarchical clustering. Once the whole network is set up, a description of the data communications phase is given in the last section.

In this paper, the network has similar configuration as standard wireless sensor networks. There are V stationary nodes in the wireless sensor network:

$$V = N \cup S \quad (1)$$

The N is the number of sensor nodes, and S is the number of sink nodes. They are randomly distributed in the target field. To simplify the simulations, we assume one sink node, but many potential source nodes are contained in the network. Each sensor node is able to determine its approximate relative distance from other nodes by analyzing received signal strength. All nodes have the same capability and initial energy. Each node has two communications radios to allow either multi-hop or single-hop message passing, but, the single-hop option is the default mode. However, the transmission range is fixed for either case.

A. Overview of Hierarchical Clustering Design

The core idea of this hierarchical clustering design is to generate an energy efficient hierarchical cluster topology for the network. This approach is motivated by the "hot spot" problem, which is faced by many other routing protocols. It is an inevitable situation for such network configuration. The approach used to relieve this problem is to minimize energy consumption difference among cluster heads in each layer; especially those near the sink nodes (one-hop nodes), which take-on most of the burden of this problem. This approach automatically serves to prolong network lifetime. Furthermore, the proposed algorithm also intends to reduce general transmission delay, which includes sleep delay and hop delay.

In order to simplify the description, the clusters which have level ID of *n* are referred to as *Level n* clusters. In the network, each node can hold at most two level IDs, which are *CM_Ln* and *CH_L(n+1)*. The *CM_Ln* means the node is a cluster member of *Level n* cluster, and *CH_L(n+1)* means it is a cluster head of *Level n+1* cluster.

As described, nodes closer to sink node are more likely to consume more energy during service. This condition also applies to clusters near the sink node. Cluster heads dissipate energy with higher rate than cluster members, due to their responsibility of managing and gathering data from cluster members. Furthermore, cluster heads near the sink node bear more relay tasks than those far away from the sink node. Therefore, this paper proposes hierarchical clustering design to balance energy dissipation by cluster heads in each layer to maximum network lifetime.

The proposed protocol consists of hierarchical clusters that mainly perform multi-hop data communications. However, such links can be terminated whenever a relay node on the link breaks down. To reduce data loss and improve network robustness against disconnection, this paper presents two options when such situation appears. One way is to choose an alternative transmission link, which may be considered as a fast recovery solution. Another way is to transmit data directly to the sink node with single-hop transmission mode. Though the former solution may cost more delay than

single-hop, it is selected as preference option in terms of energy consumption.

In wireless sensor networks, the energy consumption for different transmission schemes is calculated as [HeCh02]:

$$\mathrm{E}_{Tx}(l,d) = \begin{cases} lE_{elec} + l\varepsilon_{fs}d^2, & d < d_0 \\ lE_{elec} + l\varepsilon_{mp}d^4, & d \geq d_0 \end{cases} \quad (2a)$$

$$E_{Rx}(l) = lE_{elec} \quad (2b)$$

The E_{elec} is the electronics energy dissipation and it is determined by related operations during service (*e.g.,* modulation/demodulation, filtering, and coding/decoding). The energy consumption of transmit amplifier $\epsilon_{fs}d^2$ and $\epsilon_{mp}d^4$ are exclusively chosen according to the distance threshold d_0. From previous work [HeCh00] [HeCh02], the difference of amplifier energy dissipation is considerable. Therefore, a proper value of d_0 is important for wireless sensor network design.

The proposed protocol consists of two stages, which are network clustering stage and data communications stage. In the network clustering stage, the hierarchical layering starts from the lowest level cluster; *Level 0* cluster is first created and cluster head is the sink node with level ID of 0 (*CH_L0*). Each member in the *Level 0* cluster has a level ID: *CM_L0*. Each *CM_L0* node is able to create, one degree higher, *Level 1* cluster. By this, each *CM_L0* can also have another ID of *Level 1* cluster head (*CH_L1*). In practice, whether a *CM_Ln* becomes a cluster head (i.e., *CH_L(n+1)*) depends on several criteria; a full description is given later in this paper. The reason of such network design is that interference and cluster overlap should be minimized in the network. Since each node is able to calculate approximate distance with another node, it is better to only let cluster members near geometrical boundary of a cluster create higher level clusters. The formation of hierarchical clusters grows like a tree. Starting from the *Level 0* cluster (i.e., *CH_L0* is the sink node), potentially several branches to the next higher level clusters are created. The next level clusters, in turn, create branches to next higher level clusters. This continues until no other higher level clusters can be created.

In the data acquisition stage, time division multiple access (TDMA) technique is applied for intra-cluster communications. The time schedule is denoted by *TS_Ln*, where *n* represents the level of cluster. Cluster head with level ID of *CH_L(n+1)* collects data from its cluster members using *TS_L(n+1)*, and then relays data to *CH_Ln* during time slot *TS_Ln*. The inter-cluster TDMA slot is subdivide into *m* sub-slots for interference-free cluster member communications. By this approach, signal interference is minimized and sleep delay is substantially reduced.

B. Control Packets

During the network clustering stage, two packets, SYNC and $\mathrm{SYNC}_{\mathrm{reply}}$ are utilized. These two packets are similar to SYNC packet in S-MAC [KiLe09] and ECLP [YeHe04]. The proposed protocol uses the SYNC packet for cluster formation, management and synchronization; the $\mathrm{SYNC}_{\mathrm{reply}}$ is used for leaf node confirmation operations. Another essential component for each sensor node is the node management table. The table mainly includes elements that indicate the relation between this sensor node and its neighbor nodes.

Elements of the SYNC packet also inlcude network routing information. *c_id* indicates the packet sender, and *c_lv* is the level of cluster created by the sender (*i.e., CH_Ln*). *c_parent* points out the parent node of sender. *c_cost* is calculated on the basis of the sender's energy situation. It is written as,

$$SYNC_{c\_cost\_i} = \frac{E_{T\_i}}{E_{R\_i}} \quad (3a)$$

where

$$E_T = E_{Tx} + E_{Rx} \quad (3b)$$

The E_R is the sender's residual energy, and E_T is the average energy cost for one round relay task. Furthermore, a $\mathrm{SYNC}_{\mathrm{reply}}$ packet contains two items, which are *c_lv_max* and *node_total*. *c_lv_max* is the maximum or deepest level of cluster that exists along one branch. The *node_total* is the total number of nodes along this branch. Both values are updated by each cluster head that relays this $\mathrm{SYNC}_{\mathrm{reply}}$ packet to the sink node.

C. Network Clustering Stage

At the beginning of the network clustering stage, an initialization process is executed. Each node broadcast its information with randomly short delay whenever it starts working. Nodes that receive the message calculate relative distance and store information in their node management table. If a new node joins the network during service, it also broadcasts its node information, and surrounding nodes then broadcast their node information to the new node. The sink node is able to broadcast to all nodes in the network. In this case, initially, each node has a node management table that stores information of surrounding nodes and the sink node.

The sink node marks itself as *Level 0* cluster head (*CH_L0*), and updates the *c_cost* and *c_lv* = 0 into the SYNC packet. The sink node then sends the SYNC packet to its one-hop neighbors using the short range transmission mode. The one-hop neighbors which receive the SYNC packet choose the sink node as their cluster head, and then they store the value of *c_id*, *c_lv* and *c_cost* into their node management table. In this case, the one-hop neighbors of the sink node become cluster members of the *Level 0* cluster (*CM_L0*). According to the assumption of network configuration, only

one sink node means there is only one *Level 0* cluster in the network. Therefore, layer 0 is comprised of only one cluster. The cluster members of *Level 0* then individually decide whether they should be the ones to create the next level cluster.

At the Layer *n* construction stage node *i* is designed to accomplish two goals via its own SYNC packet. The first goal is to confirm to the cluster head of joining the *Level n* cluster. The second purpose is to decide whether it should generate the next level cluster. Whether a node with level ID of *CM_Ln* should create another level ID of *CH_L(n+1)* depends on the distance between itself and its cluster head *CH_Ln*, called boundary distance. An equation of boundary distance is given to determine the eligible nodes,

$$P\_CH\_L(n+1) = \begin{cases} 0.9, & bd < \dfrac{d(CH\_Ln, node\_i)}{R} < 1 \\ 0, & otherwise \end{cases} \quad (4a)$$

$$bd = \min(max_{bd}, min_{bd} * \rho * (n+1)) \quad (4b)$$

The *P_CH_L(n+1)* is the possibility of a *CM_Ln* to be a *CH_L(n+1)*. *R* is the multi-hop based radio transmission range of nodes, and it is also the radius of each cluster. $d(CH\_Ln, node\_i)$ is the approximate distance between node *i* and its cluster head, it is calculated by node *i* and stored in the node management table during initialization process. The *bd* is the boundary distance threshold that distinguishes possible cluster heads of next level from pure cluster members. The *n* is the level of current cluster. To improve energy utilization efficiency, lower level clusters are designed to have more nodes that could be next level cluster heads by assigning looser distance restrictions. For instance, the boundary distance threshold allows more nodes with level ID of *CM_L0* to become a potential cluster head than nodes with higher level ID. Furthermore, tighter boundary distance restriction is applied to reduce redundancy of higher level clusters. If node *i* does not meet the distance requirement, it only finishes the first SYNC packet goal, which is to confirm cluster membership to the cluster head. Otherwise, node *i* is said to establish a higher level cluster with a certain probability.

As shown in Fig. 1, all cluster heads (denoted with solid circles) except the sink node are also cluster members of lower level clusters. The solid triangles represent pure cluster members.

A parameter ρ is used to adjust the boundary distance so that the annulus of a cluster decreases with increasing layer number. Because the annulus decreases, the number of nodes that qualify for cluster head status decreases with increasing cluster layer. The reason for doing this is because the nodes that are farther away for the sink have a smaller chance of suffering from the "hot spot" problem. Accordingly, mitigation of the "hot spot' problem is decreasingly required for load balancing for cluster layers located further away for the sink. Therefore, the number of new clusters created at layers with increasing distance from the source decreases.

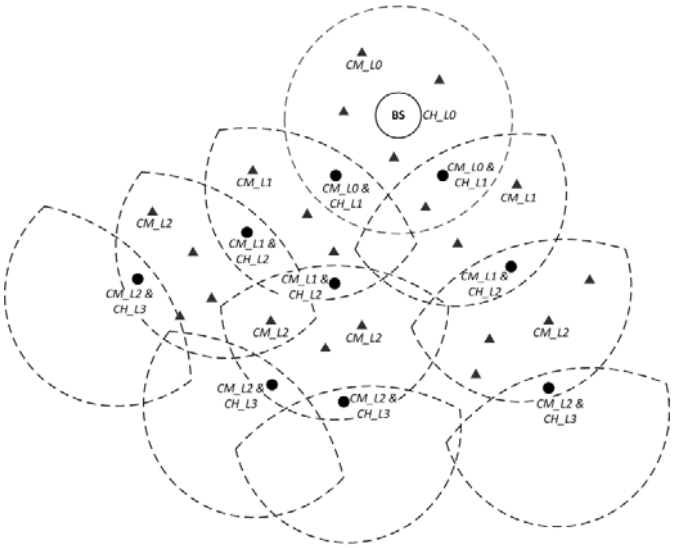


Fig. 1 Network clustering stage.

Similar with the sink node, node *i* first calculates *c_cost* and upgrades *c_lv* by one, thus, indicating it is the cluster head of *Level 1* cluster (*CH_L1*). *c_parent* is assigned with ID of the sink node. These values are then updated into SYNC packet and broadcast, both sink node and neighbor nodes of *i* are able to hear this packet: (i) sink node takes a look at the packet, it would add node *i* to its cluster member group if node *i* indicates the sink node is its cluster head. (ii) as soon as node *j* receives SYNC packet, it set up a contention timer. Once the timer is out, *j* chooses the SYNC packet sender that has lowest value of *c_lv* as its cluster head. If more than one SYNC packets have the same value of *c_lv*, then *j* selects the one that has lowest value of *c_cost*. It also stores another node that has the second lowest value of *c_cost* as alternative parent. The value of *c_lv* in the SYNC packet determines which level of cluster that node *j* becomes a cluster member of. For instance, if node *j* chooses node *i* as the cluster head, it then becomes cluster member of *Level 1* cluster (*CM_L1*). Node *j* then adds cluster head information into node management table. By this approach, each node is a cluster member of lower level cluster and possibly to be a cluster head of higher level cluster. This procedure is executed until the leaf nodes are encountered.

If a cluster head does not hear SYNC packet that indicates this node as a parent node, it then marks itself as a leaf

cluster head and transmits a $SYNC_{reply}$ packet along the formed branch. Data of *c_lv_max* and *node_total* is updated. *c_lv_max* is the level of this leaf cluster head, and *node_total* is the sum of direct and indirect child node, obviously, the value is 0 for a leaf cluster head. The $SYNC_{reply}$ packet is then relayed by each cluster head all the way to sink node. Each cluster head then updates the value of *node_total* of first received $SYNC_{reply}$ packet by adding the number of its direct child nodes. The cluster head then compares *c_lv_max* with its up to date highest level information. Cluster head updates its own highest level information if *c_lv_max* is larger, otherwise *c_lv_max* is updated. As a result, all cluster heads are able to know the highest cluster levels of this branch, and the amount of direct and indirect child nodes. Therefore, cluster heads have a sense of logical position in the network and relay load. This is for the purpose of load threshold function, which is stated in the data communications stage. When all $SYNC_{reply}$ packets reach sink node, the network clustering stage is formally finished and switched to data communications stage.

D. Data Communications Stage

In the duration of data communications stage, TDMA is applied for intra-cluster communications. Each *Level n+1* cluster member transmits data based on local time schedule *TS_L(n+1)*, and cluster head of *Level n+1* cluster relays data to lower level cluster head according to lower level cluster time schedule *TS_Ln*. The network topological architecture looks like an inverted triangle, and data is transmitted from highest side all the way to the lowest vertex, which is the sink node.

A cluster head has two energy thresholds, one is called base threshold and another is load threshold. The load threshold reveals relay load degree of a cluster head. Cluster heads that have large number of direct and indirect child nodes and more close to sink node (*i.e.,* lower cluster level) are likely to dissipate much more energy. In order to prevent rapid energy consumption of individual nodes, if the load threshold is breached, it means this cluster head is overloaded; it broadcast such message to all cluster members. Any cluster member that has alternative parent then broadcast message to join alternative cluster. It relays data to alternative cluster head once the request is confirmed. In the meanwhile, both the cluster head and alternative cluster heads recalculate load threshold. The load threshold is calculated as,

$$c\_load\_thresh = max\_thresh * (1 - \alpha \frac{n\_child * c\_lv}{node\_total * c\_lv\_max}) \quad (5)$$

where *max_thresh* is the maximum load threshold a cluster head can have. *n_child* is the number of direct child nodes and *c_lv* is the level of this cluster head. Parameter α is chosen with proper value to adjust results. From the equation, a cluster head that close to sink node or has more direct and indirect child nodes has larger load threshold. However, whenever the base threshold is violated, the cluster head disconnects all links and only sends its own data. In such situation, nodes that lost links switch cluster heads if they have alternative parent nodes. Otherwise, they automatically switch radio to single-hop transmission mode and send data to sink node in order to guarantee quality of service until next network clustering stage.

By using proposed single-hop transmission approach, a proper duration period of data acquisition is also important. A large period may degrade performance of minimize energy consumption difference. However, a small period could influence network efficiency and increase unnecessary overhead rate.

III. SIMULATION AND RESULTS

A simulation of the proposed hierarchical clustering protocol (HCP) was implemented in Matlab. Three comparison algorithms were demonstrated using the same parameters with hierarchical clusters design. The simulation was conducted to evaluate network performance that covers system survival period, link quality, and energy consumption. Parameters used in the simulation are given in Table 1. The parameters related to energy consumption are the same as those in [HeCh00]. Base threshold is set to 10% of maximum energy capacity. Max load threshold of cluster head is 30% of maximum energy capacity. Minimum and maximum boundary threshold are 20% and 80% of transmission range of multi-hop mode.

Table 1 Simulation Parameters.

Parameter	Value
Network coverage	(0,0)~(200,200)m
Base station location	(100,100)m
Number of nodes	300
Initial energy	0.5J
R	30m
E_{elec}	50nJ/bit
E_{fs}	10pJ/bit/m^2
E_{mp}	0.0013pJ/bit/m^4
E_{DA}	5nJ/bit/signal
Data packet size	4000 bits
ρ	2

The simulation was conducted to compare proposed algorithm with three previous network routing algorithms. They are the low-energy adaptive clustering hierarchy (LEACH), energy-efficient unequal clustering mechanism

(EEUC), and enhanced cross-layer protocol (ECLP). LEACH is selected since it is a typical clustering model for wireless sensor networks. Deep and systemic research has proved its value. The EEUC routing algorithm is designed to assign distinct cluster size to relieve energy pressure problems of nodes near sink node, which is also one of research goals of this paper. Furthermore, both LEACH and EEUC apply two radio transmission modes, which are the same as proposed by the present HCP algorithm. However, nodes in LEACH and EEUC switch radio mode to single-hop depend on distance between itself and the sink node, whereas in HCP, only nodes that loose link during service are allowed to send data directly to the sink node to maintain quality of service. Since LEACH, EEUC and HCP all belong to subcategory of clustering approach, the comparison is positive and worthy. ECLP is selected since it also applies layer design. A meaningful comparison can be given between ECLP and HCP.

As Fig. 2 illustrates the number of live nodes changes over time. HCP demonstrates significant longer service period than LEACH and EEUC. However, the ECLP shows better performance.

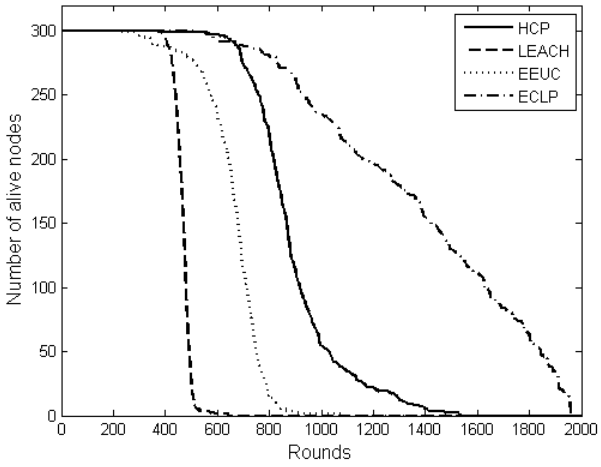


Fig. 2 Number of live nodes.

The reason is analyzed from a node efficiency aspect, which is shown in Fig. 3. Curves of LEACH and EEUC in Fig. 3 are the same with Fig. 2. It is because the link of these two algorithms is guaranteed since all nodes are able to reach the sink node by single-hop transmission. However, the node efficiency (percentage of nodes that are able to reach the sink node) of ECLP varies wildly, and then drops dramatically to a low level at a particular time spot. This is because that node in ECLP is link oriented. In multi-hop transmission mode, the sink node receives data from nodes closer to itself. This leads to huge energy consumption of these one-hop nodes, and obviously, the number of available one-hop nodes decrease during service. As a result, branches in the network tend to converge to one-hop nodes that are still available for data relay task. All nodes that have distance larger than one-hop distance lose links at the time the last one-hop node's residual energy is below its threshold, i.e., disconnecting all links and only send its own data. Though some nodes are still alive in ECLP, the network cannot service any more.

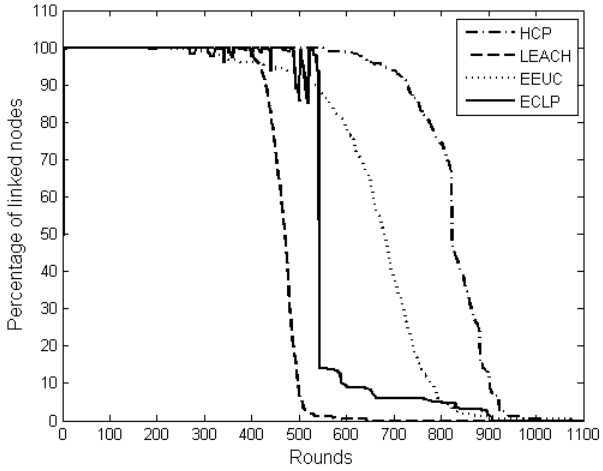


Fig. 3 Node efficiency.

The comparison between ECLP and HCP mainly concentrates on data relay efficiency, in other words, number of nodes that are hierarchical to the sink node during service. The HCP exhibits good performance on balancing energy consumption for each layer. Compared with ECLP of node efficiency, the variance of HCP is much smaller and smoother. The reason for such comparison result is that layers in ECLP categorize nodes, but layers in HCP classify clusters. Moreover, ECLP routing protocol does not apply single-hop transmission, nodes lost link broadcast error message and when reconnection attempt fails, they stop relaying data until next network configuration stage. In HCP, a node that lost link transmits data directly to the sink node.

A node efficiency comparison of one-hop nodes between ECLP and HCP is illustrated in Fig. 4. According to the simulation, the number of one-hop node is less than 10% of total number of nodes. As we can see, the percentage of available one-hop nodes in the figure decreases over time. The curve of HCP represents that the number of available one-hop nodes decreases smoothly. HCP provides longer network service time in terms of percentage of available one-hop nodes.

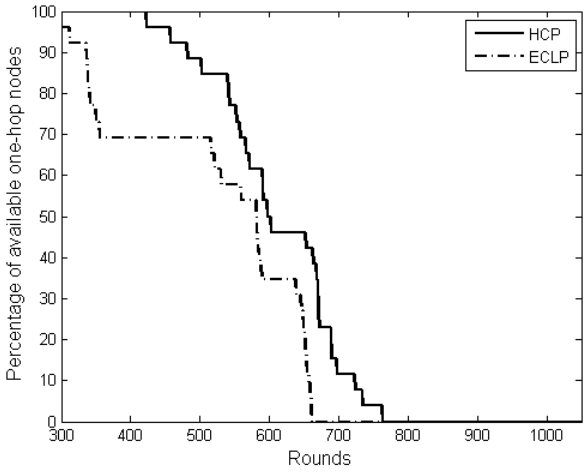


Fig. 4 Efficiency of one-hop nodes.

IV. CONCLUSIONS

This paper presents an innovated routing approach called hierarchical clusters design for wireless sensor networks. The goal of this proposed algorithm is to minimize energy dissipation difference in each layer, especially to relieve energy consumption pressure of nodes near sink node. Furthermore, a novel utilization of single-hop transmission mode is described in order to maintain node efficiency in the network. The hierarchical cluster design demonstrates promising network performance through simulation. The network lifetime is significantly prolonged compared with other algorithms. By using TDMA communications protocol and layer network architecture, the transmission delay is also reduced. The future work is needed to continue improve node efficiency, and the algorithm should be enhanced in larger scale of wireless sensor networks.

REFERENCES

[HeCh00] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," *in Proc. 33rd Hawaii Int. Conf. System Sciences*, vol. 8, pp. 8020, Jan 4-7, 2000.

{doi: 10.1109/HICSS.2000.926982}

[HeCh02] Wendi B. Heinzelman, Anantha P. Chandrakasan, and Hari Balakrishnan, "An application-specified protocol architecture for wireless microsensor networks," *in Proc. IEEE Trans. Wireless Communications*, vol. 1, no. 4, pp. 660-670, Oct 2002.

{doi: 10.1109/TWC.2002.804190}

[RaSc02] Vijay Raghunathan, Curt Schurgers, Sung Park, and Mani B. Srivastava, "Energy-aware wireless microsensor networks," *in IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 40-50, Mar 2002.

{doi: 10.1109/79.985679}

[ChCh01] SeongHwan Cho and Anantha P. Chandrakasan, "Energy efficient protocols for low duty cycle wireless microsensor," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2041-2044, May 7-11, 2001.

{doi: 10.1109/ICASSP.2001.940392}

[YeHe04] Wei Ye, John Heidemann, and Deborah Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *in IEEE Trans. Networking*, vol. 12, no. 3, pp. 493-506, Jun 2004.

{doi: 10.1109/TNET.2004.828953}

[DaLa03] Tijs van Dam and Koen Langendoen, "An adaptive energy-efficient MAC protocol for wireless sensor networks," *in Proc. the First Int. Conf. Embedded Networks Sensor Systems*, pp. 171-180, 2003.

{doi: 10.1145/958491.958512}

[LiYe05] Chengfa Li, Mao Ye, Guihai Chen, and Jie Wu, "An energy-efficient unequal clustering mechanism for wireless sensor networks," *in IEEE Int. Conf. Mobile Adhoc and Sensor Systems Conference*, pp. 604-612, Nov. 7, 2005.

{doi: 10.1109/MAHSS.2005.1542849}

[KiLe09] Jaehyun Kim, Jaiyong Lee, and Seoggyu Kim, "An enhanced cross-layer protocol for energy efficiency in wireless sensor networks," *in 3rd Int. Conf. Sensor Technologies and Applications*, pp. 657-664, Jun 18-23, 2009.

{doi:10.1109/SENSORCOMM.2009.106}

A Cluster Based Delay Tolerant MAC Protocol for Underwater Wireless Sensor Network

Zhanyang Zhang
Computer Science Department , College of Staten Island/City University of New York
Staten Island, New York, USA

Abstract - *the propagation speed of an acoustic signal is much slower than the speed of a terrestrial radio signal due to the physical characteristics of an underwater acoustic channel. This large delay can impact the throughput of the channel. There is also a very high delay variance which presents unique challenges to the designs of efficient protocols. The inconsistent delay renders many traditional communication protocols insufficient since they rely on accurate estimations of the round trip time (RTT) between two communication nodes. We presented a unique approach to develop a Cluster based delay-tolerant protocol (CBDTP) to address these problems by predicating a value for a sensor node if its data were not received at the sink node instead of having the sensor node retransmit its data. The CBDTP can reduce data traffic over the networks and uses the network resource more efficiently. In this paper we improved our CBDTP protocol by introduced a self-adoptive algorithm at MAC layer protocol to better address the inconsistent delay problems. It improves the performance of CBDTP protocol in a dynamic deployment environment.*

Keywords: *Underwater Wireless Sensor Networks, Acoustic Channel, Delay Tolerant Protocols.*

1 Introduction

We live on the surface of the earth of which close to 70% is covered by water. Underwater wireless sensor network (UWSN) is a promising technology to enable applications such as: oceanographic data collection, pollution monitoring, offshore exploration and tactical surveillance systems for homeland security. Ocean bottom sensor nodes can collect, process, and transmit data to surface or onshore stations via direct link or multi-hop relays. Compared to most terrestrial wireless sensor networks, the major difference is that UWSN uses acoustic wave instead of radio wave communication due to the poor propagation capability of radio waves under water. The physical characteristics of the underwater acoustic channel present a new set of challenging problems, such as, low bandwidth, long and variable propagation delays (5 times longer than radio) and high failure rate [1]. Many terrestrial radio-based wireless network protocols might not work well for underwater acoustic sensor networks.

Our research focuses on a major type of underwater wireless sensor network application, underwater environment monitoring, and security surveillance in rivers, lakes and oceans. This type of network requires a long duration of reliable operation with consideration of sensor failures and intermittent connections. Our goal is to explore new concepts in network protocol design that lead to more efficient use of scarce resources while tolerating longer delays and disrupted connections.

The rest of the paper is organized into the following sections. In section 2, we summarize some of the key issues and challenges in protocol design and implementation for underwater sensor networks due to long and inconsistent delay time. We reference some of the recent works related to delay-tolerant protocols to provide a background for our research. In section 3, we define a system model with abstraction and assumptions about the operation environment for the type of UWSN applications we are studying. In section 4, we present the design philosophy of the CBDTP as well as its finite state machine representation and the algorithms used in the CBDTP protocol including a newly designed self adoptive algorithm to adjust the expected packet arrival window based on past delay profiles. Simulation study and performance analysis are presented in section 5. We conclude the paper in section 6 to highlight the major contributions of our study and some open problems for future research.

2 Background and Related Work

Recent advances in communications and electronics have enabled the development of economical, low-power, multifunctional underwater sensor nodes that are small in size and can communicate for short distances via wireless links [2,3]. These sensor nodes are usually scattered underwater at different depths or at the bottom of the ocean floor. Each of these sensor nodes has the capability to collect data and route data back to the surface or to onshore sinks. Data is routed to the sinks by a multi-hop infrastructure-less architecture. Then the sink sends the data to a data processing center that could be many miles away [1]. One of the typical underwater wireless sensor network deployments is shown in Figure1.

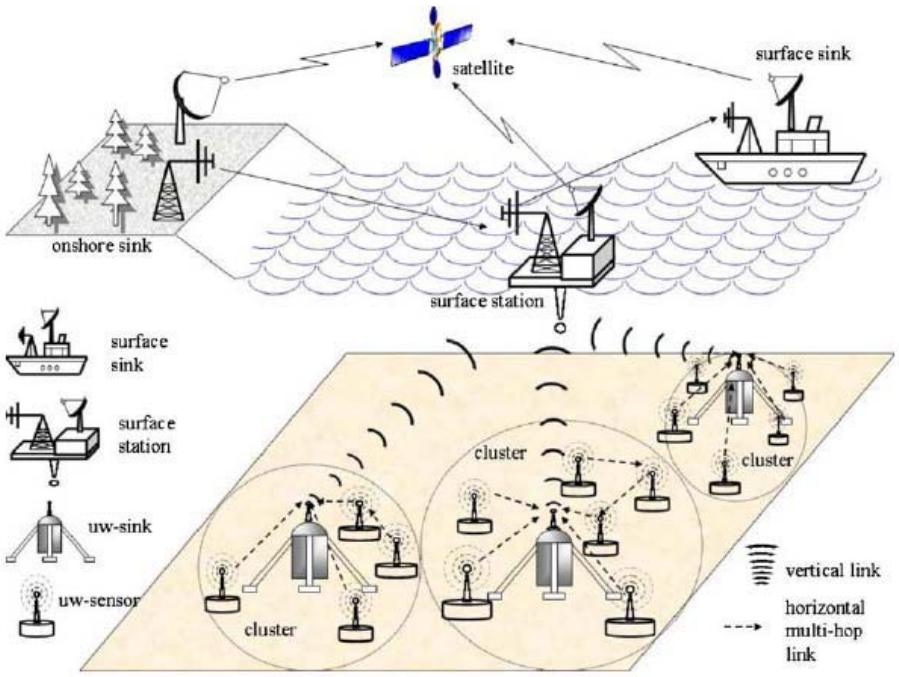


Figure 1. An example of underwater wireless sensor network architecture

Figure1 shows a UWSN network that has a 3-D topology and consists of heterogeneous nodes. Some nodes are more powerful and can perform more functions than others. For example, there are one or more sensor nodes (uw-sensor) in each cluster. But there is only one cluster head (uw-sink) which has more resources and can perform more functions than the sensor nodes, such as collecting data, controlling sensor nodes, and relaying aggregate data to the surface sinks via a vertical communication link. A sensor node only performs a few simple tasks with limited resources. Most sensors have the components that provide data collection/storage, data processing, communication functions and power supply (Fig. 2).

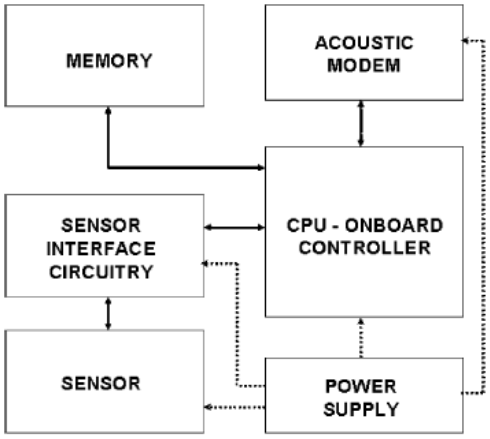


Figure 2. Internal architecture of an underwater sensor node

These nodes can be deployed from a few feet to 10,000 feet deep underwater. In the case of an oil company's offshore exploration and production platforms, sensor nodes are deployed up to 12,000 feet deep. These sensors can cover 400 square km of ocean floor [3]. They are used to collect data about water temperature, ocean current motion, strain gauges, long term fatigue and pipe corrosion.

The underwater acoustic propagation speed is five orders of magnitude slower than the speed in a terrestrial radio channel. The speed of acoustic signal is approximately 1500 m/s [2]. This large delay can reduce the throughput of the channel. There is a very high delay variance which makes efficient protocol designs even more difficult. The inconsistent delay renders many traditional communication protocols insufficient since they rely on accurate estimation of the round trip time (RTT) between two communication nodes.

Let's consider TCP protocol as an example. TCP is a reliable communication protocol between nodes. Node A sends a packet to node B. Node B sends an acknowledgement (ACK) back to node A when it receives the packet. If node A did not receive the ACK from node B after waiting for a period of time (based on RTT estimation), node A will assume the package was lost and it will send the packet again. This hand-shaking protocol guarantees a reliable communication between the two nodes. But the traditional TCP protocol does not work well in the case of long and inconsistent delays for underwater acoustic networks due to high delay variance and disrupted communication link. Considering the same example above, node A does not know if the packet were lost or how long it should wait for the ACK from node B. There is no way to know how long the RTT will be in the case of an extremely long delay. Therefore node A will keep re-transmitting the same packet many times. it overloads the

channel, wastes resources, and increases the risk of packet collisions.

There are a few previous published papers on delay-tolerant networks addressing the long delay problems in underwater acoustic networks. In [4], a propagation-delay-tolerant collision avoidance protocol was introduced to improve throughput performance while avoiding collision. The authors of [4] conducted a simulation study to show their protocol has better performance than traditional CSMA/CA protocol which is a widely used in Wi-Fi networks. The authors of [5] address the long delay problem in acoustic networks from network routing perspective. In [6], a data transport ferries concept was introduced as a delay-tolerant solution to address both long delay and low channel reliability problems. A data ferry is used to store and forward data from one node to another node. It works better in case of long delay and disrupted connection between certain nodes. Their work achieved a common goal, which is overcoming or tolerating the long propagation delay. Their proposed protocols still rely on a fair accurate round-trip-time (RTT) estimation. They do not work well in case of high delay variance since it would be difficult, if not impossible, to fair accurately forecast the RTT value in such case.

We have developed a prediction based delay-tolerant protocol (PBDTP) to address this problem at TCP layer [7]. PBDTP can tolerate long and inconsistent delays as well as connection disruptions by predicating a value for a sensor node if its data was not received at the sink instead of asking the node to retransmit its data. PBDTP significantly reduces data traffic over the networks and uses the network resource much more efficiently. In this paper, we present our recent study on applying PBDTP concept in designing a cluster based delay tolerant protocol (CBDTP). CBDTP is an enhancement to the most commonly used CSMA/CA MAC layer protocol (IEEE 802.11). Our simulation study shows that it increases the throughput and reliability of communications within a cluster.

3 The Cluster Based Network Model

The system model we specified here reflects the operation of a typical UWSN application for environment monitoring and detection. The system model is constructed based on acoustic channel characteristics that are applicable to underwater monitoring applications from a few hundred feet to ten-thousand feet. These UWSN applications are deployed in oceans or lakes with a large body of water. It is a heterogeneous sensor network with a three dimensional hieratical topology as depicted in Figure 1. A basic underwater operation unit is a cluster which consists of a number of sensor nodes with less resources and one cluster head node with more resources (Fig. 3).

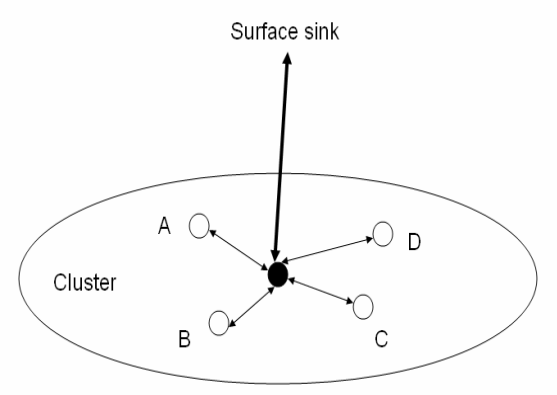


Figure 3. Cluster Based Delay Tolerant Protocol

Sensor nodes associate themselves to a cluster according to their close proximity to the cluster head. In most of the case, sensors in a cluster are located at relatively the same depth. Sensors only communicate with their head node directly (in one-hop distance). The cluster head collects data from sensor nodes or issues control commands to sensor nodes via a half-duplex acoustic channel. The cluster head also sends data to the surface sink vertically via one hop or multiple relay nodes. Due to effective communication ranges of the nodes, there may be a number of clusters in order to cover a large underwater area. In deep water applications, it may be necessary to deploy multiple relay nodes at different depths to relay data to a surface sink in multiple hops.

To keep the system model simple, we assume there is a direct link between cluster head and surface sink (Fig. 3). In most of the case, the cluster head has more resources and transmission power to send acoustic signal over much longer distance than a sensor node. We further assume that all nodes are stationary (anchored without mobility, except for drafts in a short range due to undersurface water current). Sensor nodes are deployed by surface or submarine vesicles. The vertical position (depth) is determined by the length of the anchor chains. We assume the sensors' horizontal positions are known to the UWSN operators via localization techniques [8]. However, the underwater horizontal positions may vary from their targeted positions due to drafts as a node is sinking. We assume such drafts are in small range and can be negligible.

From a real-world UWSN operation perspective, we describe how the system model works in 3 phases as listed below:

- Pre-deployment configuration - Before sensor nodes are deployed underwater, sensor node IDs are registered. Each sensor node needs to know the cluster head ID (CID) and the cluster head needs to know the entire sensor node IDs (SID) which belong to the cluster.
- Cluster formation (convergence) – During this phase, each sensor will exchange a probing message with the cluster head in order to confirm the existence of the

sensor node and to profile the communication link at the cluster head.

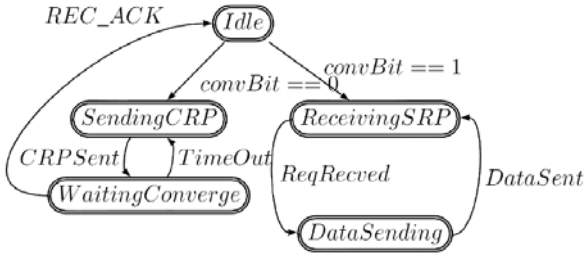
- Normal Operation – The cluster head starts to collect data from the sensor nodes and occasionally exchanges control/status messages with sensor nodes after forming the cluster. The system model can support three major groups of communication protocols for data communication between sensor nodes and the cluster head node, namely, a TDMA based protocol, a cluster head based query protocol (poll), and a sensor node event-driven protocol (push).

In this paper, we present our initial work on the cluster head based query protocol. We investigated this protocol first because of its simplicity and its wide applications in UWSN environment monitoring. We will report our work on TDMA and event-driven protocols in the near future.

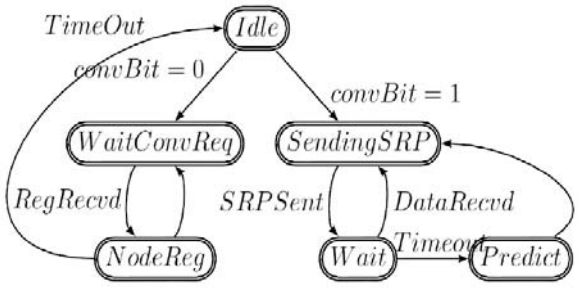
4 The Protocol Design

Underwater communication channels are subject to acoustic propagation characteristics including higher rate of link failure, long and inconsistent propagation delays [9]. Our design objective is to develop protocols that can tolerate these challenges presented in UWSN to improve performance and conserve system resources.

Since sensor nodes and the cluster head node in a cluster exhibit different behavior at different phases, it is better to illustrate the CBDTP protocol using two finite state diagrams, one for sensor nodes and another one for cluster head nodes [Fig. 4].



(a) State diagram for sensor nodes



(b) State diagram for cluster head nodes

Figure 4. Finite state machine representation of CBDTP protocol

There are 5 states for the sensor nodes (Fig. 4a), they are:

1. *Idle state* – This is the initial state. All the sensor nodes are set to this state at the time of deployment. The converge bit is set to zero (convBit=0)
2. *SendingCRP state* – If a sensor is in the Idle state and convBit=0 then it enters the SendingCRP state. This is the state where sensor nodes send a probe message (CRPSent-cluster register packet) to the cluster head.
3. *WaitingConverge state* – After sending a probe message, a sensor node enters the WaitingConverge state to wait the reply from the cluster head. Upon receiving the reply (REC_ACK), the sensor node returns to the Idle state and set the convBit=1.
4. *ReceivingSRP state* – While at the Idle state, if convBit=1 then a sensor node enters the ReceivingSRP state. In this state, a sensor node waits for the data request packet from the cluster head node. When the sensor node receives the data request (ReqRecved), it enters the DataSending state.
5. *DataSending state* – While in this state, a sensor node assembles a data packet then it sends the data packet to the cluster head node. Afterward, it returns to the ReceivingSRP state and repeats the same cycle.

There are 6 states for the cluster head nodes (Fig. 4b), they are:

1. *Idle state* - This is the initial state for cluster head node. It works in the same way as in the case of sensor nodes.
2. *WaitConvReg state* – This is the state where a cluster head node listens for the probe messages coming from all the sensor nodes. It sets a timeout period that controls how long it will wait for the probe messages.
3. *NoteReg state* – When the cluster head node gets a probe message, it enters the NoteReg state to register the sensor node ID in its registry. Then it checks the timeout to determine if it should return to the WaitConvReg state for the next probe message or return to the Idle state and set the convBit=1.
4. *SendingSRP state* – While in the Idle state, if the convBit=1, the cluster head node enters this state. It will send a data request packet (SPRSent) to the sensor node which is in front of a registered sensor node queue. Then the cluster head node enters Wait state and sets up a timeout clock.
5. *Wait state* – While in the Wait state, a cluster head node waits to receive the data packet from the sensor node. It will log the data (DataRecved) if the data packet arrives within the timeout period. Then it returns the SendingSRP state and repeats the same cycle with the next sensor node in the queue. Otherwise it enters the Predict sate.
6. *Predict state* – A cluster head node enters this state because it did not receive the expected data packet from the target sensor node within the timeout period. The data packet could be lost or it has taken an unusual long delay. The cluster head node predicts the data value based on previous data from this sensor node. Then it returns the SendingSRP node.

The proposed CBDTP protocol design has two key components. Here we describe briefly how each component works.

4.1 The Predication Algorithm

As shown in Figure 1, sensor nodes are grouped in clusters. Each cluster has a cluster head (black dot) and a few sensor nodes. Sensor nodes send their data to the head node via one-hop horizontal links. The head node sends the aggregated data to a surface sink via a vertical link (Fig. 3). The head node collects data from sensor nodes in a round-robin style at fixed cycles (called rounds). The data flow is from sensor nodes to the cluster head then to the surface sink node. The control flow is the reverse of the data flow. Since both data and control share the same half-duplex link, the link is shown as bi-directional in Fig. 3.

Published research results show that the delay variance in horizontal acoustic links is generally larger than in vertical links due to the nature of underwater links [10]. The links between sensor nodes and the cluster head node are weaker links due to this fact. In the case of a long delay or an unexpected interruption, the head node will predict a value for the sensor nodes using the prediction algorithm. Based on our observations, we make two assumptions, (1) a sensor node most likely will report the same value or a similar value with a small change over a short time window. (2) Adjacent sensor nodes will most likely report the same values or similar values with small changes. The prediction algorithm predicts a value for a sensor node either based on the previous values of the sensor node, or its neighbor sensor nodes' values. It also can predict the value based on a combination of both. The lost data packets are compensated at the cluster head node with the predicted values without resending the data. The CBDTP protocol can tolerate long delays and irregular disruptions with efficient usage of network resources.

4.2 The Adjustment Algorithm

If the head node receives data from its sensor nodes after making the prediction, it can replace the predicted value and use the real data value as the basis for future predictions for better accuracy. If the difference between the received value and predicted value exceed certain thresholds, the head node will send the newly received value to a surface sink so that the sink can make necessary adjustments based on application requirements. Otherwise no actions are taken.

5 Simulation and Performance Analysis

A preliminary simulation study was conducted to validate the system model and to measure the CBDTP protocol performance against our design objectives. We measure the communication link performance between a sensor node and the cluster head in term of throughput (number of packets per unit of time) and the data accuracy (the difference between the data sent by sensors and estimated data received at the sink in the case of predictions were made).

We compared the performance of the CBDTP protocol and the traditional CSMA/CA protocol under the same simulation conditions to test if the proposed CBDTP protocol can achieve reasonable performance gains over CSMA/CA protocol. We implemented a simulation study at MAC layer to compare the CBDTP protocol with a few other protocols including CSMA/CA protocols [3, 4, and 5].

5.1 Simulation Model

There are many ways to conduct a simulation study. In this paper, we present a set of simulation scenarios that closely reflect the class of UWSN applications for environment monitoring and detection. Since a cluster is the basic operation unit of a UWSN application and it presents the most challenges in its horizontal communication links, therefore we focus our simulation study on communications links between sensor nodes and the cluster head node within a cluster. We set the cluster size with 7 sensor nodes and one cluster head node. All the sensors are within the maximum distance of 500 meters from the cluster head. The baseline propagation delay (*dt*) is about 1/3 second or 333ms. The propagation delay variance (Δdt) can be 0 to 100ms in addition to the baseline. In order to simulate the reliability factor of the channel, we introduce a probability of channel failure (*p*). We set the channel failure value in the range of (0.01, 0.05, 0.10, 0.15, and 0.20) to study the impacts on channel performance.

The simulation program is written in C++ with distributed processes to emulate the sensor nodes and the cluster head node. Each process is a self-control entity. The sample size of each simulation run is 1500 rounds of data collection.

There are two specific functions that we implemented in the simulation.

- We adopted the Newton method to estimate the data value in the prediction algorithm using three data values in previous rounds to achieve better accuracy.
- Due to propagation delay inconsistency, a fixed timeout window set for an acoustic channel may not best reflect the dynamic changes over time. We introduced a dynamic self-adoptive timeout window management schema based on the well known Jacobson-Karn's Algorithm [11] which adjusts the timeout window based on past delay profiles.

5.2 The Adaptive Round Trip Time (RTT) Estimation Algorithm

The Jacobson-Karn's Algorithm in its original form is defined as:

$$D = \alpha D + \beta |RTT - M| \quad (1)$$

$$Timeout = RTT + 4D \quad (2)$$

Where RTT is the expected value for round trip time. D is the standard deviation of RTTs. M is the observed value for round trip time. |RTT – M| is the difference between the expected and observed values. In our simulation we set α=1, β=1/8 and use 2D instead of 4D for a better RTT estimation pertain to unreliable underwater channels.

There is a known problem with Jacobson-Karn's Algorithm when it is applied to a highly unreliable channel. When expected data does not arrive during the timeout window, it is not clear whether it is due to congestion or due to lost data. Making a wrong judgment can be costly. For example, increasing timeout window may improve performance in case of congestion. But it can impact performance negatively in case of data lost. To address this problem in our simulation, we developed an algorithm to offset the timeout window calculated by Jacobson-Karn's algorithm with a dynamic factor based on past history of data lost ratio.

5.3 Performance Analysis

Our initial simulation results show that the proposed CBDTP protocol performs better than traditional CSMA/CA protocol in the above scenarios that are used for the simulation. Figure 5 shows the throughput (data packet received by the cluster head node per minutes, 500 bits/packet) in a typical environment monitoring application with low to moderate data rate. There are three lines in the chart. They represent the effective throughput (square-dot line), the real throughput (diamond-dot line) and the CSMA/CA throughput (triangle-dot line). We use both the effective throughput and real throughput to measure the performance of the proposed CBDTP. The effective throughput includes the predicted data packets which were lost during transmission. Since the CBDTP uses the estimated data value to make up the lost data packet. The net effect of throughput is higher than the real throughput which only counts the packet that went through the channel. The performance gain over the CSMA/CA protocol is amplified as the channel quality gets worse (as channel failure probability increases from 0.01 to 0.20). In other word, the CBDTP protocol can better tolerate the channel quality degradation than the CSMA/CA protocol does.

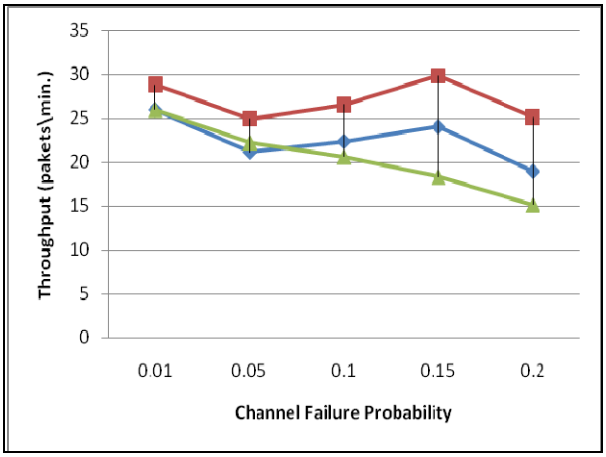


Figure 5. Cluster throughput comparison

While the CBDTP protocol has a notable performance gain in terms of throughput, we are also very interested in the quality of CBDTP in terms of data accuracy. We define data error as the differences between the received data values and the original values the sensor nodes collected and sent. In order to measure the data error, we extracted the average value of sensor data from all the nodes, which is the ocean water temperature and the average value of data received at the cluster head node (includes both real data and estimated data). Our simulation data shows that the two average values are very close (Figure 6). The average value of sensor data is shown as square-dot line and the average value of received data is shown as diamond-dot line. To measure data error variance and distribution, our study shows the mean value of data error is 0.12 and the standard deviation is 0.90 with 15000 samples.

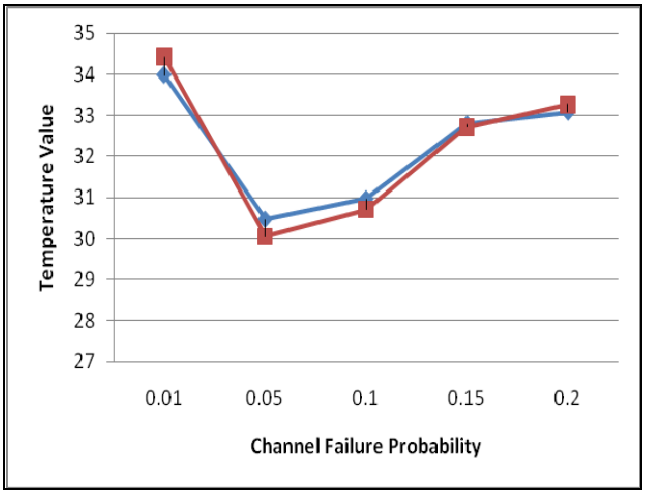


Figure 6. CBDTP protocol average temperature data accuracy

6 Conclusions

In this paper, we point out the challenges and problems that UWSN design engineers and operators are facing, in particular, the problems of link reliability, long delay and inconsistent delay. Traditional protocols do not work well with these problems. We proposed a new CBDTP protocol to overcome these challenging problems. We made two key contributions in the proposed CBDTP protocol, namely, data prediction and adjustment algorithms. Our simulation study shows the CBDTP protocol outperforms the traditional CSMA/CV protocol while maintaining good data quality.

There are a few open issues that require future research within the CBDTP framework on TDMA and event-driven protocols as well as further refinement of the query based CBDTP protocol to better manage the dynamic of timeout windows. CBDTP also faces a scalability issue. As the number of sensor nodes increases, the duration of a data collection round increases proportionally. In our simulation, we used a cluster with 10 sensor nodes. But in a real UWSN application, there can be a few hundreds sensor nodes in a cluster. We also plan to investigate CDMA/OFDA multi carrier channels for each sensor node sending data to cluster head at same time in a single RTT window.

7 References

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, "State of the Art in Protocol Research for Underwater Acosutic Sensor Networks," ACM Mobile Computing & Communication Review, vol. 11, no. 4, pp. 11–22, October, 2007.

[2] C. Liu and J. Wu, "An Optimal Probabilistic Forwarding Protocol Delay Tolerant Networks," in Proc. of ACM MOBIHOC, May 2009.

[3] J. Partan, J. Kurose and B. N. Levine, "A Survey of Practical Issues in Underwater Netowrks," ACM 2006 Wireless Underwater Networks (WUWnet'06), Los Angeles, September 2006.

[4] X. Guo, M. R. Frater, and M. J. Ryan, "A Propagation-delay-tolerant Collision Avoidance Protocol for Underwater Acoustic Sensor Networks," MTS/IEEE Conference and Exhibition for Ocean Engineering, Science and Technology (OCEANS), Boston, MP, September, 2006.

[5] D. Pompili, T. Melodia, and I. F. Akyildiz, "Routing Algorithm for Delay-insensitive and delay-sensitive Applications in Underwater Sensor Networks," ACM Conference on Mobile Computing and Networking (MobiCom), Los Angeles, CA, September, 2006.

[6] W. Zhao, M. Ammar, and E. Zegura, "Controlling the Mobility of Multiple Data Transport Ferries in a Delay-tolerant network," IEEE INFOCOM, 2005.

[7] Z. Zhang, S. Lin, and K. Sung "A Predication-Based Delay Tolerant Protocol for Underwater Sensor Networks". Proceeding of IEEE Wireless Communication and Signal Processing Conference (IEEE WCSP2010). Suzhou, China October, 2010.

[8] V. Chandrasekhar, W. K. Seah, Y. S. Choo, and H. V. Ee, "Localization in Underwater Sensor Networks – Survey and Challenges," In proceedings of ACM International Workshop on Underwater Networks (WUWNet), Los Angeles, CA, September, 2006.

[9] J. Preisig, "Acoustic Propagation Considerations for Underwater Acoustic Communications Network Development," ACM Mobile Computing & Communication Review, vol. 11, no. 4, pp. 1–10, October, 2007.

[10] M. Stojanovic, "Acoustic (Underwater) Communications," In J. G. Proakis, editior, Encyclopedia of Telecommunications, John Wiley and Sons, 2003.

[11] A. Tanenbaum, "Computer Networks" (4th edition). Prentice Hall, 2003, Page 550-555.

Acknowledgment

This research was supported, in part, under National Science Foundation Grants CNS-0958379 and CNS-0855217 and the City University of New York High Performance Computing Center at the College of Staten Island.

A Multi-hop Source Routing based on the Topology Matrix in Cluster Sensor Networks

Mary Wu, Woosuck Chang†, and ChongGun Kim[1]
Department of Computer Engineering, Yeungnam University
†Kumi College

Abstract - *Sensors have limited resources in sensor networks, so efficient use of energy is important. Representative clustering methods, LEACH, LEACHC, TEEN generally use direct transmission methods from cluster headers to the sink node to pass collected data. If clusters are located at a long distance from the sink node, the cluster headers exhaust a lot of energy in order to transfer the data. As a consequence, the life of sensors is shorten and re-clustering is needed. In the process of clustering, sensor nodes consume some energy and the energy depletion of the cluster heads meets another energy exhaustion. Many routing studies have been for saving energy. Generally, sensor nodes exchange routing information with each other in order to create a routing table. Due to this, a large amount of bandwidth is consumed and a large overhead is occurred. In this paper, we propose a Multi-hop Source Routing(MSR) method without routing tables in cluster sensor networks. This method uses the topology matrix which presents cluster topology. The experiment results show that proposed routing method is energy-efficient that any other routing method based on the routing table in clustered sensor networks.*

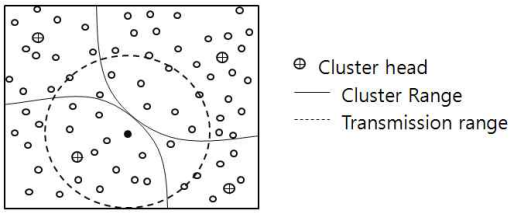
Keywords: Multi-hop source routing, Clustering, Sensor networks, Topology Matrix, Energy Efficiency

1 Introduction

Wireless Sensor Networks collect data on the surrounding environment and can be applied to a variety of purposes such as intrusion detection in military areas, area security, and environmental monitoring of temperature and humidity. Sensor nodes become aware of the resulting symptoms, and transmit the measured data to a base station, which in turn analyzes the data. A limitation arises due to the limited resources of sensor nodes on the wireless sensor networks. Many studies on the efficient use of energy have been conducted to overcome this problem [1-17]. Typically, neighboring sensor nodes collect similar information, leading to large energy wastage because of duplicated transmission of similar information. Consequently, many cluster methods on sensor networks have been studied. Clustering, in which the sensor network is divided into non-overlapping groups of nodes, is an effective method for achieving high levels of energy efficiency and scalability. In clustering, each node belongs to a local cluster and a cluster head integrates the data collected from members of the cluster, and then transmits it to a sink node. This prevents duplicate transmission of similar information and gives low-power networking in the sensor networks[2-14]. As a representative cluster protocol, there are protocols such as LEACH, TEEN, APTEEN[2-4]. Such protocols are based on the assumption of LEACH network environments. In the LEACH network environments, all sensor nodes are possible to communicate directly with the sink node, located in the far distance and cluster heads send data to the sink node directly. It causes a lot of energy consumption and is likely to be shortened lifetime of the cluster head. In clustered sensor networks environment, multi-hop transmission for energy efficiency measures is needed. Generally, cluster heads can perform multi-hop routing based on the information in the routing table[8-13]. Sensor nodes exchange routing information with each other in order to create a routing table. Due to this, a large amount of bandwidth is consumed and a large overhead is occurred. In this paper, we propose a source routing method without routing tables in cluster sensor networks. This method uses the topology matrix which presents cluster topology. Section 2 introduces the topology matrix in a cluster sensor network, section 3 introduces our proposed Multi-hop Source Routing(MSR) method and section 4 introduces the entire networking and routing procedures and section 5 presents the results of performance evaluation. Finally, section 5 concludes conclusion.

2 Related Works

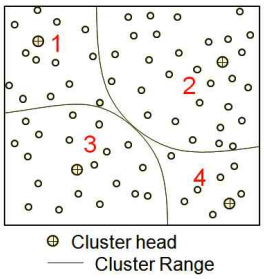
The data transmission of nodes in the cluster boundary can cause interference for data transmission neighbor clusters. Fig. 1 shows this situation that the node, located on the boundary of the cluster, affects to neighboring clusters and the data transmission can cause the collision and interference.



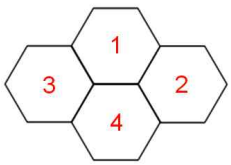
[1] Corresponding Author

Fig. 1 Collision and interference of boundary nodes between neighboring clusters

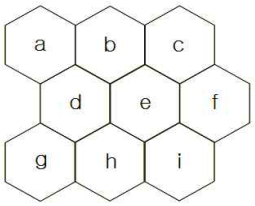
In [14,15], a method allocating different frequency channels between neighboring clusters is proposed in order to solve the problem and it generates the topology matrix which represents the cluster topology in a cluster sensor network.



(a) Cluster sensor network



(b) The hexagonal model of (a)



(c) Nine cluster hexagonal model

Fig. 2 Topology and hexagonal models in cluster sensor networks

Cluster topology matrix is generated based on a hexagon model. Fig 2(b) presents the topology of 2(a) as a hexagonal model. (1) is the topology matrix of 2(c)[16].

$$T=\begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} & t_{16} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} & t_{26} \\ t_{31} & t_{32} & t_{33} & t_{34} & t_{35} & t_{36} \end{bmatrix}=\begin{bmatrix} a & 0 & b & 0 & d & 0 \\ 0 & d & 0 & e & 0 & f \\ g & 0 & h & 0 & i & 0 \end{bmatrix}. \quad (1)$$

$$RA=\begin{bmatrix} 1 & \infty & 0 & \infty & 2 & \infty \\ \infty & 2 & \infty & 1 & \infty & 0 \\ 1 & \infty & 0 & \infty & 2 & \infty \end{bmatrix},$$

$$where \quad ra_{ij}=\begin{cases} (3i+j)\ \%\ 3, & for\ t_{ij}\in T,\ if\ t_{ij}\neq 0, \\ \infty, & otherwise. \end{cases} \quad (2)$$

After the formation of clusters, each cluster collects the information of its neighboring clusters and transfer it a server or a gateway. The server generates a cluster adjacency matrix using the information received from all clusters in the network and then a topology matrix T based on A. Resource allocation matrix RA is generated based on the topology matrix and allocates non-overlapping among neighbor clusters in (2).

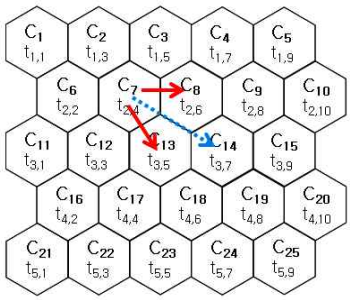
3 Multihop Source Routing(MSR) based on Topology Matrix

As a way to transmit data to the sink node, each cluster head uses DV(Distance Vector) value which is calculated by subtracting from the destination index to the source index of the topology matrix. DV appears as follows.

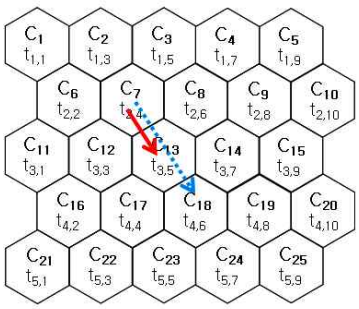
$$DV=(i_d,\ j_d)-(i_s,\ j_s)=(i_d-i_s,\ j_d-j_s),$$

where the destination element of T matrix is $t_{i_d,\ j_d}$ and the source element of T matrix is $t_{i_s,\ j_s}$.

The cluster head calculates DV value and determines the next cluster head using the DV rule of table 1. When source cluster $t_{2,4}$ transfer data to destination cluster $t_{3,7}$, DV is (3, 7)-(2, 4)=(1, 3). (1, 3) is applied by the rule (i>0, j>i) and the next cluster is determined by $t_{i+1,\ j+1}$ or $t_{i,\ j+2}$. Source cluster is $t_{2,4}$, so the next cluster is $t_{3,5}$ or $t_{2,\ 6}$ in fig 3(a). One of the both is randomly determined as a next cluster. When source cluster $t_{2,4}$ transfer data to destination cluster $t_{4,6}$, DV is (4, 6)-(2, 4)=(2, 2). (2, 2) is applied by the rule (i>0, 0<j≤i) and the next cluster is determined as $t_{i+1,\ j+1}$. Source cluster is $t_{2,4}$, so the next cluster is $t_{3,5}$ in fig 3(b).



(a) Right-bottom direction destination 1



(b) Right-bottom direction destination 2

Fig. 3 Routing based on DV pattern

Table 1 Next cluster determination rule based on DV pattern

DV-i	DV-j	Next Cluster
i>0	j>i	i+1, j+1 or j+2
i>0	0<j≤i	i+1, j+1
i>0	j=0	i+1, j-1 or i+1, j+1
i>0	-i≤j<0	i+1, j-1
i>0	j<-i	i+1, j-1 or j-2
i=0	j>0	j+2
i=0	j<0	j-2

i<0	j>-i	i-1, j+1 or j+2
i<0	0<j≤-i	i-1, j+1
i<0	j=0	i-1, j-1 or i-1, j+1
i<0	i≤j<0	i-1, j-1
i<0	j<i	i-1, j-1 or j-2

4 Networking and Routing Procedure

Networking and Routing Procedure for clustered sensor networks consists of cluster setup period, channels allocation period, data sensing and collection period, routing and transmission period as shown at fig. 4.

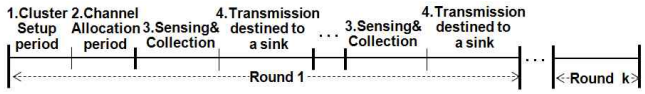


Fig. 4 A whole networking procedure in clustered sensor networks

Clustering may use the method of a representative clustering research, LEACH. It replaces cluster heads by random based on probability in order to evenly the energy consumption between nodes on the sensor network.

After a cluster setup phase, there is a channel allocation period. Each cluster head transfers the message including the information of neighbor clusters to the server and it generates a topology matrix T and a resource allocation matrix RA based on the information. The calculation for channel allocation is performed on a server and it can be the gateway of sensor networks. The server sends the matrices to each cluster head. Eech cluster has non-overlapping channel among neighbors based on RA.

In data sensing and collection period, cluster members communicate with their cluster head using their allocated channel. TDMA is used for non-collision transmission within a cluster. Cluster members transfer their cluster head sensing data in their slot time as shown fig. 5.

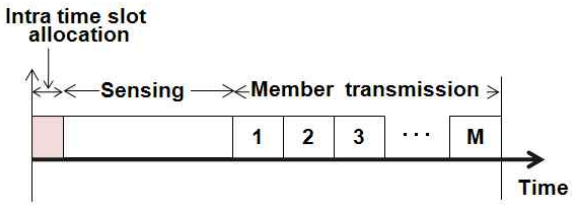


Fig. 5 A whole procedure in cluster sensor networks

In transmission period, cluster heads aggregate the collected date and transmit a sink it. Cluster heads calculate DV value and determine the next cluster based on the rules of table 1. Routing message exchange for routing tables isn't needed.

5 Performance

The experiment for the energy efficiency of MSR method based is performed. Energy consumption model[16,17] is shown by (4).

$$E_{Tx}(k,d) = \begin{cases} k*E_{elec} + k*\varepsilon_{fs}*d^2 & (d < d_0) \\ k*E_{elec} + k*\varepsilon_{mp}*d^4 & (d \geq d_0) \end{cases}$$
$$E_{Rx}(k) = E_{elec}*k \qquad (4)$$

E_{Tx} is transmission energy. If distance is smaller than threshold d_0, free space model(d^2 : energy loss) is used, if distance d is larger than threshold d_0, multipath model(d^4 : energy loss) is used. k is the number of bits, E_{elec} is electron energy, ε_{fs}(free space) and ε_{mp}(multi-path) are amplification energy to maintain acceptable SNR. d is transmission distance, d_0 is calculated by $\sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. E_{Rx} is receiving energy.

Table 2 Experimental elements

Elements	Explanation	Value
k	The number of bits	512
d	The distance between cluster centers	20m
d_0	$\sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$	87m
E_{elec}	Electron energy	50 nJ/bit
ε_{fs}	Free space amplification energy	10 pj/bit/m^2
ε_{mp}	Multipath amplification energy	0.013 pj/bit/m^4

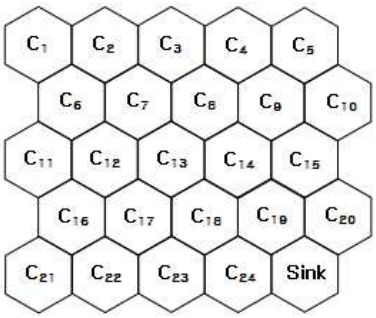
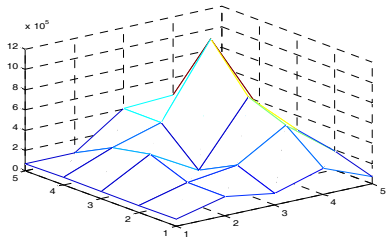


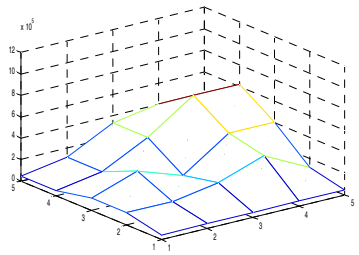
Fig. 6 Experimental topology model

Table 3 Direct transmission and routing based on the topology matrix

	C1	C2	C3	C4	C5
Routing table	71680	71680	71680	71680	71680
MSR routing	46080	117760	117760	46080	46080
	C6	**C7**	**C8**	**C9**	**C10**
Routing table	189440	189440	332800	261120	71680
MSR routing	46080	117760	261120	117760	117760
	C11	**C12**	**C13**	**C14**	**C15**
Routing table	117760	261120	71680	404480	404480
MSR routing	46080	189440	117760	332800	332800
	C16	**C17**	**C18**	**C19**	**C20**
Routing table	261120	547840	691200	1121280	430080
MSR routing	46080	332800	404480	619520	404480
	C21	**C22**	**C23**	**C24**	**Sink**
Routing table	71680	189440	261120	332800	614400
MSR routing	46080	117760	404480	619520	614400



(a) Routing table



(b) MSR routing

Fig. 7 Energy consumption

Table 3 and Fig. 7 show the amount of energy used for the routing based on routing tables, MSR routing in the topology of fig. 6. The energy efficiency of MSR routing is better than that of the method based on routing tables. The total energy consumption of MSR routing shows 78.8% compared with that of routing based on routing tables. This method is simple and can be easily applied in practical systems.

6 Conclusions

This paper proposes a multi-hop source routing(MSR) method based on topology matrix. This method uses DV value without routing table for energy efficiency when transmitting a sink data. It shows the improved performance in the energy consumption which compared with that of the routing method base on routing tables and our proposed method.

7 References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks : a survey", Computer Networks 38, 2002

[2] W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", The Hwaii International Conference on System Science, pp. 1-10, Jan, 2000

[3] Arati Manjeshwar, Dharma P. Agrawal, "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks", The 15th Parallel and Distributed Processing Symposium, pp. 2009-2015, Apr. 2000

[4] Manjeshwar, "APTEEN: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks", Parallel and Distributed Processing Symposium, IPDPS 2002, pp. 195-202, 2002

[5] Ossama Younis, Sonia Fahmy, "HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad-hoc Sensor Networks", IEEE Transaction on Mobile Computing, Oct. 2004

[6] Liang Ying, Yu Haibin, "Energy Adaptive Cluster-Head Selection for Wireless Sensor Networks", The 6th International Conference on Parallel and Distributed Computing Applications and Technologies, pp. 634-638, Dec. 2005

[7] J. s. Kim, J. H. Lee, and K. W. Rim, "3DE : Selective Cluster Head Selection scheme for Energy Efficiency in Wireless Sensor Networks", The 2nd ACM International Conference on Pervasive Technologies Related to Assistive Environments, 2009

[8] Dae-Young Kim, Jinsung Cho, "A Method to Support Mobile Sink Node in a Hierarchical Routing Protocol of Wireless Sensor Networks", The Journal of Korea Information and Communications Society, 08-01, Vol. 33, No. 1, pp. 48-57, 2008

[9] Sung-Hwa Hong, Byoung-Kug Kim, Doo-Seop Eom, "Flooding Level Cluster-based Hierarchical Routing Algorithm For Improving Performance in Multi-Hop Wireless Sensor Networks", The Journal of Korea Information and Communications Society, 08-03, Vol. 33, No. 3, pp. 123-134, 2008

[10] Choon Sung Nam, Young Shin Han, and Dong Ryeol Shin, "Multi-Hop Routing-Based Optimization of the Number of Cluster-Heads in Wireless Sensor Networks", The Journal of Sensors 2011, 11(3), pp. 2875-2884, 2011

[11] Guihai Chen, Chengfa Li, Mao Ye, Jie Wu, "An Unequal Cluster-Based Routing Protocol in Wireless Sensor Networks",

The Journal of Wireless Networks, Vol. 15, Issue 2, pp. 193-207, 2009
[12] G. Santhosh Kumar, Sitara A., K. Poulose Jacob, "Energy Aware Cluster-based Multihop Routing Protocol for Sensor Networks", International Journal of Information Processing, 4(3), pp. 9-16, 2010
[13] J. S. Rauthan, S. Mishra, "An Improved Cluster Based Multi-hop Routing in Self-Organizing Wireless Sensor Networks", International Journal of Engineering Research & Technology, Vol. 1, Issue 4, Jun. 2012
[14] Mary Wu, InTaek Leem, Jason J. Jung, ChongGun Kim, "A resource reuse method in cluster sensor networks in ad hoc networks", The 4th Asian conference on Intelligent Information and Database Systems, Vol. II, pp. 40-50, 2012
[15] Mary Wu, Byungchul Ahn, ChongGun Kim, "A Channel Reuse Procedure in Clustering Sensor Networks", Applied Mechanics and Materials Vols. 284-287, pp. 1981-1985, 2013
[16] M. Ye, C. F. Li, G. H. Chen, and J. Wu, "EECS: AnEnergy Efficient Clustering Scheme in Wireless Sensor Networks", IEEE International Performance Computing and Communications Conference (IPCCC), pp. 535-540, 2005.
[17] Bencan Gong, Layuan Li, Shaorong Wang, Xuejun Zhou, "Multihop Routing Protocol with Unequal Clustering for Wireless Sensor Networks", ISECS International Colloquium on Computing, Communication, Control, and Management, pp. 552-556, 2008

Study on Wireless Sensor Network Based Livestock Farm Integrated Control System

Hyungi Kim[1], **Hyun Yoe**[2]
[12]Department of Information and Communication Engineering Sunchon National University,
Suncheon, Jeollanam-do, Republic of Korea
kimhyungi@sunchon.ac.kr, yhyun@sunchon.ac.kr

Abstract - *This paper proposes livestock farm integrated control system based on wireless sensor network. The conventional livestock farm control monitoring system based on the WSN realizes a higher value-added livestock farmhouse, however, it has a disadvantage that is difficult to carry out the integrated control of multiple livestock farmhouses by a system that could use for a single livestock farmhouse. Therefore, the proposed system stores the environmental information received from the individual livestock farm management server of the integrated livestock farm control system based on the distributed processing platform in a DB, carries out the integrated management and comparative analysis for the collected environmental information, and provides information on the integrated environmental information and the abnormal symptom of the livestock farm through the text service and the GUI of devices that could communicate when there is an abnormal symptom in the livestock farm. It could carry out the integrated management of individual livestock farm by applying the proposed integrated livestock farm control system based on the distributed processing platform, and could quickly cope with a dangerous situation of the livestock farm by informing it in real time when there is an abnormal symptom.*

Keywords: WSN, Livestock, Distributed Framework

1 Introduction

The WSN(Wireless Sensor Networks) is a technology that deploys sensor nodes with computing and wireless communication capabilities to the application environment, forms a network autonomously, and then collects physical information acquiring from the sensor nodes by wireless to utilize it for the purpose of monitoring/controlling etc. The WSN technology has been applied through a variety of fields including defense, medicine, road transport, security, and realizes advancement of the living standard[1,2,3,4].

In particular, the agriculture field improves productivity and increases customer's reliability by utilizing the RFID/WSN technology to apply into production, shipment and distribution stages of agricultural products, and carry out the advanced enhancement of agriculture[5,6].

For the recent domestic livestock industry, its scale of breeding and the number of entities is increased to grow greatly in quantitative terms, however, it experiences difficulties due to the feed price advance caused by rising the international grain price, and it is unavoidable to have a head-to-head contest with the advanced livestock countries due to signing of the FTA. In particular, it suffered a vast damage by various livestock diseases such as the foot-and-mouth disease, AI etc., and it led to the increase of mortality rate to bring economic damages of livestock farmhouses[7,8].

In order to solve such problems of domestic livestock industry, studies have been actively carried out on the livestock farm environment monitoring and livestock disease forecasting system utilizing the WSN technology, and this system could increase productivity and produce high-quality livestock products by creating the optimum livestock breeding environment and reducing the mortality rate and the production cost[9,10].

However, the existing developed systems have been developed in terms of a single livestock farm to be difficult to carry out the integrated management of multiple livestock farmhouses, and there is a problem that the rapid initial reaction is difficult because the information on the disease could not been understood in advance when the livestock diseases occur[11].

Therefore, this paper proposes a livestock control system based on the distributed platform in order to solve these problems. The proposed system connects computers with peripheral devices distributed in multiple places via the communication networks, could construct the complicated information system rapidly through modularization of components, and provides user management, output management, environmental control based data management, notification event management and data output management. Thus, by carrying out the integrated management of multiple livestock farmhouses' monitoring system and sharing their information, the rapid reaction is possible when the livestock

[2] Corresponding Author

diseases or the abnormal symptoms of the livestock farm occur, so that it could minimize damages.

This paper is organized as follows. Chapter 2, 3 and 4 describe the livestock control system's structure, service process and system implementation, respectively, finally, Chapter 5 finishes with a conclusion.

2 Design of the proposed system

2.1 System structure

The livestock control system proposed in this paper is based on the distributed processing system, and its structure is as the figure 1.

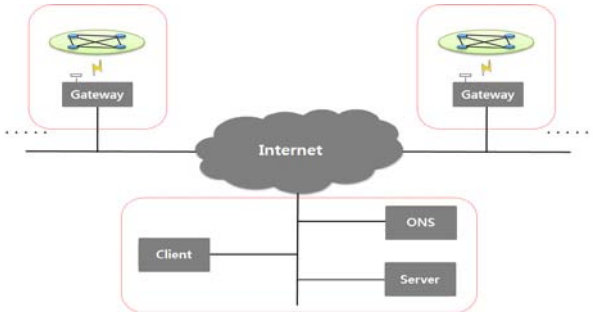


Figure 1. The Structure of the distributed processing system

The overall system consists of environmental sensors collecting environment data in a livestock farm, sink nodes collecting them to transmit, and a control system providing control and various information with the collected data.The control system not only carries out the integrated management of data collected in real time but also analyzes the collected information to provide information on functions, which could keep the optimum environment through the environmental control devices in the livestock farm, and whether or not to be abnormal, and the component for supporting various application services to user's environment in the livestock control system is composed of 5 modules such as the user management, device control management, environment control reference data management, notification event management and data output.

The figure 2 shows the components and interface structure of the livestock farm control system. The user management module provides information on users managing each livestock farm, the environment data management module is one to carry out the integrated management of each livestock farm's environmental information, the environmental information reference data management module provides reference values for creating the optimum environment in the livestock farm, the notification event management module is one to inform each user via a communication method such as SMS etc. when the important event such as livestock diseases occurs, and the data output management module is an interface one to deliver information necessary to users and represent it on the screen through the GUI.

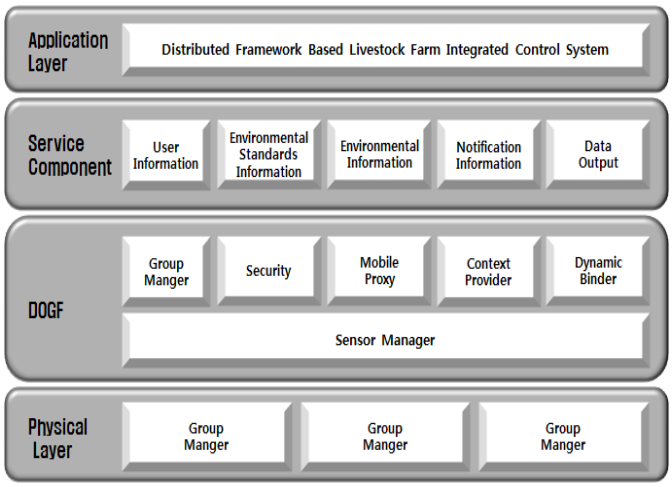


Figure 2. Components and interface structure of the livestock farm control system

2.2 System process

Data is collected through the environment sensors installed in the livestock farm, and then the collected data values are stored in the DB, the livestock farm's environment is controlled as the optimum environment through the environmental control devices. At this time, the operating procedure of each component is as follows. The user management module distinguishes the corresponding user with the stored data, the data output management module instructs to output data to the GUI screen of the corresponding user, the environment control reference data management module analyzes whether the collected data value is compared with the reference value to carry out the environment control through the control devices, the notification event management module informs users when it is different from the reference value or there is a problem in the livestock farm, and data is output on the user's GUI screen through the data output management module. The figure 3 shows the process of the distrivuted processing system.

3 Service process

This distributed platform based livestock farm control system provides the livestock farm management service, which offers the environment control and monitoring service for individual livestock farm, and the event notification service that quickly informs all users when capturing an abnormal symptom based on the collected environmental information.

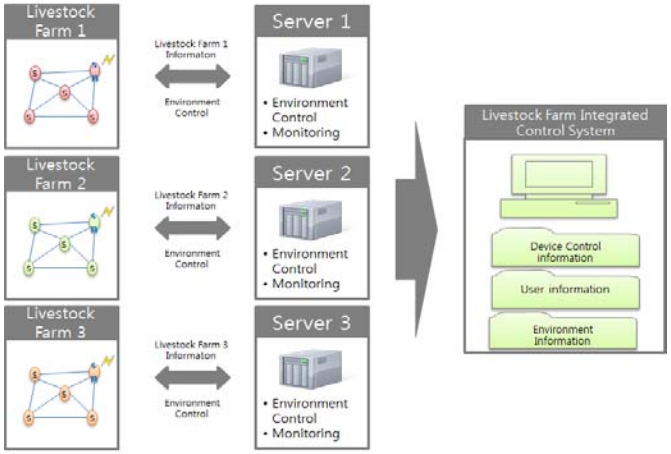


Figure 3. The Process of the distributed processing system

3.1 Livestock farm management service

The livestock farm management service collects temperature, illuminance and humidity etc. of each livestock farm through the environment sensors installed in the livestock farm, and the collected environmental information is compared and analyzed on the server managing each livestock farm to use for controlling as the optimum environment.

The automatic control service in the figure 4 stores information collected from the livestock farm into the DB, the control system calls it to compare the reference values stored in the DB with the collected data, and then it carries out the environment control suitable to the reference values.

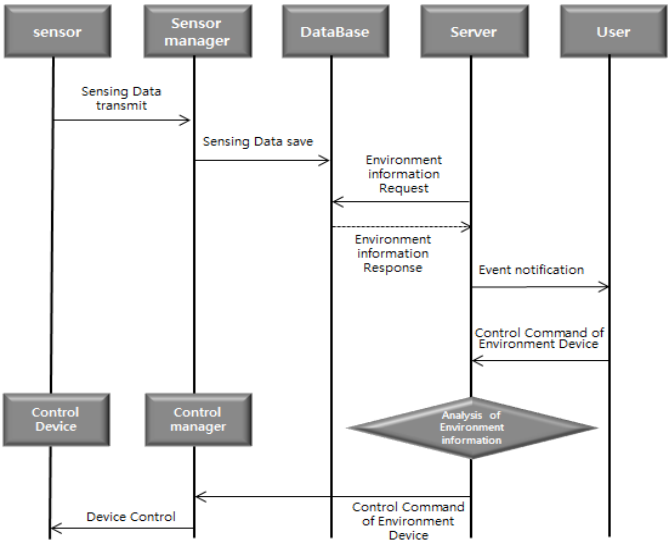


Figure 4. Auto control service of livestock farm

The manual control service in the figure 5 collects environmental data inside and outside the livestock farm through the environment sensors, the control system transmits to users in real time, and users could control the facility after seeing the received data.

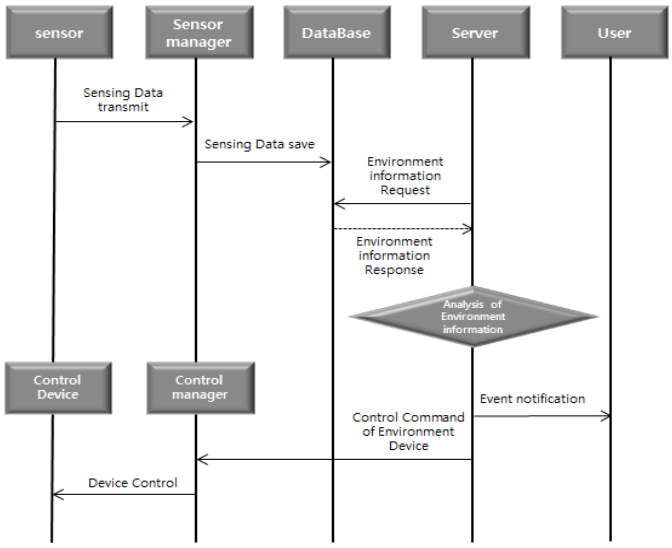


Figure 5. Manual control service of livestock farm

3.2 Event notification service

The event notification service is one to prevent the livestock diseases or the dangerous condition of the livestock farm in advance by informing users in real time when capturing an abnormal symptom in the livestock farm. The environmental information collected from individual livestock farm is transmitted to the control system, and the control system analyzes the collected environmental information to send an event to all users if capturing an abnormal symptom in the certain livestock farm. In addition to this, a user could inform all the users by requesting an event directly. The figure 6 shows the operating process of the event notification service.

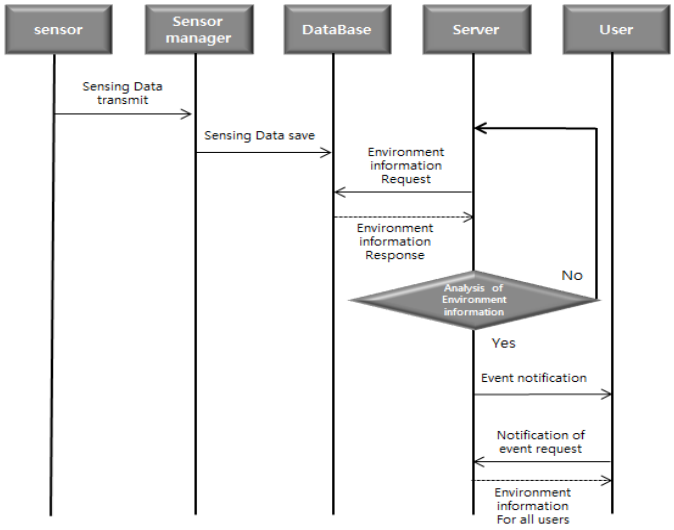


Figure 6. The process of event notification service

4 Implementation of the proposed system

4.1 Individual livestock farm management system

In order to implement the distributed platform based livestock farm control system proposed in this paper, the control system is constructed first for each individual livestock farm. The figure 7 is the block diagram of individual livestock farm management system.

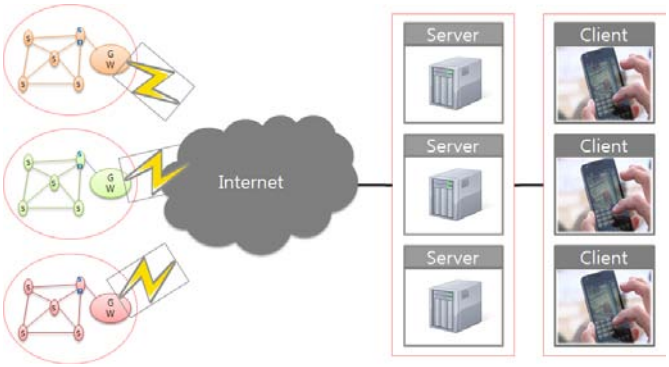


Figure 7. Block diagram of individual livestock farm management system

In the individual livestock farm management system, environment sensors are installed in the livestock farm to collect the livestock farm environment information such as illuminance, temperature, humidity etc. affecting the breeding of livestock, and a GW is installed to send the environmental data collected from the sensors to the individual livestock farm management server. The installed environment sensors transmit the collected livestock farm environmental information at regular intervals, which is sent to the livestock farm management server through the GW. The transmitted livestock farm environment information is processed into the format that could be stored in the DB, its unit is converted to correspond with the measurement element, and the processed data is stored in the DB by using the update query through the sensor manager of the livestock farm management server. In addition, the IP based surveillance cameras are installed to collect image information of the livestock farm and the livestock, and they are used to investigate the cause when there is a theft or accident, or to check current conditions of the livestock farm. Images collected from the surveillance camera are transmitted to the livestock farm management server, and they are divided into the livestock farm ID and the camera number etc. to store in the DB. In order to control the livestock farm conditions such as illuminance, temperature, humidity and CO2 etc. affecting the breeding of livestock, the livestock farm control facility such as lighting, humidifier, fan heater, air conditioner, ventilator etc. are installed, and the relay modules are installed to control them by wireless. The figure 8 is the livestock farm model.

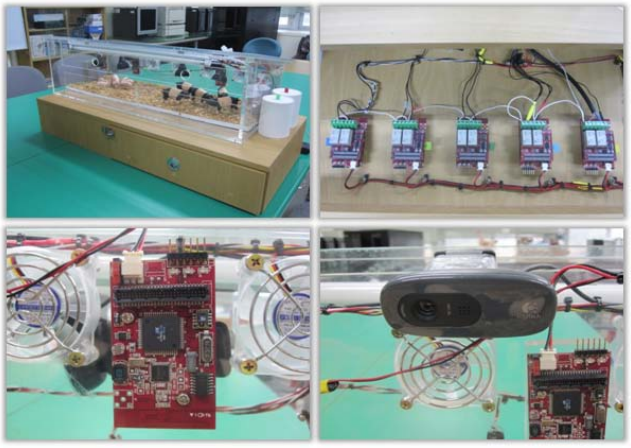


Figure 8. A prototype of livestock farm

The figure 9 is livestock farm management system's GUI applying the individual livestock farm management system.

Figure 9. Livestock farm management system's GUI

4.2 The Integrated livestock farm control system

The integrated livestock farm control system stores the livestock farm environment information collected from the individual livestock farm management system in the DB, and carries out the integrated management and comparative analysis of the collected environment data values to determine the abnormal symptom of the livestock farm. When an abnormal symptom occurs, the server could inform it to every user through the message transmission service and the GUI of devices that could communicate, and could inform the certain livestock farm's abnormal symptom to each user according to a user's request.

In order to construct the integrated livestock farm control system, the WAS used Tomcat-6.0.20, and the DB used the Mysql of version 5.0, the most stable one of currently released ones. The figure 10 is the GUI of the integrated livestock farm control system, which ① provides information on each user and the livestock farm information, ② represents the livestock farm environment information collected from the environment sensors installed in each livestock farm, and ③ indicates conditions such as the occurrence time, environment information etc. when an abnormal symptom occurs in the certain livestock farm.

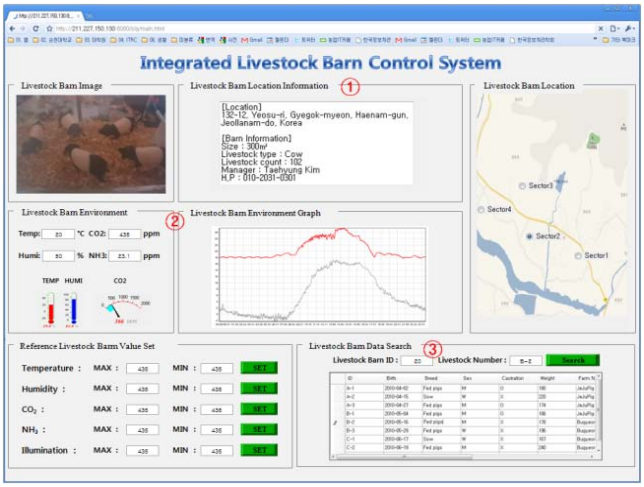


Figure 10. The GUI of the integrated livestock farm control system

The services provided through the GUI are offered by 5 modules mentioned earlier, the user management module distinguishes the corresponding user, the data management module carries out the integrated management of each livestock farm's environmental information, the environment value reference data management module compares the collected environmental information with the reference value to determine whether or not to be abnormal, and the event management module indicates it to users on the GUI through the data output management module when it is greatly different from the reference value or there is a problem in the livestock farm. The livestock farm control system works with these operations.

The figure 11 is a graph that represents the environment data measured for each livestock farm as the proposed integrated livestock farm control system was operated.

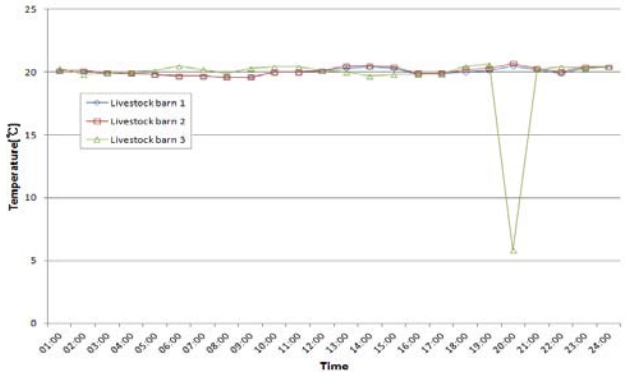


Figure 11. The graph of measured environment data

It could be known that the livestock farm A and B continue to keep uniform temperature and humidity condition, and the livestock farm C keeps uniform temperature and humidity like A and B but the rapid temperature change occurs at 20:02 through the graph above.

5 Conclusions

This study constructed the integrated livestock farm control system based on the distributed platform as a control system to carry out the integrated management of livestock farm environment in the ubiquitous agriculture environment.

It listed building blocks required for implementing the proposed system, interaction between the building blocks, operating process of the designed system, implementation and results of the system, the software structure is based on the distributed framework, the services to support the livestock farm control service were defined, it made the complicated information system could be rapidly constructed by modularizing the system components, and it is composed of the user management, output management, environment control reference data management, notification event management and data output management modules.

It is considered that the proposed system could give much assistance to high productivity and production of high-quality livestock by carrying out the integrated management of multiple livestock farmhouses' monitoring systems and sharing the information, in addition to this, it is expected to minimize damages by enabling to rapidly deal with when there is livestock disease or an abnormal symptom in the livestock farm. It is expected that domestic livestock industry would be competitive because this system could provide reduction of labor force, production of high quality livestock and improvement of productivity etc. to the labor-intensive livestock industry.

6 ACKNOWLEDGEMENTS

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

7 REFERENCES

1. Ian F. Akyildiz et al., "A survey on Sensor Networks," IEEE Communications Magazine, Vol.40, No.8 **(2002)**
2. Seok-soo Kim, Gilcheol Park, Kyungsuk Lee, Sunho Kim, "Ubiquitous Military Supplies Model based on Sensor Network", International Journal of Multimedia and Ubiquitous Engineering, Vol. 1, No. 1, March, **(2006)**
3. Minseong Ju, Seoksoo Kim, "Logistic Services Using RFID and Mobile Sensor Network", International Journal of Multimedia and Ubiquitous Engineering, Vol. 1, No. 2, June, **(2006)**
4. Paul Golding, Vanesa Tennant, "Evaluation of a Radio Frequency Identification (RFID) Library System: Preliminary Results", International Journal of Multimedia and Ubiquitous Engineering, Vol. 3, No. 1, January, **(2008)**
5. Hwang J.H., Shin C.S., Yoe H., "Study on an Agricultural Environment Monitoring Server System using Wireless Sensor Networks", Sensors 2010, 10, 11189-11211 **(2010)**
6. Boo-man Jeong, "Foreign u-Farm Service Model casebook", Korea National Information Society Agency, NCA V–RER-06005, Seoul, Korea **(2006)**
7. Yong-hl Yoo, Doo-hwan Kim, "The current state of automation in pig house establishment and prospection", Korea society for livestock housing and environment, p29-p47(19) **(2006)**
8. Hwang J.H., Yoe H., "A Study on the Context-Aware Middleware for Implementing Intellingent Service in Ubiquitous Livestock Barn based on Wireless Sensor Networks", Sensors 2011,11,4539-4561 **(2011)**
9. Hwang J.H., Yoe H., "Study of the Ubiquitous Hog Farm System Using Wireless Sensor Networks for Environmental Monitoring and Facilities Control", Sensors 2010, 10, 10752-10777 **(2010)**
10. Yongyun Cho, Hyun Yoe, "A Service Scenario Based on a Context-Aware Workflow Language in u-Agriculture", Communications in Computer and Information Science, Volume 151. 240-244 **(2011)**
11. Jeong-Young, Soo-Young kim, Dong-oun Choi, "Developement of Cattle Activity Monitoring System Model", Korea Entertainment industry Association, pp. 152-155 **(2011)**

Hierarchical Polling-based MAC scheme for Wireless Body Sensor Network

Shusaburo Motoyama
Faculty of Campo Limpo Paulista (FACCAMP)
Campo Limpo Paulista, São Paulo, Brazil

Abstract - *A hierarchical polling-based access scheme for Wireless Body Sensor Network (WBSN) is proposed in this paper. The proposed access scheme is structured in hierarchy to collect data from sensor nodes inserted in human body. In first level of hierarchy the sensor nodes are divided into groups and sensor nodes of each group communicate with a sink node which collects data by using polling technique. In second level, the sink nodes communicate with master node which collects data by also using polling technique. The sensor nodes from first level are provided with only single buffer to store data and intending to save energy by using small capacity buffer. The sink nodes have larger buffers and the master node uses exhaustive polling technique. The performance of the proposed scheme is studied using mathematical models known in the literature. The numerical analyses show that the proposed scheme can be efficient for WBSN application.*

Keywords: wireless, body sensor network, MAC, polling, mathematical modeling.

1 Introduction

Wireless Body Sensor Network (WBSN) is composed of tiny electronic devices called sensors which are attached to the human body for remote monitoring of vital signs of human being. A sensor with processing and communication capabilities is denoted sensor node.

Since the sensor nodes of a WBSN can be placed under the human skin of difficult accesses and due to the small size of the nodes and the limited energy storage capacity of battery, the sensor nodes must mainly save energy.

One of the tasks performed by sensor node that most spends energy is the communication. The sensor nodes must communicate externally with some device (sink node) for the transmission of collected data. Since many sensor nodes can be placed at human body, if more than a sensor node begins to transmit packets simultaneously, collisions will occur, and packets must be retransmitted. The packet retransmission can be an energy consuming process. Thus, an efficient medium access control (MAC) mechanism for collision reduction or elimination is fundamental for good operation of a WBSN. Furthermore, the use of sink nodes as centralized nodes for data collection from sensor nodes is more convenient because it simplifies the communication protocol, and it is appropriate for the collision reduction.

A MAC scheme based on polling technique and using sink nodes in a hierarchical structure is proposed in this paper.

This paper is divided into five sections. In the second section, the related works to this paper are presented. The proposed hierarchical polling based MAC scheme and some operations of the sources for WBSN are described in section three. In the fourth section, the mathematical modeling and performance analyses of proposed MAC scheme are carried out. Finally, the main conclusions are presented in section five.

2 Related Works

Many MAC schemes proposed in the literature for the WBSN are based on the standard 802.15.4 with beacon - enabled star configuration which provides very low energy consumption [1]. However, since the scheme is not designed for WBSN applications some drawbacks have been pointed out [2], and recently many schemes of MAC protocols specifically for WBSN have been proposed [2-16]. Some proposals are based on the variations of standard 802.5.4 [5], [8] and [11], and others are based on TDMA access technique [3], [4], [7], [10], [14], [15] and [16]. Each of the proposals explores some special features based on medical needs. For instance, in [3-4] to deal with the light and heavy loads in normal and urgent situations, a context aware MAC is proposed. To guarantee QoS of a WBSN, a MAC protocol based on random access technique is proposed in [12]. In the proposal presented in [10], the heart beating is used for the purpose of clock synchronization. In [6], the beacon used for wake-up sensor nodes is used for battery charging, increasing the network life time.

Recently, the standard 802.15.6 has been proposed for the wireless body area network [17]. This standard has three modes of operation: beacon mode with beacon period superframe boundaries, non-beacon mode with superframe boundaries and non-beacon mode without superframe boundaries [17]. The beacon mode is designed for medical and non-medical applications and has been the object of main standardization.

The non-beacon mode without superframe boundaries has been less explored. In [20] and [21] MAC schemes using this mode were proposed. Both proposals are based on polling access scheme that avoids the need for periodical synchronization. In [20] a flexible a scheme that exchanges the normal polling operation mode to the urgent polling operation mode in case of emergency needs is proposed. In [21] the

polling scheme using realistic sensor node models for WBSN are investigated by simulation.

The main objective of this paper is to propose an efficient polling access scheme for WBSN with a hierarchical structure of sink nodes to cope with the fact that the human skin is not a good electrical conductor and a sensor node may not have a direct communication with a sink node.

3 WBSN and Proposed Mac scheme

A WSN composed by biological sensors designed to monitor vital signs of the human body is usually called Wireless Body Sensor Network (WBSN). This network, composed of many sensor nodes with processing, communications and limited energy capabilities, has the function of monitoring various activities of the human body, facilitating the assistance of patients who require remote medical attention.

Many sensors can be inserted into different regions of human body as the head, the thorax, the upper members, the abdomen and lower members.

Basically, there are two classes of MAC mechanisms: ordered and random access. In the former, a centralized node (or sink node) is used to organize the conflict for the access of the output link. In the latter, each node transmits packets randomly to the physical medium and collisions may occur. A centralized node is more convenient for the WBSN because collisions can be avoided, thus saving energy.

It is known that human skin is not a good electric conductor so that some implanted sensor nodes may not have direct communication with sink node. For instance, the sensor nodes implanted at back of body may have some difficult to communicate to the sink node placed in a belt at front of human body. To cope with this problem we propose the use of two or more sink nodes placed in a belt at different locations, so that a group of sensor nodes at back can communicate with the sink node located at back and a group of sensor nodes in front can communicate with sink node placed at front. To collect the data from sink nodes it is provided another node denoted master node. To collect data, the sink nodes as well master node use the polling technique. This structure will be denoted hierarchical polling-based access scheme as shown in Fig. 1.

The communication protocol for hierarchical polling-based access scheme proposed in this paper can be simplified using the fact that the sensors are located in close proximity to the sink node. The sink node broadcasts a packet carrying the sensor node number to be investigated, i.e., it is sending an authorization to a sensor node to transmit the packets. This authorization packet has in its header enough bits for bit and frame synchronizations of a sensor node. If a sensor node has packets to transmit, it recognizes its sensor node number and starts to transmit. After the transmission, the sensor node waits for acknowledgment in case of the need for retransmission. If a sensor node doesn't have packets to transmit, it can keep the transceiver in an off state and only switches to an on state in the case of packet transmission. The sink node recognizes that a sensor node is in off state after the transmission of an authorization packet and waits for a while. If the data packet from the polled sensor node doesn't arrive, the sink node infers that the node doesn't have packets to transmit and goes to other sensor node in sequence to poll. Thus, in this proposed protocol, the sink node does almost all of the communication functions, leaving the sensor node only the packet transmission function. This same communication protocol can be used in second level, that is, when the master node polls sink nodes to get the data. For WBSN, just two hierarchical levels may be enough.

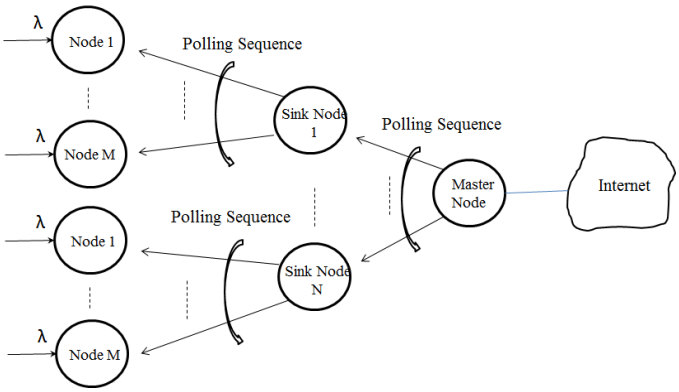


Figure 1. Hierarchical polling-based structure.

A sensor node can save energy by keeping the transceiver in an off state when is not transmitting packets. Another way to save sensor node energy is to implement functions at the node in which the sensors send only relevant information to the event observer. For instance, a sensor monitoring body temperature sends only measurements which are above a certain value. The other criterion could be to transmit just the packets that are outside of a certain range.

Some different types of functions as shown in Tab. 1 can be implemented. In a sensor node implemented with threshold function type, only packets carrying information above a threshold are sent. In the case of an out-range function type, the sensor nodes sends packets with information that is outside of a certain range. For example, in a sensor node responsible for the heart-beat monitoring, it is desirable that only the measurements representing risks for a patient's life be sent. For instance, the normal heart beat for a particular patient is 100 beats per minute and it can vary between 80 and 120 beats per minute, then should be sent only measurements less than 80 or greater than 120 beats per minute.

In the above-mentioned criteria, it is possible that there may be a hiatus where the nodes do not transmit any packet because no measurement satisfies the specified criteria for the transmission. Thus, to avoid a long silence of the sources, the discarded packets are counted and when this counting reaches a certain value, the next packet is sent, regardless if the measurement satisfies the criteria established or not. These functions represent the controlled threshold and controlled out-range functions in Tab. 1.

Table1. Function Types.

Function type	Description
Threshold	Send packets carrying information above a threshold.
Controlled Threshold	Send packets containing information above a threshold or next packet when discarded packets reached a predefined number.
Out-range	Send packets carrying information outside a certain range.
Controlled Out-range	Send packets satisfying Out-range criterion or next packet when discarded packets reached a predefined number.

4 Performance Evaluation

To analyze the proposed WBSN based on the hierarchical polling access control, the following assumptions are adopted. Each sensor node is using some kind of function type described in Tab.1, so that to only relevant packets containing measurements above a certain value or outside a certain range are randomly sent by sensor nodes. Using this approach, the Poisson distribution of rate λ can be approximately adopted at output of each sensor node. A deterministic packet length distribution with average $E\{X\}$ bits long is adopted and is the same for all nodes. The channel capacity from sensor nodes to the sink node (or vice versa) or sink nodes to the master node (or vice versa) is R bits/sec. The number of sensor nodes in each group will be considered M and the number of sink nodes is N.

The walk time, w, between two consecutive sensor nodes in the polling is constant and the same for all nodes. The propagation time of a sensor node to the sink node is the same for all nodes and is included in walk time.

For the performance analysis of hierarchical polling, it can be considered that each group of sensor nodes and a sink node together is independent of each other, so that each group can be analyzed independently. Only a case considering single buffer in the first level will be analyzed. In that case, just a packet will be transmitted when a sensor node is polled and the input of sink node can be considered as approximately Poisson distribution with rate $(1 - P_L)M\lambda$. Assuming a large buffer size at sink nodes, an exhaustive service case will be considered in the second level, as shown in Fig. 2.

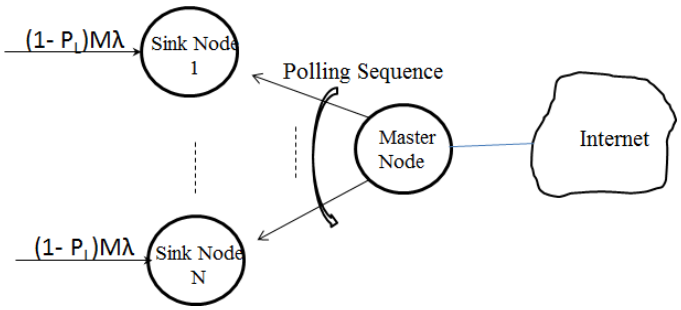


Figure 2. Second level performance model.

4.1 Firt level performance - single-buffer case

The use of a small buffer size in a WBSN is important because the energy can be saved. In this subsection the analysis of one size buffer for hierarchical WBSN is carried out.

The expression of waiting time for the polling technique using single buffer case has been solved and the expression is given by [22]

$$E\{W_1\} = (M-1)\frac{E\{X\}}{R} - \frac{1}{\lambda} + \frac{M^2 w}{E\{Q\}}, \quad (1)$$

where

$$E\{Q\} = \frac{M\sum_{n=0}^{M-1}\binom{M-1}{n}\prod_{j=0}^{n}\{\exp[\lambda(Mw + jE\{X\}/R] - 1\}}{1 + \sum_{k=1}^{M}\binom{M}{n}\prod_{i=0}^{k-1}\{\exp[\lambda(Mw + iE\{X\}/R] - 1\}}. \quad (2)$$

$E\{Q\}$ represents the mean number of packets served in a polling cycle.

The mean cycle T_{c1} for first level is given by [22]

$$E\{T_{C1}\} = Mw + E\{Q\}\frac{E\{X\}}{R}. \quad (3)$$

The transfer time for first level is given by

$$E\{T_1\} = \frac{E\{X\}}{R} + E\{W_1\}. \quad (4)$$

The packet loss probability for the blocked packets when the buffer has already stored a packet is given by [22]

$$P_L = \frac{E\{T_1\}}{E\{T_1\} + 1/\lambda}. \quad (5)$$

For illustration of the above equations, a numerical example will be given. Let the packet length be $E\{X\}$ = 900 bits, the number of sensors be M = 20, the channel capacity from nodes to sink node or vice-versa be R = 20 kbps and the authorization packet length be 10% of data packet $E\{X\}$. The above data packet length used is the average packet length obtained from [2], [18] and [19]. The transmission time of authorization packet is 90 / 20000 = 4.5 msec. Assuming the bit synchronism time at a node is equal to 2 msec, the walk time can be calculated as w = 4.5 + 2 = 6.5 msec.

Defining the first level input load as

$$S_1 = \frac{M\lambda E\{X\}}{R}, \quad (6)$$

Eqs. 1, 2 and 4 can be calculated numerically for each value of S_1. For instance, for S_1 = 0.5, the value of λ is 0.556, and using these values in Eq. 2 and solving numerically the value of

$E\{Q\}$ will be 2.411 packets and $E\{W_1\}$ = 133.25 msec. The transfer time is 178.25 msec and the mean cycle time and packet loss probability are 238.51 msec and 9.01%, respectively.

The effective arrival rate λ_{eff} is given by

$$\lambda_{eff} = \lambda(1 - P_L). \quad (17)$$

Thus λ_{eff} is 0.506 packets/sec and the number of packets waiting in the buffer is by Little rule $E\{Nq\} = \lambda_{eff} E\{W_1\}$ = 0.0674245 pckts.

To overcome high loss of packets and long transfer time, smaller values of input load can be used. Figs. 3 and 4 show the transfer time and loss probability for various values of input load and number of sensors nodes.

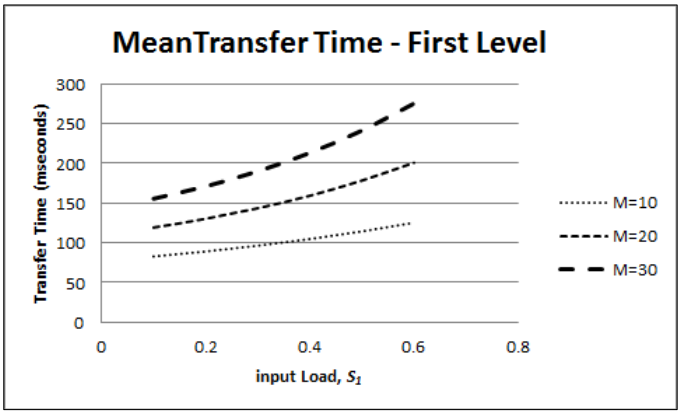


Figure 3. Mean transfer time of first level in function of input load.

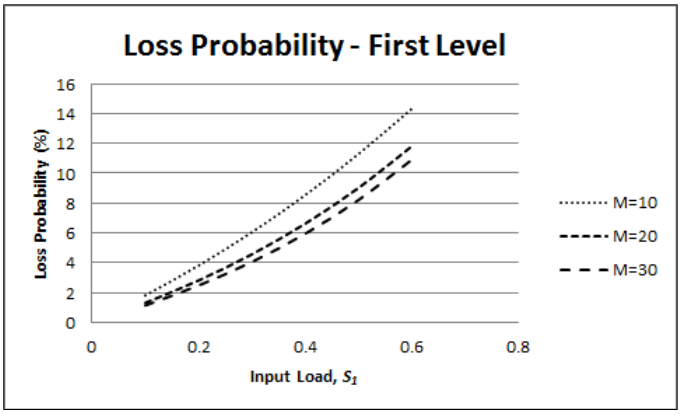


Figure 4. Loss probability of first level in function of input load.

As can be noticed, the packet loss is higher for M =10 than M =30 because the input rate λ is inversely proportional to M obeying Eq. 6. For total input rates about 0.1 and 0.2 the loss probabilities are less than 2% and 4%, respectively, regardless of the value of M, which could be a good operation of the WBSN whilst saving energy.

The difference between input rate and effective input rate at each node becomes greater as the total input rate increases as shown in Fig. 5.

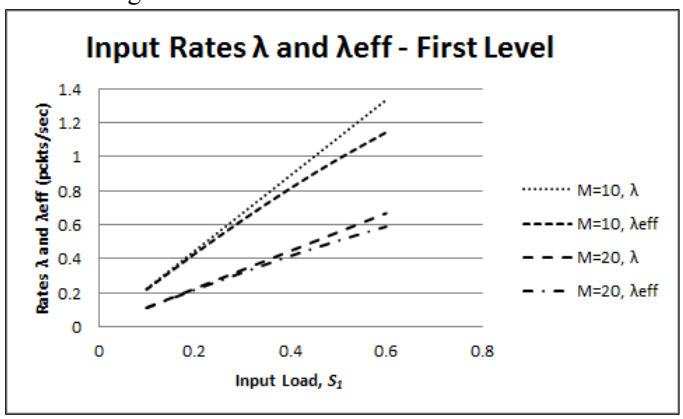


Figure 5. Input rates λ and λeff of first level in function of input load.

On the other hand, the mean polling cycle time is directly proportional to the number of nodes and increases as the nodes grow as can be seen in Fig. 6. The cycle time also increases as the total input load is increasing.

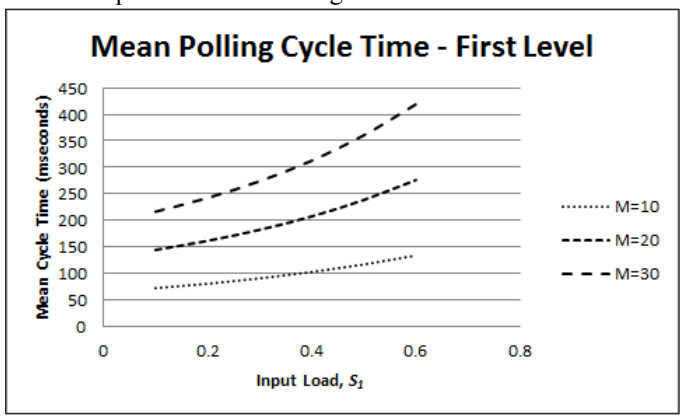


Figure 6. Mean polling cycle time of first level in function of input load.

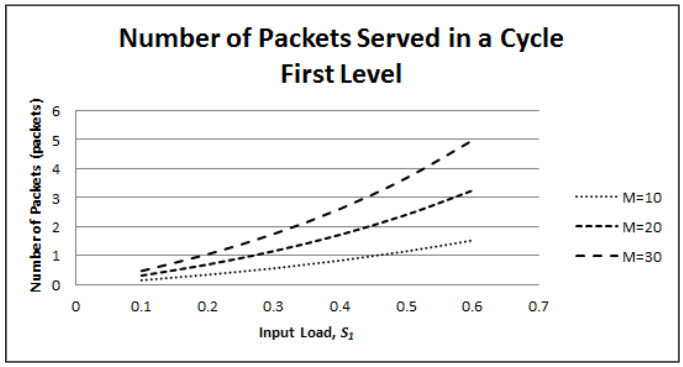


Figure 7. Number of packetes served in a cycle versus input load.

Figure 7 shows the number of packets served in a cycle. The mean number of packets served is small, ranging from 0.15 packets for S=0.1 and M=10 to 5 packets for S=0.6 and M=30 .

4.2 Second level performance - exhaustive service case

By assuming an exhaustive polling service, i.e., when a sensor node is inspected all the packets are served including those arriving during the service time, the average cycle time is given by

$$T_{c2} = \frac{Nw}{1-S_2}, \quad (7)$$

where S_2 is given by

$$S_2 = \frac{NM(1-P_L)\lambda E\{X\}}{R} = S_1 N(1-P_L) \quad (8)$$

The stability condition is given by

$$S_2 < 1 \;\Rightarrow NM(1-P_L)\lambda < \frac{R}{E\{X\}}. \quad (9)$$

The stability condition means that the polling scheme can complete the cycles without any buffer at nodes having packets waiting for long (infinite) times.

The queuing time in a buffer in the second level is given by [22]

$$E\{W_2\} = \frac{Nw(1-S_2/N)}{2(1-S_2)} + \frac{S_2 E\{X\}}{2R(1-S_2)}, \quad (10)$$

for deterministic packet length.

The assumption of constant walk time between two nodes adopted in Eq. 10 can be explained by the fact that the distances from sensor nodes to the sink node in a WBSN are very small and the authorization packet can simultaneously reach almost all the sensor nodes.

The packet transfer time for the second level is given by

$$E\{T_2\} = \frac{E\{X\}}{R} + E\{W_2\}. \quad (11)$$

The propagation time is neglected in the above equation, assuming that the distance from a sink node to the master node is very short reaching only a few meters.

For illustration of the above equations, a numerical example will be given. Let the packet length be $E\{X\}$ = 900 bits, the channel capacity from sink nodes to the master node or vice-versa be R = 20 kbps and the authorization packet length be 10% of data packet $E\{X\}$, as used in single buffer case, and the number of sink nodes be N = 2. Assuming the bit synchronism time at a node is equal to 2 msec, the walk time can be calculated as w = 4.5 + 2 = 6.5 msec. Assuming total input load of 10%, the numbers of sensor nodes of M = 10 and M = 20 and by using Eq. 5, it can find out that loss probabilities P_L are 1,81% and 1.31%, respectively. Using these values in Eqs. 8, 10 and 11, the transfer times in a buffer are 57.79 msec and 57.83 msec, for M = 10 and M =20, respectively. For an input load of 50%, and M = 10 and M = 20, the transfer times are 254.00 msec and 311.53 msec, respectively. For a load of 50%, the mean polling cycle times using Eq. 7 are 115.23 msec and 144.27 msec for M = 10 and M = 20, respectively.

Figures 8 and 9 show the mean transfer time and mean polling cycle time for other values.

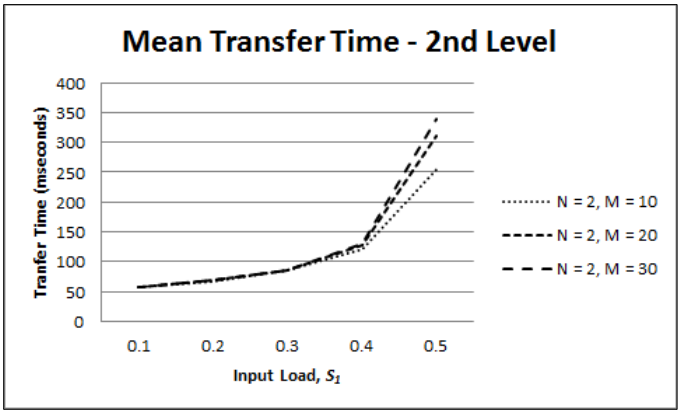


Figure 8. Mean transfer time of second level in function of total input load S_1.

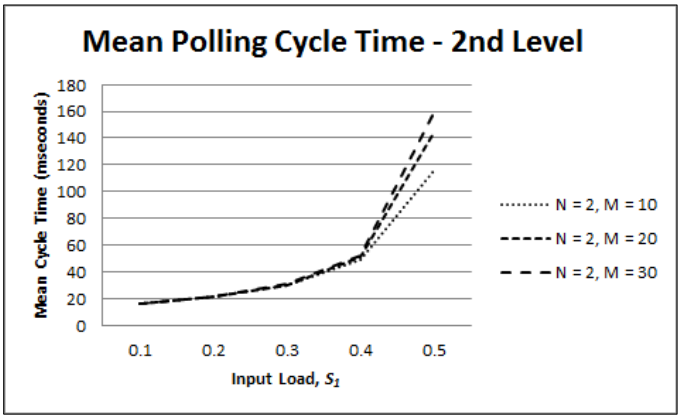


Figure 9. Mean polling cycle time in function of total input load.

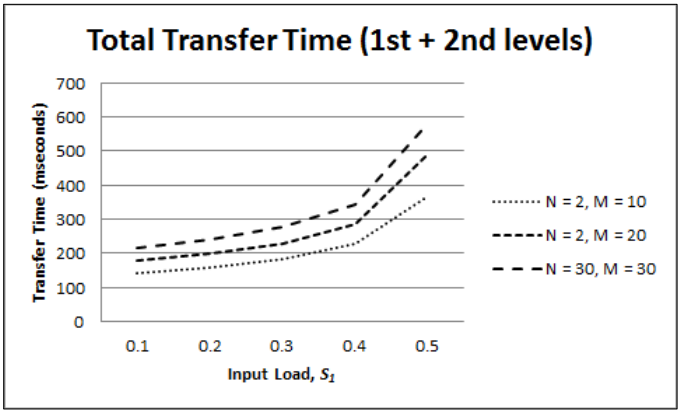


Figure 10. Total meam transfer time in function of total input load.

Figure 10 shows the total transfer time adding first and second levels of hierarchical polling-based MAC scheme. The

curves of Fig. 10 indicate that the transfer times still keep very small, less than 0.5 seconds for almost all input load so that the polling technique may be convenient for the application in WBSN.

5 Conclusions

A hierarchical polling-based access scheme for Wireless Body Sensor Network (WBSN) was proposed in this paper. The main technical advantage of the polling access mechanism is the non-necessity of frame synchronization as the TDMA technique requires, and it has centralized control of sensors convenient for WBSN. The proposed access scheme uses the sink nodes in a hierarchical structure so that only the sensor nodes having direct communication with a sink node are served. The proposed communication protocol is simple, giving to the sink node majority of controls and leaving with the sensor nodes only the function of packet transmission, thus saving energy.

The mathematical modeling of the proposed scheme was done using single buffer at each sensor node in the first level of hierarchy and an infinite buffer for sink nodes in second level.

The analysis showed that for the first level using single buffer the transfer times can be kept very small. However, packet loss for high load (above 0.4) is prohibitive and must be avoided. The analysis for second level using exhaustive service is dependent of number of sink nodes and also the number of sensor nodes of first level. Considering only two sink nodes, which we consider appropriate for WBSN, for any value of M (10, 20 and 30) and input load S_1 of up to 0.5 the transfer times are less than 350 ms for all cases. However, for values above 0.5 the operation is becoming unstable and the transfer times are very larger. The total transfer times considering the first and second levels for load up to 0.5 are less than 500 ms for all cases, showing that the hierarchical polling scheme can be convenient for WBSN applications. It must be pointed out that used link capacity is not high, mainly in the case of communication between sink nodes and master node which was only 20 kbps. In this segment a higher transmission capacity can be provided so that a better network performance can be expected.

6 Acknowledgment

This work was supported by São Paulo Research Foundation (FAPESP) under grant No 2011/12463-0.

7 References

[1] B. Latré, B. Braem I. Moerman, C. Blondia and P. Demeester, "A survey on wireless body area networks," Wireless Networks, Volume 17 Issue 1, January, 2011, Kluwer Academic Publishers Hingham, MA, USA

[2] B. Otal, L. Alonso and C. Verikoukis, "Towards energy saving wireless body sensor networks," in Health Care Systems" Proceedings of IEEE International Conference on Communications (ICC 2010), Second International Workshop on Medical Applications Networking (MAN 2010), Capetown, África dos Sul, 2010.

[3] Z. Yan and B. Liu, "A context aware MAC protocol for medical wireless body area network," in Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International, pp. 2133-2138.

[4] B. Liu, Z. Yan and C. W. Chen, "CA-MAC: A Hybrid context-aware MAC protocol for wireless body area networks," in 13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom), 2011, pp. 213-216.

[5] X. Zhu, S. Han, P. Huang, A.K. Mok and D. Chen, "MBStar: a real-time communication protocol for wireless body area networks," in 23rd Euromicro Conference on Real-Time Systems (ECRTS), 2011, pp. 57-66.

[6] D. Layerle and A. Kwasinski, "A power efficient pulsed MAC protocol for body area networks," in IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2011, pp. 2244-2248.

[7] Y. Tselishchev, "Designing a medium access control protocol for body area networks," in IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011.

[8] L.M. Borges, F. J. Velez and A.S. Lebres, "Performance evaluation of the schedule channel polling MAC protocol applied to health monitoring in the context of IEEE 802.15.4," in 11th European Wireless Conference - Sustainable Wireless Technologies (European Wireless), 2011, pp. 94-101.

[9] S. Kutty and J.A. Laxminarayan, "Towards energy efficient protocols for wireless body area networks," in International Conference on Industrial and Information Systems (ICIIS), 2010.

[10] L. Huaming and T. Jindong, "Heartbeat-driven medium-access control for body sensor networks," IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 1, January 2010, pp. 44-51.

[11] K. A. Ali, J.H Sarker and H.T Mouftah, "Urgency-based MAC protocol for wireless sensor body area networks," in IEEE International Conference on Communications Workshops (ICC), 2010.

[12] A. A. Khaled, H. S. Jahangir and T. H. Mouftah, "QoS-based MAC protocol for medical wireless body area sensor networks," in IEEE Symposium on Computers and Communications (ISCC), 2010, pp. 216-221.

[13] X. Zhang,H. Jiang, X. Chen, L. Zhang, Z. Wang, "An energy efficient implementation of on-demand MAC protocol," in Medical Wireless Body Sensor Networks" IEEE International Symposium on Circuits and Systems, 2009. ISCAS 2009, pp. 3094-3097.

[14] S. Marinkovic, C. Spagnol and E. Popovici, "Energy-efficient TDMA-based MAC protocol for wireless body area networks," in Third International Conference on Sensor Technologies and Applications, 2009. SENSORCOMM '09, pp. 604-609.

[15] G. Fang, E. Dutkiewicz, "BodyMAC: energy efficient TDMA-based MAC protocol for wireless body area networks," in 9th International Symposium on Communications and Information Technology, 2009. ISCIT 2009, pp. 1455-1459.

[16] S. S. Oliveira and S. Motoyama, "Applications oriented medium access control protocols for wireless sensor networks," IEEE Latin America Transactions, v. 7, Issue 5, 2009, pp. 586-593.

[17] K. S. Kwak, S. Ullah and N. Ullah, "An overview of IEEE 802.15.6 standard," in ISABEL, 2010, Rome, Italy.

[18] O. Omeni, A. Wong, A.J. Burdett and C. Toumazou, "Energy efficient medium access protocol for wireless medical body area sensor networks," IEEE Transactions on Biomedical Circuits and Systems, Volume: 2 , Issue: 4, 2008, pp. 254-259.

[19] R. Gravina, A. GuerrierI and A. Fortino, "Development of body sensor networks applications using SPINE," in IEEE International Conference on Systems, Man and Cybernetics. Singapura, 2008.

[20] S. Motoyama, "Flexible polling-based scheduling with OoS capability for Wireless Body Sensor Network" in Local

Computer Networks Workshops (LCN Workshops), 2012 IEEE 37th Conference, pp. 745-752.

[21] T.A. Pazeto, L. F. Refatti, S. Motoyama, "Polling-based Medium Access Control Scheme for Wireless Body Sensor Network" in International Conference on Wireless Networks - ICWN-12, 2012, Las Vegas. p p. 87-93.

[22] H. Takagi, "Analysis of Polling Systems", The MIT Press Cambridge, Massachusetts London, England, 1986.

An Overhearing Video Transmission for Wireless Sensor Networks

Yeonbo Kim[1], **Seokjin Byeon**[2], **and Byoungchul Ahn**[2]
[1]School of Electronic and Electrical Engineering, Daegu University, Gyungsan, Korea
[2]Department of Computer Engineering, Yeungnam University, Gyungsan, Korea

Abstract - *To transmit multimedia data on WMSNs(Wireless Multimedia Sensor Networks), it is required to use efficient protocols to reduce power consumption. This paper presents an efficient protocol to transfer multimedia data by overhearing messages of nodes and by transmitting next packets during the unused time interval. The proposed method is verified its performance by simulations and experiments. The results shows that the transmission rate of the proposed method 50% higher than that of End-to-end protocol. Also the transmission time is reduced up to 50%. The results of real measurement are very close to those of simulations. The proposed algorithm shows very good performance compared with the End-to-End transmission method. The transmission performance of the proposed method is double that of the End-to-End transmission. The experiment results show that average of the success rate is 93.78 %.*

Keywords: WMSN, End-to-end, Hop-by-hop, Overhearing

1 Introduction

WSNs(Wireless Sensor Networks) are networks consisted of hundreds to thousands sensor nodes. Each node has a microprocessor with small memory, a communication device and sensors with an energy efficient power source. The power source may be batteries or solar cells to run a life time of several months to several years. Typically sensor networks are used to monitor environment events such as temperature, humidity, pollution level and so on. The rapid development of sensors and inexpensive CMOS cameras are allowed for the emergence of so called wireless multimedia sensor networks. WMSN (Wireless Multimedia Sensor Network) is a network of wirelessly interconnected sensor nodes equipped with multimedia devices, such as cameras and microphones, and capable to retrieve video, audio, images, as well as sensor data[1, 2].

2 Related Work

Many researches of WSNs are concentrated on MAC protocols to optimize operations of WSNs. Le, Guyennet and Felea have proposed an overhearing based MAC protocol for WSNs. By using overhearing, the MAC protocol reduces redundant transmissions and energy consumption[7]. This paper is not applied to multimedia transmission. Kanzaki and *et el*. have proposed an overhearing-based transmission to reduce traffics of WSNs. By using overhearing, each node autonomously determines the temporal redundancy of its reading by applying a lightweight interpolation based on the readings acquired by itself before determining the spatial redundancy [8]. This paper does not describe application area. Data of WSNs are transmitted form a source node to a destination node, which is called End-to-End transmission. When the destination node receives a packet, it replies back ACK control signals to the source node[9].

3 Proposed Algorithm

A new algorithm using packet overhearing is proposed. First of all, packet overhearing and its problem are discussed and a reliable transmission method is developed by removing transmission errors. This reliable transmission method is to transmit packets without using control messages signals at the MAC layer to transmit multimedia data for a short transmission time.

3.1 Overhearing collisions

In wireless sensor networks, the wireless signals of a transmitting node propagate all neighbors of the sender. Sometimes it may work as interference signals or make collisions. If the interval of the transmission is adjust, wireless signals work as overhear packets, which make unnecessary consumption of node energy. Packet overhearing is used in synchronous CSMA protocols such as SMAC and TMAC, in which all nodes in a neighborhood wake-up simultaneously to listen to in coming packets. As shown in Figure 1, Node S overhears packets of Node R1, when Node R1 transmits packets to Node R2. If Node R1 relays packets of Node S to Node R2, Node S uses packet overhearing as a reliable transmission signal instead using *ACK* signals. Unnecessary signals do not be used for reliable signal.

In order to implement the overhearing packets, all sensor nodes must process them as control signals to relay packets to a destination. It is required to change the packet handling algorithm of the MAC layer. When a node receives overhearing packets, the node must handle packets although they are not transmitted to the node.

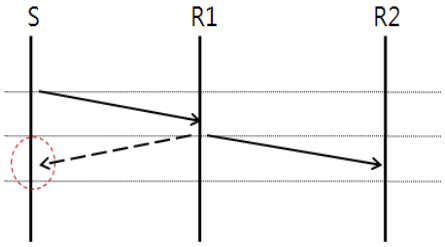


Figure 1. Packet overhearing by transmission of Node R1

After a packet is transmitted, the next packet is retransmitted sometime later. The time interval of packet transmission is important to prevent packet collision. Figure 2 and Figure 3 show examples of packet collision. Figure 2 shows that collision occurs in the middle of two nodes and Figure 3 shows that ambiguous collision occurs on Node R1 since Node R1 receives a new packet from the node S and receives overhearing packet from Node R2 at the same time. Therefore, appropriate time interval to transmit packets must be decided to avoid collision[9,10].

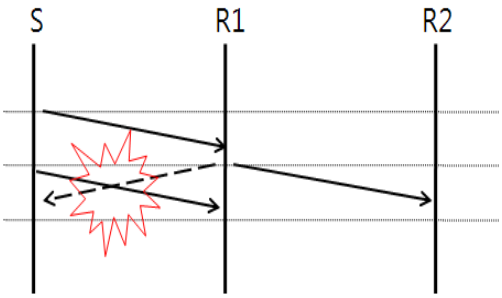


Figure 2. Packet collision by an overhearing packet

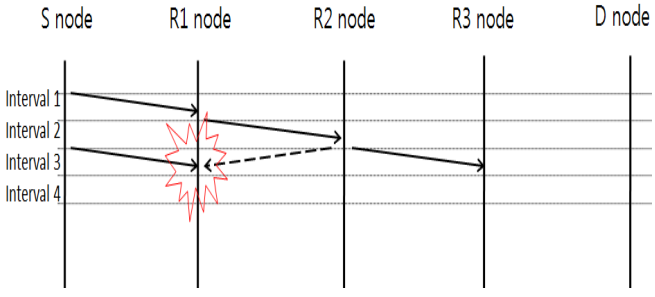


Figure 3. Packet collision by ambiguous collision problem

3.2 Time interval modeling

To avoid collision and to make efficient and reliable transmission, it is required to decide transmission time interval. At Figure 4, Node S transmits a packet to Node R1 during time t_1. When Node R1 receives the packet, it processes for time α and retransmit the received packet to its next node R2 for time t_2 . And Node R2 transmits the packet to Node R3. To avoid collisions the total time interval is calculated as Equation (1). The transmission time is represented as $t_1 = t_2 = t_3 = t$ since all packets are the same size.

$$\begin{aligned} T_3 &= t_1 + \alpha + t_2 + \alpha + t_3 \\ &= t_1 + t_2 + t_3 + 2\alpha \\ &= 3t + 2\alpha \end{aligned} \qquad (1)$$

To use packet overhearing, it is required to transmit next packet after T_i time later. It means that there is no collision or transmission fail after T_i time later.
If Equation(1) is generalized for n nodes, it is represent as Equation (2).

$$T_n = nt + (n-1)\alpha \qquad (2)$$

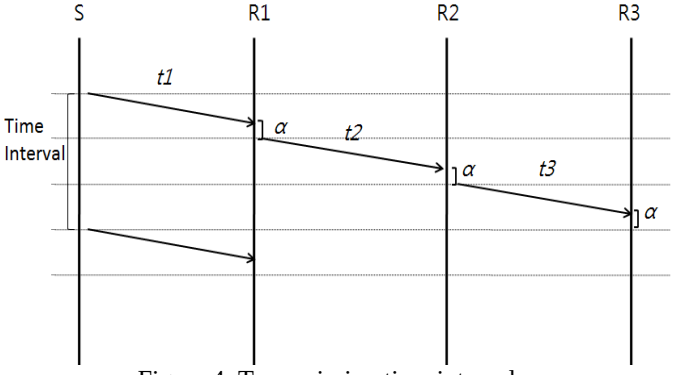


Figure 4. Transmission time interval

3.3 Proposed Algorithm

Without using control protocols, packet overhearing is used for the multimedia data transmission. By adjusting time interval of the packet transmission, the proposed algorithm is developed. The algorithm is shown in Figure 5. The algorithm is divided into two parts. One is the transmission part and the other is receiving part. In Figure 5, each node will work a source node or a destination node depending upon received packets. To avoid collisions, it is required that the timer must work accurately.

```
1.  Transmission
  if ( source node ) {
    if (packet = image message ) make packets from image
                                    or  video data;
    do until ( packet = the end of frame) {
      Start Timer;
      Transmit packets;
      Change the receiving mode;
    }
  }
  else if ( relay node ) {
    transmit received packets to the next node;
    change to receiving mode;
  }

2.  Reception
  receive packets;
  if ( not source node ) {
    if ( packet = image message) {
    find the node position if the message ;
      if( not destination node ) {
        increment message-relay-node count ;
        transmit the message to the next node;
      }
    if ( destination node ) {
      transmit Ack;
    }
    else {
      change transmission mode;
    }
  }
  else if ( source node ) {
    if ( packet overhearing ) {
      stop timer;
      calculate interval time ;
      change transmission mode;
    }
    else {
      retransmit packets;
    }
  }
```

Figure 5. Proposed algorithm

4 Simulation and Experiment

The proposed method is simulated using NS2 and its results are compared with the End-to-End transmission method. Also the algorithm is programmed on sensor network motes and measured its performance.

4.1.1 Simulation

To setup simulation environment, the topology is shown in Figure 6. Five nodes are placed equally and the total distance from the source node to the destination is 100*m*. This configuration is good to find any problems to transmit packets continuously. Also it is fair condition to compare simulation results and the real measurement results. Table 1 shows simulation parameters.

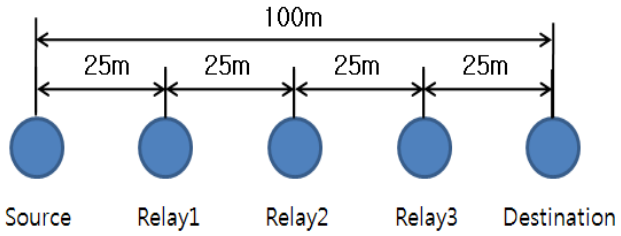


Figure 6. Network topology for simulation

Table 1. Simulation parameters

	End-to-End Transmission	Proposed algorithm
Routing protocols	AODV	
MAC and PHY	IEEE 802.15.4	
Packet size	127	
Transmission Time	5 sec	

CBR traffic is set to each transmission method and transmission time is 5seconds. The results of simulations are shown on Table 2. Transmission delay means that the interval time to transmit the next packet after a packet is transmitted. The transmission delay time of the proposed algorithm is the half of the End-to-End transmission. The number of transmitted packet is double of the End-to-End transmission. The reason is that the control signal, which is ACK, makes slow down packet transmission.

Table 2. Simulation results for 5seconds

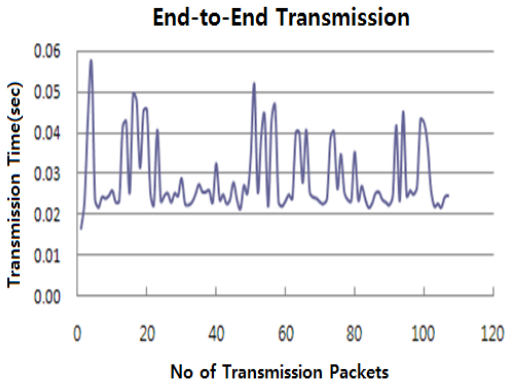
	End-to-End Transmission	Proposed Algorithm
No of Packets transmitted	52	104
No of failed packets	14	7
Average transmission time(*msec*)	28.08	24.2
Transmission delay(*msec*)	96.15	48

To apply to the multimedia sensor networks, JPEG image compression data is used for simulations. Three kinds of image size, which is 2KB(128 x 96), 8KB(320 X 240) and 16KB(640 x 480) , are used for simulations. The simulation results are shown in Table 3.

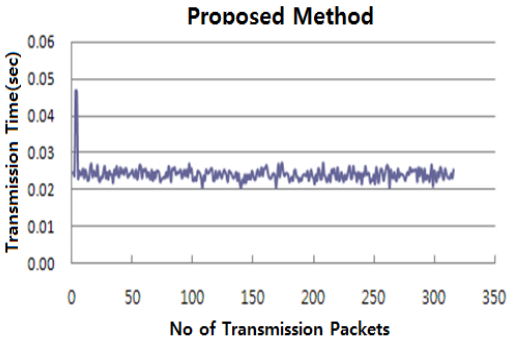
Table 3. Transmission performance for video images

	End-to-End Transmission	Proposed Algorithm
2KB (19)	1.824 *sec*	0.78 *sec*
8KB (75)	7.2 *sec*	3.1 *sec*
16KB(149)	14.304 *sec*	6.2 *sec*

The overall performance of the proposed algorithm shows double of that of End-to-End transmission. The transmission time is shown in Figure 7. By using overhearing packets and proper transmission delay time, the proposed algorithm improves transmission perpormance compared with the End-to-End transmission method.



(a) End-to-End transmission



(b) Proposed method

Figure 7. Transmission Time

4.1.2 Experiments

The proposed algorithm is programmed and implemented on a sensor node, which is USS-2400 from Huins Co. TinyOS is the operating system of the sensor node. The firmware is modified to run the proposed algorithm. The timer of the nodes is set to 1*msec* to measure transmission interval. Experiment topology is set to the same as Figure 6. 100 packets are used to test the algorithm

From Equation (2), α =3*msec* and transmission delay = 10*msec* are set to measure the transmission performance. As shown in Figure 8, average transmission delay time is 48*msec*, when 100 packets are transmitted from the source node to the destination node.

Figure 9 shows the success rate of the packets. To measure the success rate, 50 rounds of the 100 packets are transmitted and measured. Average of the success rate is 93.78 %.

From simulation results and real measurements, the proposed algorithm shows very good performance compared with the End-to-End transmission. The transmission performance of the proposed method is double that of the End-to-End transmission method.

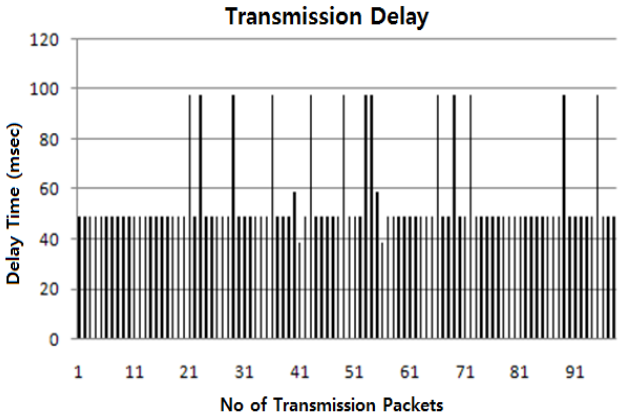


Figure 8. Transmission Delay Time

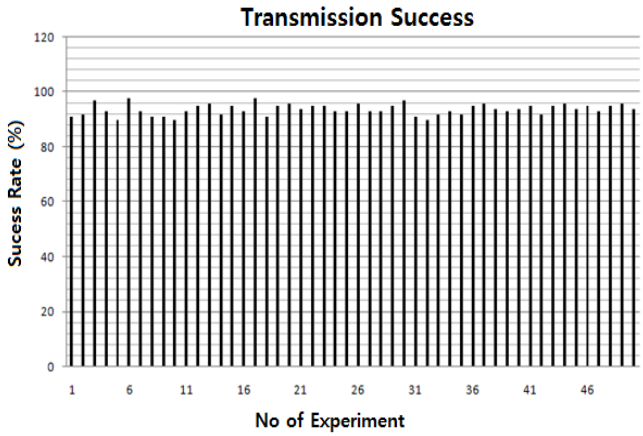


Figure 9. Packet Success Rate

5 Conclusion

To transmit multimedia data on WMSNs, it is required to use efficient protocols to reduce power consumption. This paper presents a new approach by using packet overhearing and adjusting packet transmission delay time. Simulations and real measurements are implemented to prove the performance of the proposed algorithm.

The results of real measurement are very close to those of simulations. The proposed algorithm shows very good performance compared to the End-to-End transmission. The transmission performance of the proposed method is double that of the End-to-End transmission. The experiment results show that average of the success rate is 93.78 %.

Please address any questions related to this paper to Byoungchul Agn by Email (b.ahn@yu.ac.kr).

6 References

[1] I. Almalkawi, M. Zapata, J. Al-Karaki, J. Morillo-Pozo, "Wireless multimedia Sensor Networks: Current Trends and Future Directions", Sensors, vol.10, pp.6662-6717, July 2010

[2] A. Sharif, V. Pordar, E. Chang, "Wireless Multimedia Sensor Network Technology: A Survey", Proc. of INDIN 2009, pp.606-613, June 2009

[3] R. Cucchiara, Multimedia surveillance systems. In Proceedings of the Third ACM International Workshop on Video Surveillance & Sensor Networks, VSSN'05, ACM: New York, NY, USA, 17-20, April 2005; pp. 3–10.

[4] I. Akyildiz, T. Melodia and K. Chowdhury, "Wireless Multimedia Sensor Networks: Applications and Testbeds", Proceedings of the IEEE, Vol.96, No.10, Oct. 2008

[5] L. Shu, Y. Zhang, Z. Zhou, M. Hauswirthh, Z. Yu and G. Hynes, "Transmitting and Gathering Streaming Data in Wireless Multimedia Sensor Networks Within Expected Network Lifetime", Mobile Networks and Applications, Vol.13, No.3-4, pp.306-322, August 2008

[6] A. Rowe, A, Goode, D. Goel and I. Nourbakhsh "CMUcam3:An open programmable embedded vision sensor", International Conferences on Intelligent Robots and Systems, May 2007.

[7] H-C. Le, H. Guyennet and V. Felea, "OBMAC: an Overhearing Based MAC Protocol for Wireless Sensor networks," The International Conference on Sensor Technologies and Applications (SENSORCOMM07), Valencia, Spain, October 2007.

[8] A. Kanzaki, Y. Iima, T. Hara and S. Nishio, "Overhearing-based Data Transmission Reduction Using Data Interpolation in Wireless Sensor Networks," The 5th Int'l conference on Mobile Computing and Ubiquitious Networking 2010, 2010.

[9] K. Winstein, H. Balakrichnan, "End-to-End Transmission Control by Modeling Uncertainty about the Network State", Proc. of the 10th ACM Workshop on Hot Topics in Networks, Nov. 2011.

[10] T. Roh and K. Chung, "A MAC Protocol for Efficient Burst Data Transmission in Multihop Wireless Sensor Networks", KIES: Information and Communication, Vol 35. No3, 2008.

[11] T. Chiang, J. Chang, S. Lin, "A Distributed Multicast Protocol with Location-Aware for Mobile Ad-Hoc Networks", Advances in Intelligent and Soft Computing, Vol.129, pp.691-697, Nov. 2012

[12] B. Wang, "An Individual Behavior-Based Trust Routing Model for Ad Hoc Networks", Proc. of MINES '09, Vol.2, pp.454-457, Nov. 2009.

On Smart Wireless Sensor Networks: Directions and Challenges

Dan Partynski and Simon G. M. Koo
Department of Mathematics and Computer Science
University of San Diego
San Diego, CA 92110
{dpartynski,koo}@sandiego.edu

Abstract

With the advancement of sensing technology and the increasing computational power of processors, there is an increased interest in using large, dynamically distributed wireless sensor networks (WSN) in a variety of areas. In recent years, WSN technology has proven to be beneficial in health and environmental monitoring, military applications, and many other fields. Since WSN can both sense the environment and apply algorithms to process the data, more real-time and useful information can be gathered from the physical world than using traditional sensing systems. This paper will explore the challenges faced by WSNs and how they can be addressed, from hardware and software aspects to the security of the systems.

1 Introduction

As we move toward the future of technology, computers will become increasingly prevalent in our daily lives. Advances in microprocessors and data processing will allow for the use of small, ubiquitous computing devices that may change the way we work with computers. A network of smart wireless sensors has the potential to be of great benefit across a large number of industries.

According to IEEE [11], a smart sensor node is a sensor "that provides a function beyond those necessary for generating a correct representation of a sensed or controlled quantity. This function typically simplifies the integration of the transducer into applications in a networked environment." A smart wireless sensor network thus consists of a large number of dynamically distributed smart sensor nodes. Due to the appeal of a truly dynamic system, the nodes must form an ad-hoc network robust to node failure and other changes in network topology, adding to the complexity of the underlying collective intelligence algorithms.

The main appeal of these networks, their ability to be cast into a dynamic environment for sensing purposes and require little to no human involvement, is the underlying cause of many significant research challenges. Hardware issues are especially challenging, as each sensor node needs sufficient energy and durability to perform its tasks. The underlying software and algorithms must be designed in such a way as to use the hardware to its full potential. There are also many security hurdles that must be overcome before smart wireless sensor networks can achieve widespread use. While there are many challenges to achieving this technology, the progression of sensor networks is an important goal. There are many potential uses of smart sensor networks, and there are likely many applications not yet imagined.

This paper will explore various aspects of Smart Wireless Sensor Networks. The standard architecture of a typical sensor network will be presented in Section 2. Section 3 will explore various research challenges of these networks. Section 4 will discuss various applications of sensor networks, and Section 5 concludes this paper.

2 System Architecture

Before going over research challenges and applications of smart sensor networks, we first give an overview of the architecture of both the individual sensing nodes and then the network as a whole.

2.1 Sensor Node Architecture

A typical node in a smart sensor network will consist of a few basic modules. These are the sensor module, the computation and communication module, and the power module [17]. The role of the sensor module is to gather different kinds of environmental data. Depending on the application, the nodes can be built with a variety of different sensors including temperature, acoustic, motion, moisture, magnetic, and humidity. Though it is possible that each individual node can include many of these different sensors on its own,

it is likely more cost effective to limit a node to perhaps one or two of these sensors. If an application calls for a large variety of sensors, then certain subsets of sensor nodes can be equipped with different sensor types, and a central intelligence will pool the disparate sensor data for processing. The computation and communication module will analyze sensed data and transmit information where needed. Each node should contain a sufficiently powerful microprocessor to permit the data and signal processing required for many applications. Perhaps the most crucial aspect of the architecture is the node's power module. Typically, the power source will consist of an alkaline battery, though this is not the most energy efficient solution and may lead to a large increase in node size. An overview of the challenges related to power sources for sensor nodes is presented in Section 3.

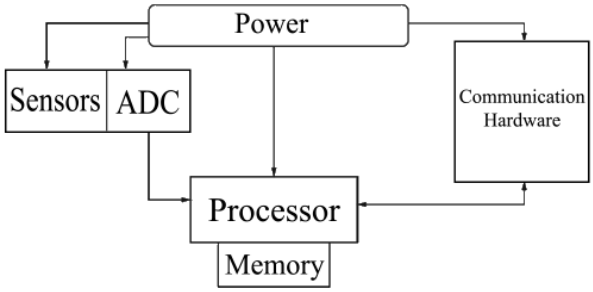


Figure 1. Basic components of a sensor node

2.2 Network Architecture

A smart wireless sensor network will consist of hundreds or thousands of these sensor nodes, dynamically distributed into an environment. It is unlikely that the network's topology can be predetermined, so the nodes will have to form an ad hoc network. It is ideal that the network be completely autonomous as to avoid any overhead and cost associated with direct human interaction. The crucial part of the network is the relaying of information to a central base station, which can be significantly far away from the nodes of the network. Having every node of the network transmit its sensed data directly to the base station works in theory, but due to the energy constraints of a single node, the combination of sensing data, processing it, and transmitting it to a base station may prove to be impractical. One possible solution is to have a central node distributed with the network, which would receive data from other nodes and relay the data to the base station. While this does solve a few energy related problems, having one central node to perform such an important task is undesirable as it leads to a single point of failure for the network. Thus networks will likely form node clusters, where a cluster is "a set of sensor nodes that surround the target phenomena and are capable of detecting and processing the data required by the users" [4]. Each cluster selects a node to be the cluster head, which will coordinate the actions of all the nodes within the cluster. Each cluster head will relay information to a central node, or directly to the base station if the central node has failed. The concept of node clustering leads to many benefits in the network architecture. They provide many points of failure rather than just one, and by coordinating behavior, they prevent redundancies among the sensor nodes within the cluster, thus preserving energy [2]. Clearly, network architecture is of critical importance to the success of a smart wireless sensor network, as a poor network design will lead to wasted energy and a short network lifetime.

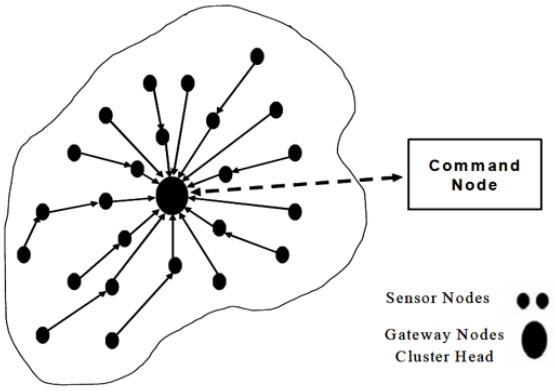


Figure 2. A typical node cluster

3 Challenges in Smart Wireless Sensor Networks

There are many interesting researches aimed at addressing challenges related to sensor networks. In this section, we will categorize such challenges into three different areas: security, hardware, and software, and discuss the challenges associated with each aspect .

3.1 Security

The nature of smart wireless sensor networks leads to several critical security issues that need to be addressed. Such issues include, but are not limited to, node compromise, unauthorized data access, and denial of service attacks.

3.1.1 Node Compromise

Since wireless sensor networks may consist of hundreds or thousands of smart nodes working together, each node is a potential point of attack. However, due to the dynamic dispersion of these nodes, it is impractical if not impossible to monitor each one of these nodes to protect them from an attacker [8].

A potential issue is the addition of a false node in the network, which would transmit corrupt information and attempt to falsify sensor data. This type of attack has been studied in other types of ad-hoc networking systems, but the countermeasures tend to be computationally expensive and too impractical for sensor networks given their energy constraints [18].

It is possible to make the individual sensing nodes resistant to physical tampering, so that it would be difficult for an attacker to alter a node‘s behavior if captured. However, this would likely be a cost ineffective solution for large networks, and does not alleviate all of the issues of node compromise. The end goal is a network that can both detect anomalies and function in the presence of a subset of corrupt or malicious nodes.

3.1.2 Unauthorized Data Access

Perhaps the most obvious security hurdle for a smart wireless sensor network is the prevention of access to network‘s transmitted information. There is a large amount of data generated by these networks, which can be easily viewed remotely by an attacker if no security measures are taken.

A standard way to combat this is to encrypt the information, but encryption algorithms may be expensive in a low energy environment. It is also worth noting that if a network wide encryption key was used for security purposes, it is possible that the compromise of a single node could allow for decryption of the entire network. Some potential encryption schemes are described in [14], though due to the complexity of data encryption and the limited availability of power, new schemes may need to be developed.

3.1.3 Denial of Service

Rather than compromise individual nodes, a malicious outsider could launch an attack on the entire network, thus rendering the network incapable of performing its task. These attacks may come in many forms, such as transmitting malicious signals into the network in order to interfere with routing protocols, or sending large amounts of useless information to sensors in order to waste their battery life. Some countermeasures to denial of service attacks are presented in [2] such as authentication to prevent unwanted signals from being processed by the network. Yet each countermeasure may in fact open up a number of other vulnerabilities, so much research needs to be done in order to construct a truly secure smart wireless sensor network.

3.2 Hardware

Wireless sensor networks are an ambitious concept, and there are many software issues that need to be taken care of before these networks can be widely used. However, there are just as many hardware challenges that will arise. These issues center around the difficulty of taking a sensing unit, a transceiver unit, a processing unit, and a power unit, condensing the entire system to the size of a cubic centimeter [3], and have the device be as energy efficient and cost effective as possible. Here we examine two major research challenges, preserving as much energy as possible and maximizing the node‘s processing ability.

3.2.1 Energy

Ideally, the nodes that make up a smart wireless sensor network would remain in a dynamic environment for long periods of time until the sensing job is done and all the required data has been collected and analyzed. However, it is inevitable that in the duration of the networks lifetime there will be small subset of the nodes that deplete their source of available energy. The underlying software of the network should be robust enough to handle these small changes in the networks topology, but the longer individual nodes can survive the better the collective data will be overall.

The question is what should be the energy source for each of the individual nodes of the network? Battery power is certainly an option, but the end goal is to have each node be as small as possible, and batteries would dominate the size of the structure. A better option is to use fuel cells, which are “rechargeable electrochemical energy-conversion devices where electricity and heat are produced as long as hydrogen is supplied to react with oxygen” [15]. Fuel cells would allow for good energy storage and power delivery, with the downside of a more complicated architecture. Perhaps the most interesting solution is to create energy scavenging devices, which gather acoustic, thermal, or solar energy for storage inside built-in capacitors. This design would allow for the initial goal of tiny sensing nodes that can survive on their own in a dynamic environment for extended periods of time with a minimized risk of node failure. Whether future networks will use these energy saving techniques or others, the task of powering sensing devices for long periods of time while keeping the size of the devices as small as possible remains a significant research challenge.

3.2.2 Processing

Due to the complexity of the sensed data across the network, each sensor node would ideally perform a certain amount of data processing before transmitting the information to another node or to the base station. Thus it is desirable that each node have a modestly fast processor. Due to the previously stated constraints of a sensing node, this may not be very easy to include in a cost effective manner, since each node should only cost around $1 in order for the entire network to be cost effective [3]. It is possible to rely on CMOS technology for each node's processor, but it may prove difficult to simultaneously achieve energy efficiency. Designing a processor that can work quickly inside of a tiny sensing node while simultaneously minimizing total energy usage remains an interesting and important research challenge.

3.3 Software

Coordinating thousands of sensor nodes in a dynamic environment with limited energy will prove to be a difficult challenge, so a robust software system and a new class of algorithms will be needed for the future of smart wireless sensor networks.

One energy related issue that can potentially be alleviated by software is the energy required for the transmission of data. Transmitting data to a base station and even to neighboring nodes can drain large amounts of a sensor node's energy supply, thus hindering its lifetime and the lifetime of the network as a whole. A possible solution is to take advantage of a node's processing ability and compress the sensed data before transmitting it to the other parts of the network. The idea is that the amount of energy needed to compress the data before sending it is far less expensive than the energy needed to transmit the full uncompressed data. While this idea has potential, existing data compression algorithms may still be too expensive to run in a single energy-constrained sensor node. Thus wireless sensor networks would benefit from a new, efficient data compression algorithm. Some sample algorithms for this task are presented in [12].

To reduce human interaction and to support the concept of an autonomous network, wireless sensor networks must be self-organizing. That is, when cast into an environment, the nodes must attempt to form clusters, assign heads to each of the clusters, and distribute tasks among the individual nodes so as to limit the amount of redundant behavior. For many applications, it may be beneficial for all or some of the nodes to know their location, but location identification techniques (such as GPS) tend to be computationally expensive. It may be possible to have only a small subset of node calculate their position, and design location algorithms to deduce the positions of the rest of the network nodes, thus limiting the energy used across the network.

A difficulty that can arise due to various instances of node failure is a coverage hole. A coverage hole is any region that is not sufficiently covered by a small number of sensors. The environment that the nodes have been deployed in ideally would be fully covered, but this clearly will not happen in the majority of cases. Closely related to a coverage hole is a routing hole, a region where either no nodes exist or are unable to transmit information across the network for various reasons. These challenges are inevitable in any network, and robust software solutions are required. A broader overview of the coverage problem and some potential solutions are presented in [1].

Perhaps the most difficult software challenges associated with these networks is software engineering, efficiently programming the nodes for a desired application. It is perhaps beneficial to break up the software development into different components in order to simplify the process. Three such components could be the sensor applications, the node applications, and the network applications [7]. The sensor application would have complete access to the operating system of each of the individual node and would be most closely linked to the hardware. The node application would be concerned with all high-level node tasks, such as data processing, data transmission, and location algorithms. The highest layer would be the network applications, which would interface with the administrator of the network.

4 Applications

This section will present a few of the potential applications of wireless sensor networks. There are many applications that these networks could be used for, and perhaps many that are not yet imagined, so here we examine a small application subset.

Because the sensors are equipped with various environmental sensing capabilities, wireless sensor networks are well-suited to environmental monitoring. For example, sensor nodes can be scattered across an environment and constantly collect information relating to temperature and humidity. In the future, the individual nodes may be small enough to literally loft about an environment by the wind, leading to even greater weather sensing opportunities. Various environmental events can also be detected with this kind of widespread continuous data gathering. The data acquired by these networks could help predict oncoming storms and earthquakes, and alert a base station of environmental fires by carefully analyzing sensed temperature data.

Wireless sensor networks have many military related applications. One possible application is to use sensor nodes as a replacement to mine fields. Sensor nodes scattered across a battlefield could detect acoustic and seismic activity to detect the presence of a hostile unit [5]. If the nodes determine that there is a threat, then they could relay in-

formation to a nearby actuator to handle the situation. Determining that a present unit is indeed an enemy personnel can be quite complex, and may rely on complex classification algorithms, which can of course be performed by the network. While this has been described as a wartime application, these concepts can be applied in peacetime for surveillance purposes.

The small size and processing capabilities of sensor nodes make these networks ideal for health related applications. Patients in hospitals can wear a large number of sensor nodes, which will be unobtrusive and carefully monitor and analyze the patients‘ physiological signs. The work can also be split up across many sensor nodes. For example, one subset of the nodes may be responsible for detecting heart rate, and another may only detect blood pressure. This will allow for detailed monitoring of a patient, and because the sensors can collectively analyze the data, they may be able to deduce context information (e.g. they can determine if an increase in heart rate is caused by exercise or a more serious health problem).

5 Conclusion

In the future, Smart Wireless Sensor Networks will become a standard informational tool throughout various industries due to their ability to gather and process data in new ways. As the technology advances, these networks will reduce in cost, and their use will be widespread across various applications. This paper has presented many of the challenges that must be addressed before the true potential of Wireless Sensor Networks can be fully realized. While issues in security, hardware, and software pose a great barrier to this technology, various research projects hope to address these issues in order to make wireless sensor networks a reality.

It is not entirely clear when these networks will become a fully functional, ubiquitous technology, but the progress in the direction of these networks is greatly important. The complex collective intelligence of smart wireless sensor networks will enable us to learn more about our world than ever before, and allow us to gather data that may be impossible to obtain from a traditional sensing system. It will take ingenuity and a large amount of dedicated research to achieve this futuristic technology, but due to the data acquisition potential of smart wireless sensor networks, it is well worth the effort.

References

[1] N. Ahmed, S. S. Kanhere, and S. Jha. The holes problem in wireless sensor networks: A survey. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(2):4–18, 2005.

[2] I. F. Akyildiz and I. H. Kasimoglu. Wireless sensor and actor networks: Research challenges. *Ad hoc networks*, 2(4):351–367, 2004.

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. Wireless sensor networks: A survey. *Computer Networks*, 38(4):393–422, 2002.

[4] A. Al-Ali, Y. Aji, H. Othman, and F. Fakhreddin. Wireless smart sensors networks overview. In *Second IFIP International Conference on Wireless and Optical Communications Networks (WOCN) 2005*, pages 536–540. IEEE, 2005.

[5] T. Arampatzis, J. Lygeros, and S. Manesis. A survey of applications of wireless sensors and wireless sensor networks. In *Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation*, pages 719–724. IEEE, 2005.

[6] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Wireless Communications*, 17(6):44–51, 2010.

[7] J. Blumenthal, M. Handy, F. Golatowski, M. Haase, and D. Timmermann. Wireless sensor networks-new challenges in software engineering. In *Proceedings of IEEE Conference on Emerging Technologies and Factory Automation (ETFA) 2003.*, volume 1, pages 551–556. IEEE, 2003.

[8] H. Chan and A. Perrig. Security and privacy in sensor networks. *Computer*, 36(10):103–105, 2003.

[9] D. Christin, A. Reinhardt, P. Mogre, and R. Steinmetz. Wireless sensor networks and the internet of things: Selected challenges. In *Proceedings of the 8th Fachgesprach Drahtlose Sensornetze*, pages 54–57. IEEE, 2009.

[10] M. Gigli and S. G. M. Koo. Internet of things: Service and applications categorization. *Advances in Internet of Things*, 1(2):27–31, July 2011.

[11] Institute of Electrical and Electronics Engineers (IEEE). *1451.2-1997 IEEE Standard for a Smart Transducer Interface for Sensors and Actuators – Transducer to Microprocessor Communication Protocols and TEDS Formats*. Piscataway, NJ 08855, Sep 1997.

[12] N. Kimura and S. Latifi. A survey on data compression in wireless sensor networks. In *International Conference on Information Technology: Coding and Computing (ITCC) 2005*, volume 2, pages 8–13. IEEE, 2005.

[13] L. Mainetti, L. Patrono, and A. Vilei. Evolution of wireless sensor networks towards the internet of things: a survey. In *Int. Conf. on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6. IEEE, 2011.

[14] A. Perrig, J. Stankovic, and D. Wagner. Security in wireless sensor networks. *Communications of the ACM*, 47(6):53–57, 2004.

[15] D. Puccinelli and M. Haenggi. Wireless sensor networks: applications and challenges of ubiquitous sensing. *Circuits and Systems Magazine, IEEE*, 5(3):19–31, 2005.

[16] R. Roman and J. Lopez. Integrating wireless sensor networks and the internet: a security analysis. *Internet Research*, 19(2):246–259, 2009.

[17] G. Song, Y. Zhou, Z. Wei, and A. Song. A smart node architecture for adding mobility to wireless sensor networks. *Sensors and Actuators A: Physical*, 147(1):216–221, 2008.

[18] J. Undercoffer, S. Avancha, A. Joshi, and J. Pinkston. Security for sensor networks. In *CADIP Research Symposium*, pages 1–11, 2002.

Clustered QoS Routing Protocol Based on Traffic Classification and Cost Expectation in Wireless Sensor Network

Haitao Wang[1], Jianzhou Li[1], and Lihua Song[2]
[1] College of Communications Engineering, PLA Univ. of Sci. & Tech., Nanjing, Jiangsu, China
[2] College of Command Information Systems, PLA Univ. of Sci. & Tech., Nanjing, Jiangsu, China

Abstract - *Quality of Service (QoS) is one of research hotspots in WSN, in which routing protocol plays a key role. On the basis of analyzing the existing QoS routing protocols, a clustered routing protocol based on traffic classification and cost expectation (CRPBCC) is proposed. In CRPBCC, packets are classified into ordinary packets and the important ones. Nodes form clusters and data fusion is carried out by cluster heads. Different data packets select the best route according to different cost expectations and queuing mechanisms. Besides, the congestion feedback mechanism is applied in CRPBCC protocol. Simulation results show that CRPBCC can effectively prolong the network lifetime, increase the packets delivery ratio and realize rapid and reliable transmission of important packets.*

Keywords: Wireless Sensor Networks; QoS Routing; Traffic Classification; Cost Expectation; Congestion Feedback

1 Introduction

WSN is composed of a large number of sensor nodes with limited computation and communication capabilities that cooperate with each other to form a network. It is widely used in fields like environmental monitoring, battlefield surveillance, earthquake relief etc. The energy, processing capacity, cache of sensor nodes in WSN are very limited, while the working environment of WSN is often poor; the network topology is easy to change, and the stability of the network link is difficult to guarantee, therefore QoS guarantee is an important and difficult problem in WSN [1].

In recent years, many QoS routing algorithms for wireless sensor network have been put forward. These algorithms either focus on cases with a single QoS parameter or consider multiple QoS indices, the latter of which could better satisfy the requirements of a specific application in WSN (such as delay and packet loss rate). SAR (Sequential Assignment Routing) protocol [2] adopts a tree structure which takes the one-hop neighbors of gateway nodes as roots and absorbs new nodes to join continually, and in the extension process of the tree nodes whose service quality is poor and the remaining power is insufficient are avoided, which guarantees the data transmission reliability, but the overhead of establishing and maintaining the routing table is large. SPEED protocol [3] is a real-time routing protocol, and to a certain extent realizes the end-to-end transmission rate guarantee, network congestion control and load balance mechanism. But this protocol does not consider the priority mechanism, unable to meet the real-time requirements furthest. EAQR protocol [4] selects the node load, average energy potential, and communication delay as QoS evaluation attributes, to normalize the data, and then calculates the weighted sum of them. QMR (QoS based multi-path routing) protocol [5] takes metric values like node residual energy, available buffer and channel quality into comprehensive consideration to choose transmission path, meanwhile, it adopts the scheduling strategy of data classification, to meet the real-time and non real-time data QoS demands. But when the two protocols are considering QoS indexes comprehensively, the QoS indexes are in linear superposition, which does not reflect the relationship between these indexes.

In proposed QoS-aware CRPBCC protocol the relation among QoS indexes including the delay, packet loss rate, energy consumption and reliability is considered and the data packets are divided into different types adopting different routing strategies and queuing mechanisms. In addition, the congestion feedback mechanism is applied to form the control loop.

2 Definitions and Descriptions

Packet delivery rate is defined as follows:

$$PDR = Nsr / Ns \quad (1)$$

Here, Ns is the number of data packets that nodes in wireless sensor network send, and Nsr is the number of data packets received by the base station in wireless sensor network. In the clustering routing protocol, the data packets that cluster head sends to the nodes in the cluster have higher degree of integration. Therefore, the delivery rate of WSN in cluster structure is defined as $PDR = Nsr / Nhs$, in which Nhs represents the number of data packets cluster head sends.

The importance of data in WSN is different, for example, the temperature, humidity and other data in the routine

This work is supported by NSFC (61072043)

monitoring are of low requirements for real-time property and reliability; meanwhile, the requirements of audio, video and command control information for real-time property and reliability are higher. And hence the packets are divided into ordinary packets and significant ones, according to the importance. We should pay attention to energy saving when transferring ordinary packets, while significant packets shall be timely and reliably transmitted.

When data packets are relayed between the nodes, the data packets in transmission may be lost, if the channel delay between nodes is small, but the packet loss rate is high; if we choose retransmission to improve reliability, time delay will increase. On the other hand, packet delay for reliable transmission may be short, if the channel delay is large, but the packet loss rate is very low. Therefore, when selecting the node for the next hop, delay and packet loss rate should be considered comprehensively. Similarly, energy consumption of the channel between nodes and packet loss should also be considered comprehensively.

If the packet is an important one, acknowledgment and retransmission have to be conducted during the relay between nodes [6]. Suppose that the packet loss rate of the channel between nodes is PLR, and then the transmission times until packet transmission succeeds are expected to be $Re = 1 / (1 - PLR)$. And the estimation of delay approximates to the product of expected transmission times and single transmission delay when the transmission of packets between nodes succeeds. The concept of estimation of delay is presented below.

Expectation of the delay Edt is defined as follows:

$$Edt = t\_delay * Re = t\_delay / (1 - PLR) \quad (2)$$

Here, t_delay is the transmission delay of data packets between nodes. The lower transmission delay, packet loss rate and estimation of delay are, the better the channel between nodes is. Therefore, estimation of delay could be the index for data packets to select next hop. Similarly, the estimation of energy consume approximates to the product of expected transmission times and the energy consumption of single transmission when the transmission of packets between nodes succeeds.

Expectation of energy consume Eec Edt is defined as follows:

$$Eec = e\_consum * Re = e\_consum / (1 - PLR) \quad (3)$$

Here, e_consum is energy consumption of single transmission between nodes. The lower energy consumption of the channel between nodes, packets loss rate and estimation of energy consume are, the better the channel is. And hence estimation of energy consume could be the index for data packets to select the next hop.

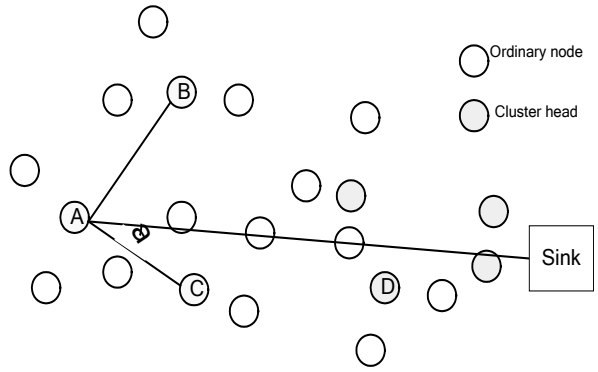


Fig.1 One example of routing angle

The efficiency of data packet transmission has something to do with routing angle [7]. Taking Fig. 1 as an example, cluster heads B and C are the nearest ones for the next hop of cluster head A, which have the same distance to A, and it is assumed that they share the same delay, energy consumption and packet loss rate, but the direction from A to C is closer to that from A to Sink or Base station (BS), which obviously make C a better choice than B. The probability of routing angle between nodes is given below.

Routing angle: If node C is the neighbor node of node A and the intersection angle between the lines that connects A to Sink and A to C is ‘a’, and then ‘a’ is the routing angle of C relative to A. As in the Figure 1 above, the included angle ‘a’ is the routing angle of C relative to A. From the perspective of the overall network, it would help reduce relay hops and communication conflict to choose the node with small angle to be the node for next hop.

3 Design of CRPBCC Protocol

CRPBCC protocol adopts the cluster network structure, in which the cluster head fuses the data packets sent by the nodes in cluster, and then based on the property of the data packets it selects the relay cluster head according to estimation of delay, estimation of energy consume or routing angle respectively. Relay cluster head adopts different queuing strategies of FIFO and LIFO for ordinary packets and important ones respectively. If there is congestion in the neighbor cluster head, we should choose suboptimal neighbor cluster head for relay. The data packet was finally relayed to the Sink. The workflow of CRPBCC is shown in Fig.2.

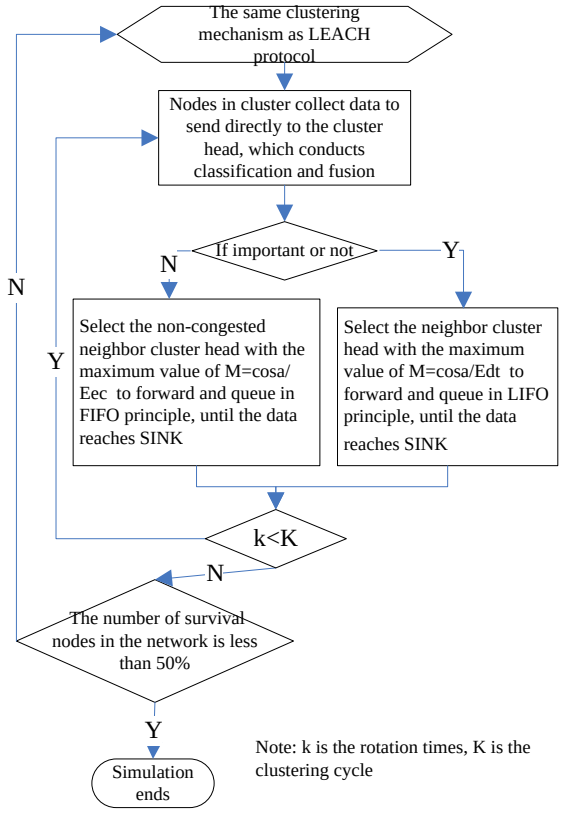


Fig.2 The workflow of CRPBCC

3.1 Clustering and Data Fusion

The clustering structure of WSN could achieve local autonomy, which reduces the amount of data transmission and interference between nodes, with management more convenient and the complexity of the protocol reduced. At the same time, since the data monitored by adjacent nodes has a strong correlation with redundancy in the data, through the data fusion in the cluster head, we can reduce the transmission of redundant data considerably, and improve the energy utilization efficiency. CRPBCC protocol adopts the same clustering method as LEACH protocol [8]. After the formation of clusters, the cluster nodes collect data, and divide it into ordinary packets and important ones according to the importance of data content. If the packet is an ordinary one, the cluster nodes wait for their own transmission slots to send packets to the cluster head; if the packet is an important one, then they can occupy the transmission time slot. Data packets in the cluster head shall get fused and then forwarded.

3.2 Routing Selection

For ordinary packets, we should select the neighbor cluster head with small route angle and estimation of energy as a relay node in the neighbor cluster heads that send no congestion alarm information, in order to improve energy utilization rate. Therefore, we take $M = cosa/ Eec$ as the selection criteria for the next hop, in which $cosa$ is the cosine of the cluster head's routing angle relative to its neighbor, and E_{ec} is its neighbor cluster head's estimation of energy consume relative to itself. When a neighbor cluster with the maximum M value sends out congestion alarm (see below), then the node with the second largest M value shall be selected, and so forth.

When a node is to transmit an important packet, the neighbor cluster head with small route angle and estimation of delay should be selected as a relay node, in order to reduce the average time consumption when the transmission is successful. Therefore, $M = cosa/ Edt$ is selected as the criteria with the same meaning of $cosa$ as above and E_{dt} being its neighbor cluster head's estimation of delay relative to itself, to choose the neighbor cluster head with the maximum M value to forward.

3.3 Queuing Mechanism

As usual, the node will put the received data packets in the buffer queue. If the received packet is an ordinary one, it will be stored at the position the queue tail pointer points, and the tail pointer moves forward. If the received packet is an important one, then a preemption mechanism shall be adopted -- the head pointer moves backward, the packets stored at the queue head.

When the node's sending module and the channel are idle, we always choose the packet that head pointer points to send. So even if a node is in congestion state, it can also forward important packets timely. The whole process is shown as in Figure 3, in which p122 represents an ordinary packet, and P127 represents an important one.

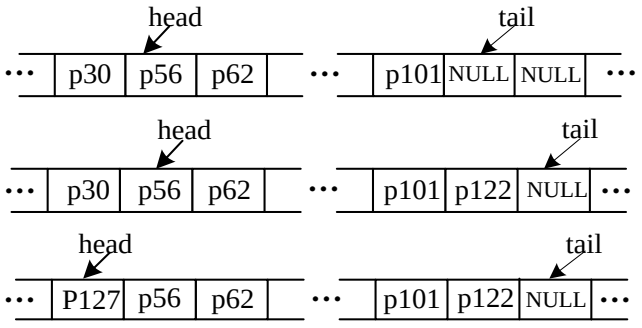


Fig.3 Queuing mechanism based on traffic classification

Besides, when the data packet at the queue head is important itself and another important packet comes up, the head pointer shall move backward, ensuring that the new packet gets relayed earlier. When all the important packets' relay is completed, the queue head pointer continues to move forward. When the new pointed data packets have been transmitted, the queue head pointer moves on until the pointer reaches the data packets that haven't been transmitted, and

begins to deal with it. So it can be seen that the preemption mechanism will not impact on the queue.

The cache of WSN's nodes is limited, and when the cache is exhausted, the data packets that come early or ordinary ones shall be discarded prior to others.

3.4 Congestion Feedback

When the channel gets into congestion, the packet sending rate of the nodes will decrease, with cache queue increased. In this case, the packet sending delay increases, or even packet loss occurs due to buffer overflow. Therefore, CRPBCC protocol adopts congestion feedback mechanism. When the transmission queue is very long or the cache is about to run out, alarm signal shall be sent to the upstream node. The upstream node will not choose the node as the next hop after receiving the alarm. But when the transmission channels around the node return to normal and the cache queue is reduced to a certain degree, a recovery signal is sent, and the upstream node can continue to choose it as a relay node. Congestion feedback mechanism has the function of balancing energy and flow and reducing time delay. The mechanism only needs to check the buffer occupancy within nodes, so the cost is quite small [9].

4 Simulation Experiment and Result Analysis

4.1 The simulation environment

OMNet++ 4.0 is chosen as the simulation tool, and main simulation parameter configurations are shown in table 1. According to the relationship between QoS expectations and the number of nodes [10], with the edge effect considered, the number of nodes is set to 160. Nodes in the region obey random uniform distribution. All working nodes send data packets in the cycle of 60 seconds, wherein the probability of important packets is 20%. Nodes' transmission delay and packet loss rate are randomly generated with the transmission delay distributed randomly in interval of (1, 5) ms and packet loss rate distributed randomly in interval of (5%, 30%). Through simulation the newly clustered network's cycle is determined to be 60*18 seconds [11]. When the number of survival nodes is reduced to half the original number of nodes, the lifetime of the network is set to end.

With a typical QoS routing protocol SAR selected as the comparison object in experiment, and several QoS indexes like the number of packets received, the delivery rate, time delay and network lifetime of ordinary packers and important ones are compared. Experiment simulations are conducted for 20 times, to compare the average value.

Table 1 Simulation parameters

Parameter	Value
network area	(500m*500m)
location of Sink	(550m, 250m)
energy consumption of wireless transceiver circuit	50nJ/bit
energy consumption of wireless transceiver	100pJ/(bit*m2)
size of data frame	256Byte
size of broadcast frame	64Byte
energy consumption of data fusion	5nJ/bit
size of buffer	512K

4.2 Simulation results and analysis

Simulation results show that, compared with the SAR protocol CRPBCC can prolong the network lifetime by an average of 15%, and the death rate becomes slower, with more balanced energy consumption. The change of number of survival nodes with time by using two routing protocols is shown in figure 4. CRPBCC balances the energy consumption by the method of periodic clustering. It belongs to the distributed protocol, with the method of local optimization for routing employed, and retransmission times are reduced when delivering important packets, to achieve better performance of energy saving.

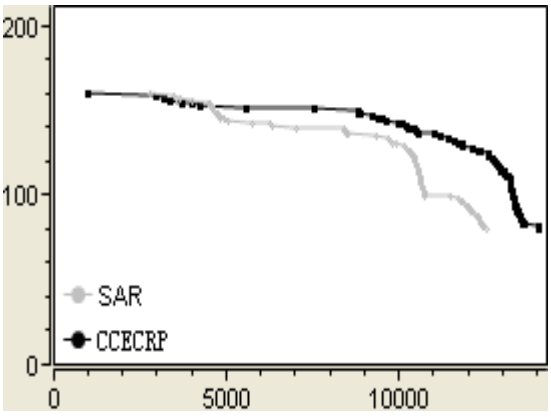


Fig.4 The change of number of survival nodes with time

As shown in Figure 5 (horizontal ordinate values 1 and 2 represents the number of ordinary packets and the important packets respectively), the number of ordinary packets and important ones increases by an average of 12.9% and 12.5% respectively, to receive more data with limited energy.

As shown in Figure 6, the delivery rate of ordinary packets and important packets improves by an average of

2.7% and 1.2% respectively, to reflect the effect with the packet loss considered.

As shown in Figure 7, the ordinary packets' transmission delay increases by an average of 14%, while that of important packets is reduced by an average of 9.5%. Ordinary packets' requirement for delay is relatively low, and the CRPBCC protocol mainly considers the energy saving and delivery rate, which means it will not necessarily choose the path with small delay for transmission; while for important packets, CRPBCC is capable of more rapid and reliable transmission.

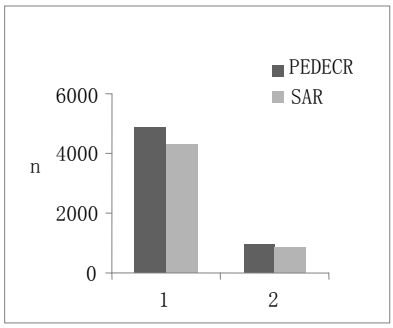


Fig.5 Average number of received packets by Sink

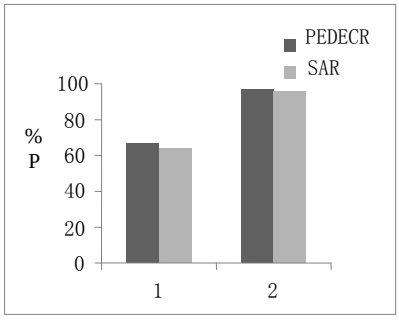


Fig.6 Comparison of average delivery rate

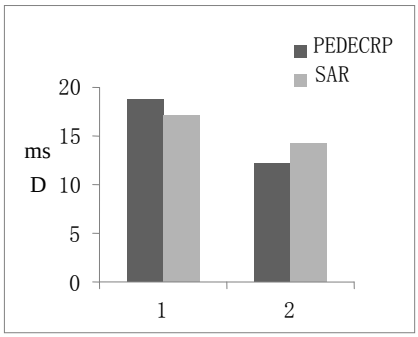


Fig.7 Comparison of average delay

5 Conclusions

The QoS routing protocol of CRPBCC is forwarded here, which clusters nodes, classifies packets, routes according to the price expectations, queues in accordance with category, and employs the congestion control mechanism. Simulation results show that, the protocol can prolong the network lifetime, receive more data, improve the packet delivery ratio, and realize fast and reliable transmission of important data, to achieve better QoS comprehensive performance. CRPBCC protocol applies to the occasions with different QoS requirements. On the basis of the CRPBCC protocol, how to reduce transmission delay of ordinary packets and combine with other QoS security strategies is the focus of the next step.

Acknowledgment

This paper is supported by National Natural Science Foundation of China (NO: 61072043).

References

[1] Haitao Wang. The clustering protocol review in Wireless Sensor Networks[J]. Sensor world, 2011,5(4):6-10.

[2] Sohrabi K，Pottie J．Protocols for self－organization of a wireless sensor network [J]. IEEE Personal Communications，2000,7(5):16-27．

[3] He T，Stankovic J A，Lu Chenyang，et al． SPEED: A stateless protocol for real-time in wireless sensor networks [J]. IEEE Transactions on Parallel and Distributed Systems, 2005,16(10): 995-1006．

[4] Weiren Shi, Mingmeng Yan, He Huang. Adaptive QoS routing algorithm in wireless sensor network based on entropy-weight coefficient method [J]. Computer application.2011,31(2): 298-300.

[5] Fangwu Yao, Chenhao Li. A wireless sensor network routing protocol based on QoS [J]. Computer technology and development, 2012,7(22):37-41.

[6] Lanchi Jiang, Ertao Li, Guoxuan Zhang. The research of wireless sensor network based on packet loss rate [J]. Journal of Hangzhou Dianzi University, 2009,8(29):42-45.

[7] Zhihen Xie, Xiangli Zhang, Long He, et al. Wireless sensor network routing scheme based on distance and angle [J].Computer engineering and applications, 2010,46(31):109-110.

[8] Heinzelman W, Chandrakasan A. An Application-specific protocol architecture for wireless micro-sensor network. IEEE Trans on Wireless Communication, 2002,1(4):660-670.

[9] W.Chieh-Yih, B.S.Eiseman, CODA: Congestion Detention and Avoidance in Sensor Networks, ACM SENSYS, Aug, 2003.

[10] Shuguang Deng, Lianfeng Shen, Xinhui Chen, etc. On-demand QoS protocol based on energy balance in large-scale wireless sensor networks [J]. Journal of circuits and systems, 2010,15(4):75-81.

[11] Jianzhou Li, Haitao Wang, An Tao. A balanced energy consumption of the WSN routing protocol [J]. Journal of transducer technology, 2013,26(3):396-401.

Delay-Efficient Data Collection with Dynamic Traffic Patterns in Wireless Sensor Networks

Phuc Nguyen Van, Vyacheslav Zalyubovskiy, and Hyunseung Choo*
College of Information and Communication Engineering
Sungkyunkwan University
Suwon, South Korea
nvphuc@skku.edu, slava@ece.skku.ac.kr, choo@skku.edu

Abstract— *Data collection is one of the most important applications in Wireless Sensor Networks (WSNs), where the data are gathered from sensor nodes to the base station. To reduce energy consumption, the sensor node may not report every sensed data sample to the base station. Thus, the network traffic of continuous data collection application often varies unpredictably over different sampling intervals. In this paper, we propose an energy-efficient scheme, Delay-Efficient Traffic Adaptive (DETA), for collecting data from sensor nodes with minimum delay according to the traffic load. The DETA scheme minimizes data collection delay by constructing a delay-efficient, collision-free schedule, and by using an adaptive mechanism to enable every node to self-adapt to the change of traffic. The simulation results show that our proposed solution could significantly decrease data collection delay and obtain reasonable values of energy consumption compared with other schemes.*

Keywords: scheduling; data collection; dynamic traffic, wireless sensor networks.

1. Introduction

Data collection from sensor nodes in a network over a tree based structure is a fundamental problem in WSNs. In many applications of WSNs, such as military surveillance [1], habitat monitoring [2], or structural maintenance [3], data collection is a key function in which the base station collects all data generated by sensor nodes in the network. Because sensor nodes are often powered by batteries that may not be recharged, reducing energy consumption has attracted great attention in recent years. Moreover, in many applications, it is crucial to guarantee the data collection time as quickly as possible. For instance, when the sensor nodes are used to detect gas, oil, or structural damage and so on, sensing data must be gathered as soon as possible. As a result, energy-efficiency and delay-efficiency are always important targets in WSNs.

The TDMA scheduling method is used quite commonly in WSNs. In a TDMA schedule, the collisions are eliminated by scheduling only non-interfering transmissions to proceed in the same transmission slot. There are many research [4]−[8] which provide the TDMA schedules for data collection in WSNs. In contrast to many realistic applications, most of these methods are designed for the static network traffic pattern where every sensor node has at least one data packet to send to the base station in a sampling interval. However, to reduce energy consumption, a sensor node may not report its data to the base station in every sampling interval. This causes the traffic network of data collection to vary unpredictably over different sampling intervals.

To the best of our knowledge, Wenbo Zhao *et al.* [9] are the first researchers to consider sensor data collection with dynamic traffic patterns. In this approach, the authors propose a TDMA scheduling algorithm which effectively deals with the change of network traffic. However, this scheme cannot guarantee good data collection delay because the parent node waits to receive data from all its children before sending its own data to its parent. In this paper, we propose a Delay-Efficient Traffic Adaptive (DETA) scheme to solve the data collection problem with dynamic traffic pattern in WSNs. The main contributions of this paper are summarized:

- Proposing an algorithm for scheduling sensor nodes to report data with minimum delay.
- Providing an adaptive mechanism to allow the sensor nodes to reduce their idle listening according to the change of network traffic.

The simulation results show that our proposed scheme can achieve up to 20% improvement in terms of data collection delay and keep energy consumption at reasonable values, compared with the existing scheme.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work. Section 3 defines our system model and assumptions. The proposed scheme is presented in Section 4. The performance evaluation is shown in Section 5. Finally, we conclude the paper in the last section.

2. Related Work and Motivation

Many research on sensor data collection in WSNs have been investigated in recent years. Data collection in WSNs

*Corresponding Author
ICWN'13, July 22–25, Las Vegas, Nevada, USA

consists of two types: (1) *Aggregate data collection* is to enable each internal node in the tree to aggregate all the data received from its children before forwarding them toward the base station. A node compresses the data from all its children and its own data into one packet, and then sends that data packet to its parent. Thus, only one transmission slot is assigned to a sensor node to send all data in its sub-tree [11]−[14]. (2) *Non-aggregate data collection* where the base station collects all data from sensor nodes individually, where a node has to send its data in multi-transmission slots. An internal node in the data collection tree needs more transmission slots than any of its children in the tree because it has to relay all the data received from those children to its parent [4]−[10].

In this paper, we consider the *non-aggregate data collection* problem in WSNs. Although there are many studies, Wenbo Zhao *et al.* in [9] are the first researchers to consider sensor data collection with dynamic traffic patterns. The authors propose a TDMA schedule algorithm, called TPO, to assign the collision-free transmission slots to all nodes to report their data to the base station. In the TPO schedule, a node can send all available data in its sub-tree at all subsequent transmission slots from the first one. For example, assuming that there are only two nodes which have data packets to send in a sub-tree of five sensor nodes, thus five transmission slots are required to be scheduled for the node to send data. In the data transmission process, two available data packets will be sent in the two earliest transmission slots assigned to that node. Data collection delay and energy consumed can be reduced by applying this salient feature. However, the TPO scheme cannot achieve the good performance of data collection delay. Since a parent node only can send its data after it has received data from all its children, unnecessary delays can occur in the data collection process.

3. System Model and Assumptions

In our approach, we model the sensor network by a unit-disk graph $G = (V, E)$, where V is the set of nodes including the sink S, and E is the set of links. An edge $(u, v) \in E$ if and only if node u is in the transmission range of node v. Due to limited transmission range, a routing infrastructure has to be constructed to transport data from sensor nodes to the base station. A common practice is to organize the sensor nodes into a tree structure rooted at the base station [4]−[8]. In this paper, we also assume that the routes from all sensor nodes to the base station have been formed by a given data collection tree T.

Using the same assumptions as in [4]−[10], we suppose that the data reported by each sensor in a sampling interval, if any, fit into one packet and the data packets created by different sensor nodes are not aggregated on the way to the base station. We divide the time into slots, and a node only can send or receive one data packet during one transmission slot. Our goal is to find a TDMA schedule strategy which can minimize data collection delay and energy consumption. As mentioned before, the network traffic frequently varies over different sampling intervals, thus the proposed schedule must effectively deal with these changes.

We eliminate collisions by scheduling only non-interfering transmissions to proceed in the same transmission slot. Note that the proposed algorithm is a scheduling strategy, thus it is independent of the interference models. To simplify the presentation, we only consider pair-wise conflict relationships in which a transmission from a node to its parent conflicts with the transmission of its siblings, parent and grandparent over the tree T, same as in [9].

4. Proposed Scheme

4.1 The Overall Approach

Our proposed scheme consists of two phases: the scheduling and the data transmission phase. In the scheduling phase, we assign transmission slots to all sensor nodes under the assumption that each of them has data to send. To achieve a minimum delay, we use a parallel strategy which enables each sensor node has the chance to send their own data as early as possible. In the data transmission phase, after the schedule information is obtained, each sensor node applies an adaptive traffic mechanism to reduce idle listening according to the real data distribution of current network traffic. Using this mechanism, the base station can conclude data collection earlier instead of listening until the end of schedule.

4.2 Delay-Efficient Data Collection Scheduling

In this section, we introduce a TDMA collision-free schedule algorithm, called a DETA schedule. We assume that each sensor node has one data packet to send to the base station and use an example shown in Fig. 1(a) to explain the algorithm. We define three states for each sensor node: $Wait$, $Ready$, and $Scheduled$. Initially, all the sensor nodes are in the $Wait$ state. A node is changed to $Ready$ state if it is a leaf node or a node whose all children have been scheduled. The DETA schedule assigns transmission slots to the nodes only if they are in $Ready$ state. After being assigned transmission slots, the state of that node changes to $Scheduled$.

Algorithm 1 presents the pseudo code of the DETA algorithm. There are some variables which will be frequently used in this presentation: T is a given data collection tree, $ReadySet$ is the set which contains all sensor nodes in $Ready$ state. We consider that sensor node v: $Ch(v)$ contains all descendants of v, $p(v)$ is a parent node of v, $CS(v)$ is a set of nodes that conflict with v when they send data in the same transmission slot with v, $ReqTS(v)$ is the number of required transmission slots of node v to send all data in the sub-tree rooted at node v, $TS(v)$

Algorithm 1 DETA-SCHEDULE

```
1: ReadySet ← φ
2: while T ≠ {s} do
3:     for each node v ∈ T do
4:         if v is a leaf node then
5:             State(v) ← Ready
6:             ReadySet ← ReadySet ∪{v}
7:         else
8:             State(v) ← Wait
9:         end if
10:    end for
11:    SCHEDULE-ROUND (T, ReadySet)
12:    if |TS(v)| = ReqTS(v) then
13:        T ← T \{v}
14:    end if
15: end while
```

Algorithm 2 SCHEDULE-ROUND(T,$ReadySet$)

```
1: for each node v in T do
2:     count(v) ← 0
3: end for
4: while ReadySet ≠ φ do
5:     for each node v ∈ ReadySet do
6:         SCHEDULE-NODE(v)
7:         State(v) ← Scheduled
8:         ReadySet ← ReadySet \{v}
9:         if |TS(v)| = ReqTS(v) then
10:            T ← T \{v}
11:        end if
12:        Inform all nodes in CS(v) about its state and its
   schedule information
13:    end for
14:    for i = 1 : |Ch(p(v))| do
15:        if State(Ch^i(p(v))) = Scheduled then
16:            count(p(v)) ← count(p(v)) + 1
17:        end if
18:    end for
19:    if count(p(v)) = |Ch(p(v))| then
20:        State(p(v)) ← Ready
21:        ReadySet = ReadySet ∪ {p(v)}
22:    end if
23: end while
```

contains all transmission slots assigned to node v. For any set of nodes V, $TS(V) = \cup_{v \in V} TS(v)$. DETA scheduling algorithm starts from leaf nodes and works in rounds. At the beginning of each round, all leaf nodes are in $Ready$ state; other nodes are in $Wait$ state. We then apply an algorithm called SCHEDULE-ROUND to assign transmission slots to all sensor nodes in the $ReadySet$ of T. Each schedule round ends when the SCHEDULE-ROUND algorithm is finished. If a sensor node is assigned the number of required transmission slots, it will be removed from the tree T.

The SCHEDULE-ROUND algorithm is presented by the pseudo code in Algorithm 2 where the goal is to schedule all sensor nodes in the $ReadySet$. First, it picks a node in the $ReadySet$, and assigns transmission slots for that node. The transmission slots are obtained by the algorithm, called SCHEDULE-NODE, which will be presented later. Then, the node changes its state to $Scheduled$, and informs its state and schedule information to its neighbors within two hops in the tree. Next, it is removed from the $ReadySet$. On receiving schedule information from its children, the parent node will check whether all its children are in $Scheduled$ state. If so, its state changes to $Ready$ and the node is put into $ReadySet$. The algorithm continues to assign the transmission slots to the $Ready$ nodes until the $ReadySet$ is empty. It means that all children of the sink are in $Scheduled$ state. As mentioned before, the schedule round also ends at this moment, the algorithm removes the nodes which have assigned the number of required transmission slots from the tree and starts the new round. In the same way, the next round repeates all the procedures with new tree structure.

All the transmission slots are obtained by the SCHEDULE-NODE algorithm, where its pseudo code is presented in Algorithm 3. There are three types of transmissions slots that can be assigned to the node. The first type is assigned to allow the leaf nodes to send their own data and the internal tree nodes to forward previously received data. There is at most one transmission slot assigned to a sensor node over a schedule round. The type two transmission slot enables the internal nodes to send their own data earlier; thus only one data packet of type two is assigned to one node in a whole algorithm. According to receive the data packets from some descendants early, the node is also responsible to forward those packets as soon as possible. The third type of transmission slot is used to do this duty. We define $T_1(v)$ as the type one set of transmission slots, $t_2(v)$ is the type two transmission slot, and $T_3(v)$ is the set of transmission slots of type three of node v. $T_r(v)$ is the set of transmission slots when a node v receives early data packets from its children. Initially, $T_1(v)$, $T_3(v)$, and $T_r(v)$ are empty, $t_2(v)$ is equal to zero. The three types of transmission slots can be obtained by the following rules:

Rule #1: "*Each schedule round is responsible to assign transmission slots to the sensor nodes to collect data from the leaf nodes*." A parent node receives data from all its children, and then forwards them one by one. Therefore, if the child has no data to send, it means that there is no remaining data in its sub-tree. In Fig. 1(a), since node B, L, I, O, N and V are the leaf nodes, they are assigned transmission slots 1, 1, 2, 1, and 1, respectively, to send their data. Then, the parent of these sensor node must be assigned transmission slots to forward data received from them, such as node U, it is assigned transmission slots 2 to forward data

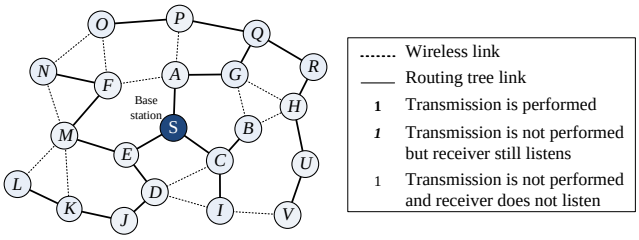
Algorithm 3 SCHEDULE-NODE(v)

1: // Assigning type one of transmission slot
2: $t_1(v) \leftarrow min\{r|r > 0, r > t_1(x), \forall x \in Ch(v), r \notin TS(CS(v))\}$
3: $T_1(v) \leftarrow T_1(v) \cup \{t_1(v)\}$
4: $TS(v) \leftarrow TS(v) \cup \{t_1(v)\}$
5: // Assigning type two of transmission slot
6: **if** $(t_2(v) = 0)$ and $(|TS(v)| < ReqTS(v))$ **then**
7: $Z \leftarrow [1, t_1(v) - 1]$
8: **if** $Z \neq \phi$ **then**
9: $t_2(v) \leftarrow min\{r|r \in Z, r \notin TS(CS(v))\}$
10: $TS(v) \leftarrow TS(v) \cup \{t_2(v)\}$
11: **end if**
12: **end if**
13: // Assigning type three of transmission slot
14: $Z = Z \backslash \{t_2(v)\}$
15: $T_r(v) \leftarrow \cup_{x \in Ch(v)}\{t_2(x), T_3(x)\}$
16: **while** $T_r(v) \neq \phi$ and $(|TS(v)| < ReqTS(v))$ **do**
17: $t_r^1(v) \leftarrow min\{t|t \in T_r(v)\}$
18: **if** $\exists z, z = min\{r|r \in Z, r > t_r^1(v), t_r^1(v) \in T_r(v), r \notin TS(CS(v))\}$ **then**
19: $T_3(v) \leftarrow T_3(v) \cup \{z\}$
20: $TS(v) \leftarrow TS(v) \cup \{z\}$
21: **end if**
22: $T_r(v) \leftarrow T_r(v) \backslash \{t_r^1(v)\}$
23: **end while**

from node V to node H. In general, the transmission slot assigned to a parent node is always greater than all of its children. Therefore, the type one transmission slot assigned to a node v, $T_1(v)$, is presented in lines 2 of Algorithm 3.

Rule #2: "*Each internal node in a current data collection tree can report its own data early such that no collision happens*." After assigning the type one transmission slot to forward data from the leaf node, the node continues to be assigned one time slot to send its own data. In Fig. 1(a), after being assigned time slot 4 as the first type, node R is assigned time slot 1 to send its own data. Similarly, after being assigned time slots $7, 4, 4$, nodes A, D, G then are assigned time slots $1, 1, 3$, respectively, to send their own data. In general, type two transmission slot of node v, $t_2(v)$, can be obtained as in lines 6−11 of Algorithm 3.

Rule #3: "*Upon receiving the children' own data early, a parent node is responsible to forward those data as early as possible*." If we consider each child node x of v, node x could send its own data early at $t_2(x)$, node v therefore has to be assigned a transmission slot to forward that data immediately. Secondly, if node x also has to forward data to some others nodes which can send their own data early, the data should be continuously forwarded by v as soon as possible. Therefore, we put all the early receiving transmission slots of v into list $T_r(v)$, thus $T_r(v) = \cup_{x \in Ch(v)}\{t_2(x), T_3(x)\}$, and sort the values in this set in ascending order. We then



(a)

Node ID	Sending TS	Node ID	Sending TS	Node ID	Sending TS
A	**1**,**4**,**7**,**10**,***13***,16,19,22,25	*H*	***3***,6,9	*O*	***1***
B	**1**	*I*	***2***	*P*	***2***,6
C	**3**,**6**,***11***	*J*	***3***,6,10	*Q*	**5**,**8**,***11***,14,17,20,23
D	**1**,***4***,8,12	*K*	***2***,5	*R*	**1**,***4***,7,10
E	**5**,**9**,***14***,15,17,18,20,21	*L*	***1***	*U*	***2***,5
F	***2***,6	*M*	***3***,7,11	*V*	***1***
G	**3**,**6**,**9**,***12***,15,18,21,24	*N*	***1***		

(b)

Node ID	Sending TS	Node ID	Sending TS	Node ID	Sending TS
A	**7**,**10**,**13**,**16**,***19***, 22,25,28,30	*H*	***3***,6,9	*O*	***1***
B	**1**	*I*	***2***	*P*	***2***,6
C	**3**,**4**,***6***	*J*	***3***,6,10	*Q*	**5**,**8**,***11***,14,17,20,23,26
D	**4**,***8***,11,13	*K*	***2***,5	*R*	**4**,***7***,10,13
E	**5**,**9**,***14***,15,17,18,20,21	*L*	***1***	*U*	***2***,5
F	***2***,6	*M*	***3***,7,10	*V*	***1***
G	**6**,**9**,**12**,***15***,18,21,24,27,29	*N*	***1***		

(c)

Fig. 1: An example execution of scheduling algorithms: (a) Network topology, (b) DETA, and (c) TPO [9]

assign the transmission slots to forward all early data packets received by those transmission slots. As in lines 14−23, type three transmission slots of a node v are obtained. For example, node G in Fig. 1(a) can send its own data at $t_2(G) = 3$. Therefore, node A is a parent of G responsible to forward that data early, and time slot 4 is assigned to forward that data, $T_3(A) = \{4\}$.

4.3 The Adaptive Network Traffic Mechanism

In this section, we describe an adaptive mechanism that enables sensor nodes to detect the completion of data transmission in their sub-tree. Note that in the same round of the DETA schedule, the first type time slot which is assigned to the parent is always greater than of all its children, which is represented in (1). Moreover, from (2)−(3), the first type of transmission slot that is assigned to a node is greater than the two remaining types in each schedule round. Therefore, if the parent node v does not receive any data from node x at any $t_1(x) \in T_1(x)$, and $t_1(x)$ is greater than the maximum value in $T_3(x)$, node v can go into sleep mode in all subsequent receiving transmission slots in $T_1(x)$. The base station, instead of listening until the end of schedule, concludes data collection once it infers that

all its children have finished transmission, thereby reducing collecting delay.

For example, suppose that only nodes A, B, C, D, E, G, Q and R generate packets to send. The actual transmissions occur as in Fig. 1(b) and Fig. 1(c) for the DETA and the TPO scheme, respectively. It is clear that our approach can significantly reduce the number of transmission slots for data collection according to the change of traffic. The DETA scheme spends only 10 transmission slots to collect all available data instead of the 16 transmission slots spent by TPO. As a result, base station S concludes that the data collection process has completed at the end of time slot 14 when no data comes from E. However, if we apply the TPO scheme, base station S can only conclude data collection until the end of time slot 19 when there is no data coming from node A.

5. Performance Evaluation

5.1 Simulation Environment

We implement TPO [9] to compare its performance with the DETA in terms of delay and energy consumption using MATLAB. We randomly distribute 100 nodes in a region of 100m x 100m. Each sensor node in the network has the same transmission range of 15m. A Breadth-First Search tree rooted at the base station has been constructed. In addition, to implement the change of network traffic in each sampling interval, we first generate 100 sets of nodes whose size varies from 1 to 100. Each set contains a random number representing the IDs of nodes that do not have data to send to the base station at that sampling interval. Each simulation runs through 200 sampling intervals. The energy costs for a sensor node to perform a transmission, and listen for transmission in a time slot are set 1 and 0.75 energy units respectively, the same as in [15]. Each node calculates the energy consumed for transmitting, receiving, and idle listening over 200 sampling intervals. Then, we add those amounts to get the total energy consumed for data collection.

5.2 Simulation Results

5.2.1 The Impact of Traffic Patterns

In this experiment, for each set of parameter settings, the reported results are the average of 200 runs on random network topologies. Fig. 2 and Fig. 5 show the data collection delay and energy consumption of the TPO and the DETA when network traffic varies from 100% nodes to no node that has data to send in a sampling interval. The collection delay of the DETA is always smaller than the TPO scheme. The DETA achieves up to 20% improvement in terms of data collection delay when 60% $-$70% nodes in the network have data to send. A disadvantage of the DETA is that we spend a little more energy for idle listening when some nodes do not have their own data to send at earlier transmission slots. However, we do not lose the energy in all the earlier transmission slots because the earlier transmission slot of the parent node also can be used to forward data from its children, *e.g.*, if a child node is assigned time slot 1 to send its own data, then a parent node will be assigned time slot 2 to send its own data; thus, if a parent node does not have data to send, the time slot of the parent node will be used to forward data from its children. In Fig. 5, we can see that the effect of that disadvantage is not too big, we only spend at most 2.8% more energy.

5.2.2 The Impact of Network Density

To investigate the impact of network size, we increase the number of nodes to vary from 100 to 600 in the same 100m x 100m region, each sensor also has a $15m$ transmission range. We evaluate the proposed algorithm with both: full and dynamic traffic patterns. In full traffic networks, each sensor node has one data to send to the base station in a sampling interval. Fig.3 illustrates the improvement of our scheme compared with the TPO scheme in terms of data collection delay. Our proposed algorithm can reduce data collection delay significantly. The improvement varies from 6.4% to 17.2%. In particular, a salient feature of the DETA is that it consumes the same energy consumption with the TPO in the case of full traffic. This is because they use the same amount of total transmission slots to collect all data, as in Fig. 6. Secondly, we assume that only 50 % nodes in the network have data to report to the base station. The impact of increasing the number of nodes in the fixed sensing area are presented in Fig. 4 and Fig. 7. Note that our scheme can achieve 6.7% to 18.5% improvement compared with the TPO scheme in terms of data collection delay as in Fig. 4. As seen in Fig. 7, the effect of the disadvantage of the DETA algorithm regarding energy consumption is very small; the DETA scheme only spends at most 2.7% more energy.

6. Conclusion

In this paper, we have presented a TDMA scheduling algorithm to schedule all nodes in the network to send their data to the base station with minimum delay. In addition, we designed an adaptive mechanism to enable sensor node to go to sleep early according to the current data distribution in a sampling interval, thereby reducing data collection delay and energy consumption. The simulation results show that our proposed scheme achieves better performance than the existing schemes in terms of data collection delay.

Acknowledgment

This research was supported in part by MSIP (NIPA,KEIT) and MEST(NRF), Korean government, under ITRC (NIPA-2013-(H0301-13-3001)), IT R&D Program [10041244, SmartTV 2.0 Software Platform], and PRCP (2012-0005861), respectively.

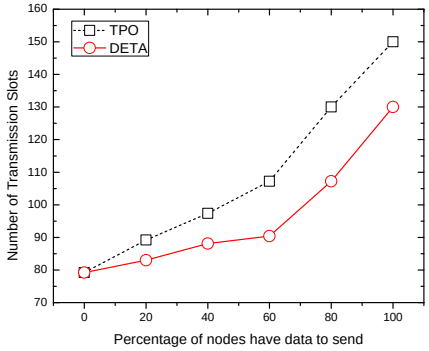


Fig. 2: Data collection delay: varying traffic patterns

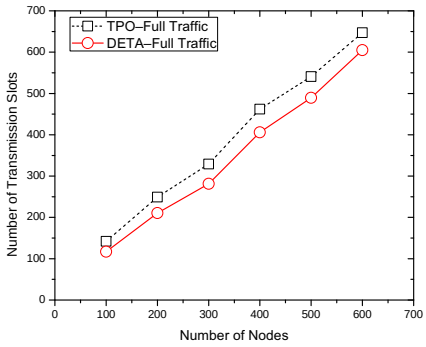


Fig. 3: Data collection delay: varying number of nodes when all nodes have data to send

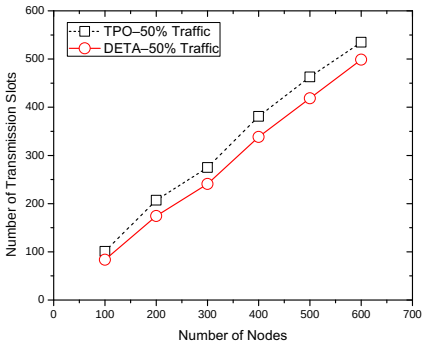


Fig. 4: Data collection delay: varying number of nodes when 50% of nodes have data to send

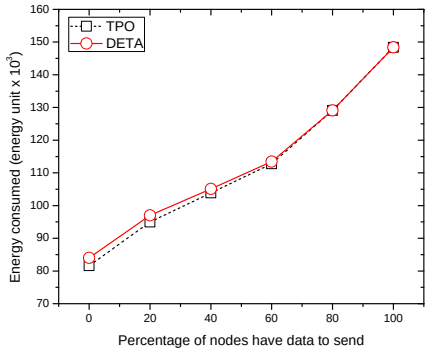


Fig. 5: Energy consumption: varying traffic patterns

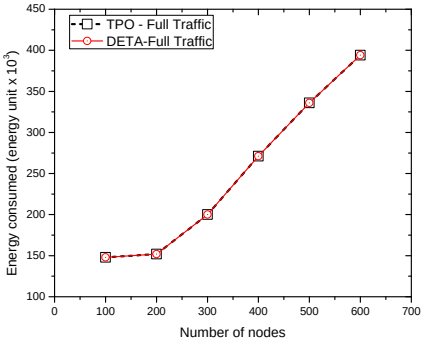


Fig. 6: Energy consumption: varying number of nodes when all nodes have data to send

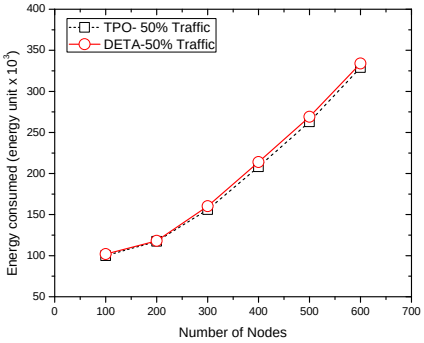


Fig. 7: Energy consumption: varying number of nodes when 50% of nodes have data to send

References

[1] I.F. Akyildiz, W.Su. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, pp. 102−114, 2002.

[2] R. Szewczyk, A. Mainwaring, J. Anderson, and D. Culler, "An Analysis of a Large Scale Habitat Monitoring Application," *In Proceedings of Sensys*, pp. 214−226, 2004.

[3] N. Xu, S. Rangwala, K. Chintalapudi, and D. Ganesan, "A Wireless Sensor Network for Structural Monitoring," *In Proceedings of Sensys*, pp. 13−24, 2004.

[4] S. Gandham, M. Dawande, and R. Prakash, "Link Scheduling in Sensor Networks: Distributed Edge Coloring Revisited," *Proc. IEEE INFOCOM*, pp. 2492−2501, 2005.

[5] S. Gandham, Y. Zhang, and Q. Huang, "Distributed Minimal Time Convergecast Scheduling in Wireless Sensor Networks," *Proc. IEEE ICDCS*, 2006.

[6] W. Song, F. Yuan, and R. LaHusen, "Time-Optimum Packet Scheduling for Many-to-One Routing in Wireless Sensor Networks," *Proc. IEEE MASS*, pp. 81−90, 2006.

[7] Y. Zhang, S. Gandham, and Q. Huang, "Distributed Minimal Time Convergecast Scheduling for Small or Sparse Data Sources," *Proc. IEEE RTSS*, pp. 301−310, 2007.

[8] L. Paradis and Q. Han, "TIGRA: Timely Sensor Data Collection Using Distributed Graph Coloring," *Proc. IEEE PerCom*, 2008.

[9] W. Zhao and X. Tang, "Scheduling Sensor Data Collection with Dynamic Traffic Patterns," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, pp. 789−802, April 2013.

[10] W. Zhao and X. Tang, "Scheduling data collection with dynamic traffic patterns in wireless sensor network," *Proc. IEEE Infocom*, pp. 286−290, April 2011.

[11] B. Yu, J. Li, and Y. Li, "Distributed Data Aggregation Scheduling in Wireless Sensor Networks," *Proc. IEEE INFOCOM*, pp. 2159−2167, 2009.

[12] Y. Li, L. Guo, and S. Prasad, "An Energy-Efficient Distributed Algorithm for Minimum-Latency Aggregation Scheduling in Wireless Sensor Networks," *Proc. IEEE ICDCS*, 2010.

[13] A. Ghosh, O. Durmaz Incel, V.A. Kumar, and B. Krishnamachri, "Multi-Chanel Scheduling Algorithms for Fast Aggregated Convergecast in Sensor Networks," *Proc. IEEE Int'l Conf. Mobile Adhoc and Sensor Systems*, pp. 363−372, 2009.

[14] X. Chen, X. Hu, and J. Zhu, "Minimum Data Aggregation Time Problem in Wireless Sensor Networks," *Proc. Int'l Conf. Mobile Ad-Hoc and Sensor Networks*, pp. 133−142, 2007.

[15] J. Ma, W. Lou, Y. Wu, X. Li, and G. Chen, "Energy-Efficient TDMA Sleep Scheduling in Wireless Sensor Networks," *Proc. IEEE INFOCOM*, pp. 630638, April 2009.

Distributed Positioning and Tracking in Cluster-based Wireless Sensor Networks

Chin-Der Wann[1] **and Chih-Ying Lee**[2]
[1]Department of Computer and Communication Engineering
National Kaohsiung First University of Science and Technology, Kaohsiung 81164, Taiwan
[2]Macronix International Co. Ltd., Hsinchu 300, Taiwan

Abstract— *In this paper, distributed positioning and tracking schemes for cluster-based wireless sensor networks (WSNs) are presented. We focus on the cases with overlapping cluster regions, which are particularly useful in applications involving inter-cluster communications, time synchronization, and target tracking. In overlapping clustering, the duplicated correlated data between adjacent clusters may, however, impair the performance in information fusion. To tackle the problem, a geometric formation with regularly distributed overlapping clusters is used for situation analysis. Each cluster is composed of one cluster head (CH) and a number of normal sensors. Different operating modes are defined for the CH and normal sensors. To resolve the problem of correlated data, Cholesky decomposition is adopted to decorrelate the measurement noises. In addition, extended information filtering (EIF) is modified for distributed target estimation. The simulations and results show that the proposed distributed schemes with the designated mode settings attain good efficiency and accuracy in target tracking.*

Keywords: Sensor networks, clustering, distributed positioning, target tracking, information filter.

1. Introduction

Wireless sensor networks (WSNs) have attracted significant attention in both academia and industry in the past decade [1], [2], [3]. With the rapid development of manufacturing and wireless communication technologies, sensors with moderate power consumption have been widely used. Among the WSN applications, target positioning and tracking are considered important issues.

The data processing in the WSNs can be divided into two categories: centralized and distributed architectures. In the centralized architecture, there exists a data processing center. All measured data at sensors are passed to the center for centralized computing. On the other hand, no data processing center is used in the distributed architecture. Measured data at each sensor are processed by the sensor itself. The processed data, derived information and estimated results can then be exchanged and shared among sensors via inter-sensor communications.

In a large-scale WSN, long-distance transmission usually leads to severe power consumption among sensors. Data collision and channel competition can also be major concerns. Sensor clustering provides an effective approach for better data aggregation and power conservation. In a cluster-based WSN, some sensors can be elected as cluster heads (CHs). The WSN is then divided into several clusters, each equipped with one CH. Inside each cluster, normal sensors send the measured data back to their corresponding CH. The CH is responsible for processing the received data, sharing the information with adjacent CHs, and fusing the exchanged information.

In the cluster-based WSN, each CH is responsible for a spatial region. The formation of cluster regions can be either disjoint or overlapping. In the formation with disjoint cluster regions, there is no overlap between adjacent cluster regions [4], [5]. The disjoint regions can be utilized to avoid redundancy in processing of duplicated data. On the other hand, in the formation with overlapped cluster regions, sensors located in the overlapped areas may be used in assisting the inter-cluster communications. A sensor located in an overlapping cluster region will generally transmit data to all of the corresponding CHs.

In this paper, the cases with overlapping cluster regions will be considered. A regularly distributed cluster-based WSN with multiple sensor clusters is studied. To improve the region monitoring and target tracking in WSN, operating modes at the CHs and sensors are designed. Since the sensors located in the overlapping area may transmit data to multiple CHs, the problem of redundant data processing need to be resolved in the distributed structure. To tackle the problem, approach with Cholesky decomposition is applied. By transforming the measurement processes at the CHs, the data correlation between adjacent clusters may be eliminated. The extended information filter (EIF) is adapted for performing data decorrelation. By using an EIF to fuse the shared information from other adjacent CHs, the CH of each sensor cluster is capable of performing distributed target positioning and tracking.

The remainder of the paper is organized as follows. In Section 2, methods related to the cluster-based wireless sensor networking for target positioning are discussed. In Section 3, a regularly distributed WSN and the operating

modes at the CHs and sensors are presented. The processing of redundant data and the modified EIF approach to implementing distributed positioning are explained in Section 4. Simulation cases and their results are provided in Section 5. Finally, conclusions and remarks are given in Section 6.

2. Cluster-based WSN for Target Positioning

For the applications of target tracking in the cluster-based WSNs, cluster formation and inter-cluster communications are essential to the overall performance. We intend to tackle the problems of data processing and inter-cluster communications in mobile positioning and target tracking in cluster-based scenarios.

We assume that the network consists of two different categories of sensors. The first category consists of highly capable sensors, which will be designated as cluster heads. Cluster heads are responsible for data processing and inter-cluster communications. The second category is a collection of normal sensors, which will be utilized for providing measured data to their corresponding cluster heads.

To conduct wireless location in a cluster of sensors with a highly capable cluster head, the time difference of arrival (TDOA) positioning method is used. In the TDOA location method, a sensor needs to be selected as a reference sensor. Since the CHs are highly capable in computing, it becomes intuitive that a CH is designated as the reference sensor in the TDOA positioning method. Measurements obtained at normal sensors are then transmitted to the CH to form the TDOA location metrics. In order to implement the distributed architecture, each CH is equipped with an EIF. The EIF at each CH performs target tracking. The derived information and estimated results are then shared with the adjacent CHs to enhance the accuracy of positioning.

2.1 Cluster-based WSN Architecture

In a cluster-based WSN, the processing burden of the CH is considered larger than that of normal sensors. The selection of the CHs is essential, and the strategies can be categorized as static CHs and dynamic CHs [6], [7], [8]. In our work, the static-CH strategy is adopted, and the CHs are assumed pre-designated. The CHs have higher computing capability, and are equipped with sufficient power.

Each CH is generally responsible for a spatial area called cluster region. In the approaches with overlapping cluster regions, sensors located in an overlapping region are used for facilitating inter-cluster communications and seamless coverage of situation changes [9], [10]. Since sensors located in the overlapping region may transmit data to multiple CHs, the duplicated data may cause redundant data processing problems. However, the processing of redundant data in cluster-based WSNs has seen few studies in the literature.

2.2 Extended Information Filtering (EIF)

Information filtering is an associated filter structure of the Kalman filtering (KF), which is useful in solving target tracking problems [11], [12], [13], [14]. Since the measurement equation of the TDOA positioning method is nonlinear, extended information filtering is adopted in the work. In addition to its capability in target tracking, the EIF also has low computational complexity, making it a good candidate in constructing a distributed architecture.

In target tracking, the state equation of a non-maneuvering target can be expressed as

$$\mathbf{X}(k) = \mathbf{\Phi}(k-1)\mathbf{X}(k-1) + \mathbf{\Gamma}(k-1)\mathbf{W}(k-1), \quad (1)$$

where the terms are listed as follows.

$\mathbf{\Phi}(k)$: state transition matrix
$\mathbf{X}(k)$: state vector
$\mathbf{\Gamma}(k)$: process noise transition matrix
$\mathbf{W}(k)$:process noise $\mathbf{W}(k) \sim N(\mathbf{0}, \mathbf{Q})$.

For the TDOA positioning at each CH, the measurement equation is non-linear and can be expressed as

$$\mathbf{Z}_i(k) = \mathbf{h}_i(\mathbf{X}(k)) + \mathbf{V}_i(k), \quad i = 1, 2, \cdots, \quad (2)$$

where i is the index for the CH, $\mathbf{h}_i(\cdot)$ the nonlinear measurement function, $\mathbf{Z}_i(k)$ the measurement data, and $\mathbf{V}_i(k)$ the measurement noise with normal distribution $\mathbf{V}_i(k) \sim N(\mathbf{0}, \mathbf{R}_i(k))$.

The iterative procedure of the EIF can be expressed by the steps listed below.
Prediction of information state vector:

$$\hat{\mathbf{y}}(k \mid k-1) = \mathbf{L}(k \mid k-1)\hat{\mathbf{y}}(k-1 \mid k-1). \quad (3)$$

Prediction of information matrix:

$$\mathbf{Y}(k \mid k-1) = [\mathbf{\Phi}(k-1)\mathbf{Y}^{-1}(k-1 \mid k-1)\mathbf{\Phi}^T(k-1) + \mathbf{\Gamma}(k-1)\mathbf{Q}(k-1)\mathbf{\Gamma}^T(k-1)]^{-1}. \quad (4)$$

Information transition matrix:

$$\mathbf{L}(k \mid k-1) = \mathbf{Y}(k \mid k-1)\mathbf{\Phi}(k-1)\mathbf{Y}^{-1}(k-1 \mid k-1). \quad (5)$$

Correction of information state vector:

$$\hat{\mathbf{y}}(k \mid k) = \hat{\mathbf{y}}(k \mid k-1) + \mathbf{i}(k). \quad (6)$$

Correction of information matrix:

$$\mathbf{Y}(k \mid k) = \mathbf{Y}(k \mid k-1) + \mathbf{I}(k). \quad (7)$$

where $\mathbf{i}(k)$ is the information state contribution, and $\mathbf{I}(k)$ is the associated information matrix. When the measurement noise $\mathbf{V}_j(k)$ between different sensors are uncorrelated, $\mathbf{i}(k)$ and $\mathbf{I}(k)$ can be written as

$$\mathbf{i}(k) = \sum_{j=1}^{m} \mathbf{i}_j(k), \quad (8)$$

$$\mathbf{I}(k) = \sum_{j=1}^{m} \mathbf{I}_j(k). \quad (9)$$

The information state vector and the information matrix at each CH are expressed as follows.

$$\mathbf{i}_j(k) \triangleq \mathbf{H}_j^T(k)\mathbf{R}_j^{-1}(k)[\mathbf{Z}_j(k) - \mathbf{h}_j(\hat{\mathbf{X}}(k \mid k-1)) + \mathbf{H}_j(k)\hat{\mathbf{X}}(k \mid k-1)], \quad (10)$$

$$\mathbf{I}_j(k) \triangleq \mathbf{H}_j^T(k)\mathbf{R}_j^{-1}(k)\mathbf{H}_j(k). \quad (11)$$

From (8) and (9), it can be seen that if the measurement noise between the sensors are uncorrelated, the correction stage of the EIF, (i.e., (6) and (7)) can be rewritten as Correction of information state vector:

$$\hat{\mathbf{y}}(k \mid k) = \hat{\mathbf{y}}(k \mid k-1) + \sum_{j=1}^{m} \mathbf{i}_j(k). \quad (12)$$

Correction of information matrix:

$$\mathbf{Y}(k \mid k) = \mathbf{Y}(k \mid k-1) + \sum_{j=1}^{m} \mathbf{I}_j(k). \quad (13)$$

Each CH calculates the information state vector and information matrix related to its cluster, and exchanges the results with other adjacent CHs. After receiving all the exchanged information, each CH can fuse the information and its predicted results by simple addition operation.

3. Regularly Distributed WSN and Operation Modes

3.1 Regularly Distributed WSN

A distributed WSN with hexagonal formation of sensors is used in analysis of target tracking performance. Without loss of generality, each hexagon represents a cluster of sensors. The analysis on a regularly distributed WSN will be easier than that on randomly distributed WSNs. In practical applications, if the sensor locations can be arranged in advance, the hexagonal deployment of sensors may be an efficient arrangement. For a given number of sensors, the WSN with hexagonal distribution will have the largest coverage of areas.

Fig. 1 depicts the composition of four clusters. The WSN is composed of two different types of sensors: high-capability sensors and normal sensors. The high-capability sensors are assigned as cluster heads (CHs), responsible for data processing and inter-cluster communication. The number of normal sensors is generally much larger than that of CHs. Normal sensors are designated to provide measured data to the CHs. In the hexagonal formation, each cluster consists of six sensors and one CH. It can be seen in Fig. 1 that each normal sensor belongs to three different clusters. The redundant data processing problem arisen from this arrangement will be discussed in Section 4.

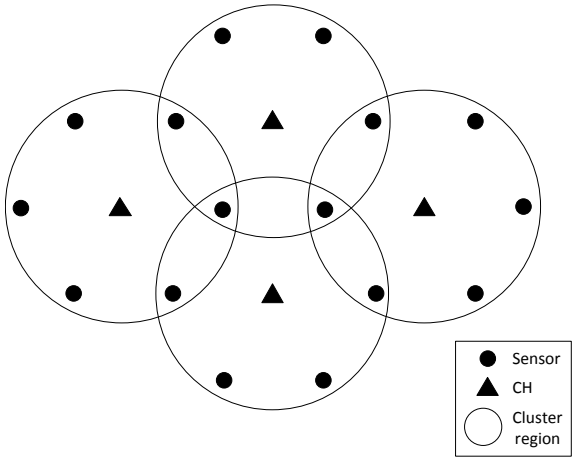


Fig. 1: Cluster composition

In the hexagonal distributed WSNs, the following terms are defined.

- *Sensing region* is the region in which CHs and sensors are designated to receive signals from a target. The sensing ranges of CHs and sensors are assumed the same. For short-range transmission, the radius of the sensing region is chosen to be 10 meters.
- *Communication region* is determined by the transmitting range of the sensors. All sensors communicate with their corresponding CHs, while each CH communicates with its adjacent CHs. The radius of communication region of sensors and CHs are set to 10 and 15 meters, respectively.
- *Cluster region* is related to the size of a cluster. Since the communication range of a sensor is set to 10 meters, the CHs can receive the signal from the sensors within the distance. The cluster region becomes a disk with radius of 10 meters.
- *Cluster operation region* is chosen as the same as the sensing region of the CH, which is a disk with 10-meter radius. Each CH is equipped with an EIF for target estimation.

3.2 Mode Settings for Sensor Operations

In the subsection, operating modes for both CHs and sensors are defined. Through the mode switching procedure, the WSN is expected to function more efficiently.

For the CHs in the network, three operating modes are defined.

1) Sleeping mode
   - Condition: (a) Initial setting, or (b) CH does not receive any information or measurement.
     - Action: Periodical scanning to check whether a target enters the sensing region.
2) Prepared mode (divided into two stages)

- Prepared-1: The CH receives the information from its adjacent CHs.
  - Action: Continuous scanning. The CH utilizes the received information to estimate the target position.
- Prepared-2: CH receives the information from adjacent CHs and the measurements from sensors in its own cluster.
  - Action: Continuous scanning. The CH utilizes the received information to estimate the target position. Also, the CH sends message to notify the sensors in its cluster to enter Active mode.

3) Active mode
   - Condition: A target is detected by the CH.
     - Action: Continuous scanning. The CH tracks the target by using the EIF, and broadcasts the information to all adjacent CHs.

For the normal sensors in the network, two operating modes are defined.

1) Sleeping mode
   - Condition: (1) Initial setting, or (2) The sensor receives a request from the CH (Prepared-2 $\rightarrow$ Prepared-1).
     - Action: No action.
2) Active mode
   - Condition: The sensor receives a request from the CH (Prepared-1 $\rightarrow$ Prepared-2).
     - Action: Continuous scanning. If a target is detected, the sensor sends the location measurements to the CHs.

In Fig. 2, an example showing seven clusters is used for illustration of the mode switching in a cluster. The arrow in Fig. 2 represents the trajectory of the mobile target, which moves from the right upper corner toward the left lower corner. The six cross symbols on the straight line represent six different target locations, implying different scenarios of mode operations.

From the view point of the central cluster, of which CH1 is the cluster head, the status and mode switching related to the sensors and CHs are described below.

(1) The target is far away from the operation region of the cluster. All the CHs and sensors are in Sleeping mode to conserve energy. CHs scan periodically, and sensors do nothing.

(2) The target moves into the operation region of an adjacent cluster (with CH2). CH1 receives the information from CH2 (the information is computed by the EIF at CH2). CH1 enters Prepared-1 mode, and uses the received information to predict the position of the target.

(3) The target moves into the sensing regions of some sensors in the cluster, but not into the the cluster operation region. CH1 receives measurements from the sensors, and enters Prepared-2 mode. CH1 still uses the received information to predict the position of the target, and also sends informing messages to all the sensors in the cluster. Sensors will enter the Active mode after receiving the message. In this scenario, sensors and CHs perform continuous scanning.

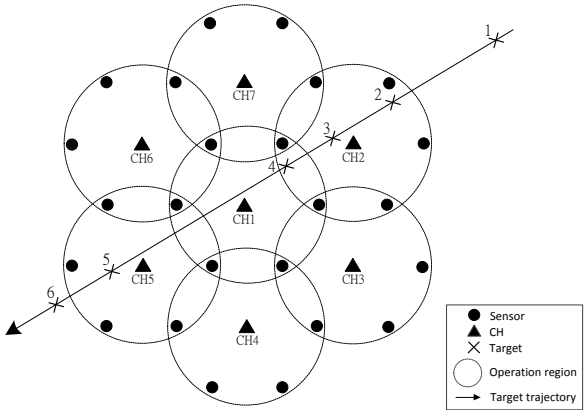


Fig. 2: Example: mobile target travels through a cluster-based WSN

(4) The target moves into the operation region of the cluster. With the continuous scanning setting of Prepared-2 mode, CH1 can detect the target immediately when it enters the cluster operation region. CH1 enters the Active mode and uses previously predicted state and the current measurements to track the target (by using the EIF). After obtaining the information of the target ($\mathbf{i}(k)$ and $\mathbf{I}(k)$ in the EIF), CH1 broadcasts the information to all adjacent CHs. If both of the CHs are in Active mode (e.g., CH1 and CH2 in Fig. 2), they will exchange the information. The positioning accuracy will therefore be enhanced.

(5) The target moves out of the operation region of the cluster. CH1 enters Prepared-2 mode. When the target moves out of the sensing regions of all the sensors in the cluster, CH1 enters Prepared-1 mode and sends messages to inform the sensors in the cluster to enter sleeping mode. Priority can be set at the the sensors: Active mode takes precedence over Sleeping mode. A part of sensors in the CH1 cluster receive active request from CH5 or CH6, and still work in Active mode.

(6) The target is far away from the operation region of the cluster. CH1 does not receive any information or measurement, thus enters Sleeping mode. The sensors sharing with the CH5 cluster are still in Active mode, while others return to Sleeping mode.

Different from the sensors, the CHs is designated with an extra operation–Prepared mode. In the Prepared mode, a CH is informed by other adjacent CHs of the possibly approaching target and its current location. For data pro-

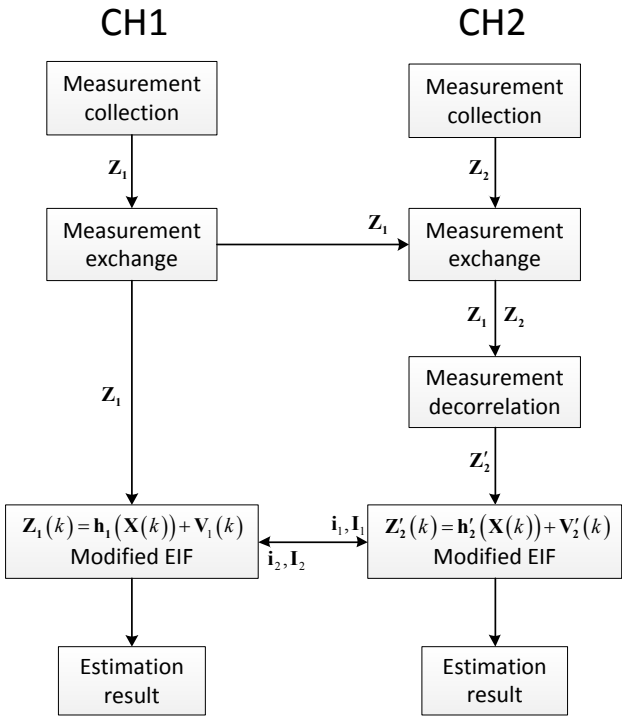


Fig. 3: Decorrelation between Two Clusters

cessing in the EIF, an initial value should be given in the beginning. If CH knows the position of the target in advance, the possibly large initial positioning error can be avoided. Accordingly, the Prepared mode can enhance the stability and the positioning accuracy of the system.

4. Processing of Redundant Data

In the cluster-based distributed WSN, a sensor may belongs to different clusters when overlapping regions are used. Different CHs may receive the same measurement from a sensor for target estimation, which leads to redundant data processing when information is integrated at the CHs. Due to the correlated measurement noises between adjacent CHs, the biased estimation results need to be corrected in the overlapping-cluster distributed approaches.

4.1 Decorrelation between Two Clusters

Suppose that the target is located in an operation region which belongs to two adjacent clusters. The sensors located in the overlapping region will transmit the TOA measurements to the corresponding CHs, which incurs correlation of the measurement noises between the two CHs.

The procedure of decorrelation between two clusters is illustrated in Fig. 3. The target state equation and the measurement equation at each CH are the same as (1) and (2). After collecting all the measurement in the cluster and calculating the TDOA measurement, a CH will broadcast the obtained TDOA measurement. Meanwhile, the CH will possibly also receive measurements from the adjacent CHs. The joint measurement equation can then be written as

$$\mathbf{Z}(k) = \mathbf{h}(\mathbf{X}(k)) + \mathbf{V}(k), \tag{14}$$

where

$$\begin{aligned} \mathbf{Z}(k) &= \begin{bmatrix} \mathbf{Z}_1^T(k) & \mathbf{Z}_2^T(k) \end{bmatrix}^T, \\ \mathbf{h}(\mathbf{X}(k)) &= \begin{bmatrix} \mathbf{h}_1^T(\mathbf{X}(k)) & \mathbf{h}_2^T(\mathbf{X}(k)) \end{bmatrix}^T, \\ \mathbf{V}(k) &= \begin{bmatrix} \mathbf{V}_1^T(k) & \mathbf{V}_2^T(k) \end{bmatrix}^T, \end{aligned} \tag{15}$$

and the measurement noise is of the distribution $\mathbf{V}(k) \sim N(\mathbf{0}, \mathbf{R}(k))$.

In real applications, the measurements which need to be decorrelated can be determined based on the chronological order of operation. As shown in Fig. 3, when the target moves towards a newly joining cluster, the newly joining CH is set as CH1, and the original CH as CH2. After the process of decorrelation, the information broadcasted from CH2 is corrected.

4.2 EIF for the CH in a Hexagonal Cluster

After the decorrelation process of the measurement noises, the extended information filtering (EIF) can be used in performing distributed target positioning. The dimension of the measurement noise covariance matrix at a CH is related to the number of TDOA measurements in the cluster. For example, if there are two TDOA measurements calculated at the CH, the dimension of the measurement noise covariance matrix will be 2×2. The number of TDOA measurements is equivalent to the active sensors in the cluster. In the hexagonal distributed WSN, the number is generally smaller than three in most cases. The least preferred situation, which involves all of the six normal sensors, would occur when a target travels through the close vicinity of the CH. In case that lower computing complexity is required, additional criteria can be applied in the algorithm to reduce the dimension of the matrices.

5. Simulation Results

Based on the hexagonal distributed WSN and the designated operating modes for CHs and sensors, some simulation results are provided in this section. The simulation scenario is shown in Fig. 4. The triangles represent the CHs, and the small circles represent the normal sensors. The dotted circle indicates the cluster operation region, and the straight line is the trajectory of the mobile target. The simulation parameters are set as follows:

– Distance between sensors: 800 cm
– Sensing region: 1000 cm
– Measurement noise: $v(k) \sim N(0, \sigma_v^2)$, $\sigma_v = 15$ cm
– Initial position of the target: $(900, 1160)$
– Velocity of the target: $(-80, -60)$, 1 m/sec
– Process noise: $\mathbf{W}(k) \sim N(\mathbf{0}, \sigma_w^2 \mathbf{I})$, $\sigma_w = 0.1$ cm
– Initial value of EIF of CH1: $(400, 693)$

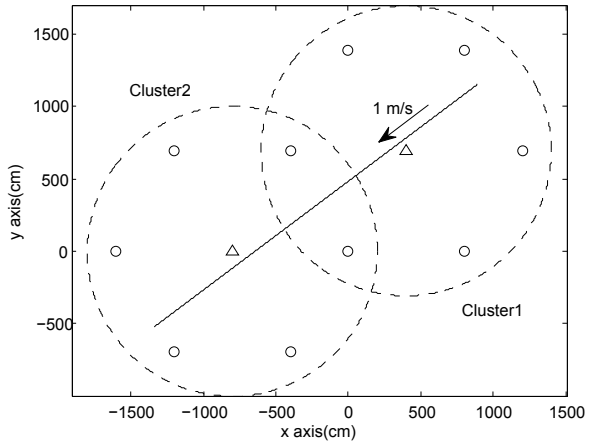


Fig. 4: Simulation case for the Prepared mode

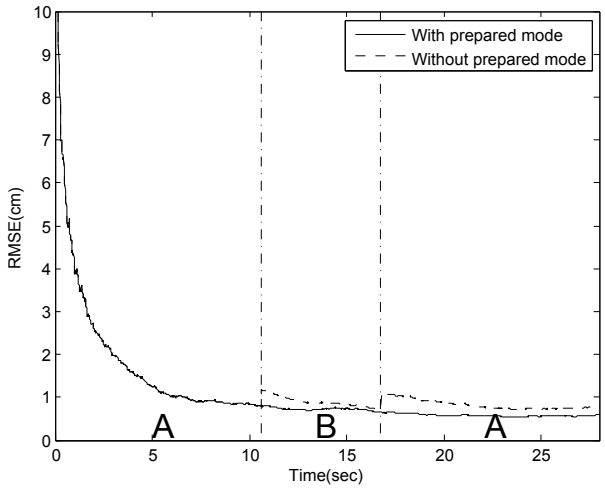


Fig. 5: Comparison of positioning errors between WSNs with and without Prepared mode

– Initial value of EIF of CH2: $(-800, 0)$
– Sample rate: 100 samples/sec

5.1 The Effects of the Prepared Mode

In Section 3, operating modes at the CHs and sensors are defined. As indicated in Section 3.2, the Prepared mode can enable a CH to acquire in advance the status and position of an approaching target. In this simulation case, the effects of the Prepared mode will be investigated and compared with the system without the Prepared mode.

The root mean square error (RMSE), obtained by using the Monte Carlo method, is calculated by

$$\hat{e}(k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i(k))^2}, \tag{16}$$

where $\hat{e}_i(k)$ is the error at the $k-$th step in the $i-$th simulation. The Monte Carlo simulation times is $n = 100$.

The estimation results of the active CHs are illustrated in Fig. 5. For a system without the Prepared mode, pre-selected initial values are used at the CHs. The capital letters **A** and **B** in Fig. 5 imply the number of operating clusters. Letter **A** means that the target is located in the operation region of a cluster, while letter **B** indicates that the target is located in an operation region covered by two clusters. The curves show the RMSE at CH1 for the first 17 seconds, followed by the RMSE at CH2 from the 17th second and on. The joint measurement noise covariance matrix turns out to be a diagonal matrix after the process, indicating the elimination of the correlation between any two adjacent clusters. Though the dotted curve in the first **A** interval is almost the same as the solid curve. In the **B** interval and the second **A** interval, the process with Prepared mode outperforms that without the designated mode.

It is noticed that there are two raises for the system without Prepared mode at the 11th second and 17th second, respectively. The first RMSE raise (at the CH1) is affected by the erroneous initial information sent from CH2. The second RMSE raise (at the CH2) indicates the estimation status, which has not yet achieved convergence since no beforehand target information is available in the case. It can be seen that due to the availability of approaching target information for CH2 in advance during the Prepared mode, the network with the Prepared mode operation provides better stability and better positioning accuracy than the one without Prepared mode.

5.2 Measurement Noise Decomposition

The measurement decorrelation is also investigated by using the scenario illustrated in Fig. 6. The solid circle represents the sensors located in the overlapping area of the operation regions, the remaining symbols are the same as described in Section 5.1. The number of clusters involving the operation region during the course of moving target changes in the following sequence: $1 \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 2 \rightarrow 1$. The simulation parameters are set as follows:
– Normal measurement noise: $v(k) \sim N(0, 15^2)$
– Biased measurement noise: $v'(k) \sim N(0, 45^2)$
– Initial position of the target: $(890, 1300)$
– Velocity of the target: $(-70, -70)$, about 1 m/s
– Initial value for the EIF: $(500, 1200)$
The settings of distance between sensors, sensing region, process noise, sample rate, simulation times of Monte Carlo method are the same as those in Section 5.1.

We assume that the measurements at all sensors are biased. The simulation contains the scenarios where the target is in the operation region of single cluster, two clusters, and three clusters, indicated by letters **A**, **B** and **C**, respectively. The redundant data processing problem occurs in the intervals **B** and **C**. Simulation results are shown in Fig. 7. It is seen that the process of decorrelation successfully reduces the RMSE caused by biased measurement problem.

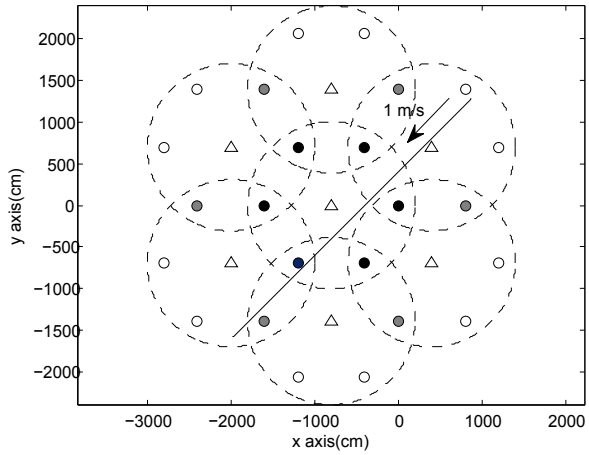


Fig. 6: Example for measurement noise decorrelation

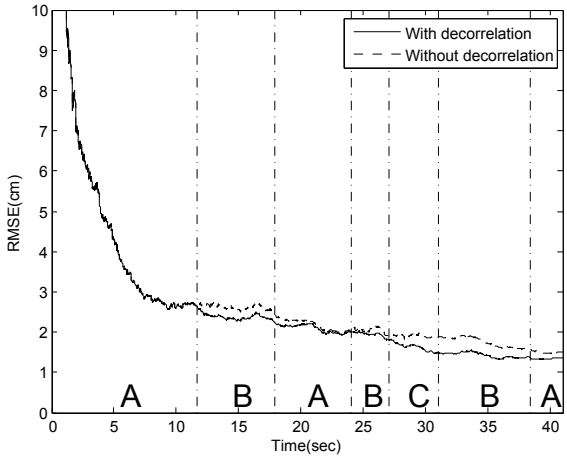


Fig. 7: Comparison of positioning errors for a scenario where all sensors are biased

In sensor networks, since a mobile target usually move through different types of regions, the process of decorrelation can effectively improve the position accuracy.

6. Conclusions

By using an overlapping hexagonal formation for performance analysis in cluster-based wireless sensor networks, we present a distributed positioning and tracking scheme to estimate the state of a mobile target. We assume that the network consists of two kinds of sensors. The first type represents the high-capability sensors, which are assigned as cluster heads. The cluster heads are responsible for data processing and inter-cluster communications. The second type is the collection of normal sensors, which are utilized for providing the measured data to their corresponding cluster heads. The cluster heads are designated to process measured data, to fuse the shared information from other CHs, and to perform distributed target positioning. We have designated the operating modes at the CHs and sensors for real-time information processing. The details of the mode settings and the operations are discussed.

In dealing with the redundant data processing problem, which may exist in the overlapping regions of WSNs, Cholesky decomposition has been used for decorrelating the measurement noises. The simulation results of a target moving through the region with overlapping clusters show that the overlapped cluster-based WSN with the designated modes provides more efficient and stable performance. The results for three-cluster cases also show that estimation with decorrelation process exhibits better tracking performance.

Acknowledgment

This work was supported under Grants No. NSC101-2221-E-327-003- and NSC98-2221-E-110-044- from the National Science Council, Taiwan, ROC.

References

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] D. Culler, D. Estrin, and M. Srivastava, "Overview of sensor networks," *IEEE Computer*, vol. 37, no. 8, pp. 41–49, Aug. 2004.

[3] K. Sohraby, D. Minoli, and T. Znati, *Wireless Sensor Networks: Technology, Protocols, and Applications*. New Jersey: John Wiley & Sons, 2007.

[4] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 2000, pp. 1–10.

[5] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, Oct. 2004.

[6] W.-P. Chen, J. Hou, and L. Sha, "Dynamic clustering for acoustic target tracking in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 3, pp. 258–271, Aug. 2004.

[7] B. Huang, F. Hao, H. Zhu, Y. Tanabe, and T. Baba, "Low-energy static clustering scheme for wireless sensor network," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, 2006, pp. 1–4.

[8] O. Boyinbode, H. Le, A. Mbogho, M. Takizawa, and R. Poliah, "A survey on clustering algorithms for wireless sensor networks," in *Proceedings of the 13th International Conference on Network-Based Information Systems*, 2010, pp. 358–364.

[9] M. Youssef, A. Youssef, and M. Younis, "Overlapping multihop clustering for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 12, pp. 1844–1856, Dec. 2009.

[10] D. Yan, J. Wang, L. Liu, and B. Wang, "Dynamic cluster formation algorithm target tracking-oriented," in *Proceedings of the International Conference on Computer Design and Applications*, vol. 4, 2010, pp. 357–360.

[11] A. G. Mutambara, *Decentralized Estimation and Control for Multisensor Systems*. Florida: CRC Press LLC, 1998.

[12] H. Durrant-Whyte, B. Rao, and H. Hu, "Toward a fully decentralized architecture for multi-sensor data fusion," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 2, 1990, pp. 1331–1336.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. New Jersey: Prentice Hall, 1993.

[14] G. Welch and G. Bishop, "An introduction to the Kalman filter," University of North Carolina at Chapel Hill, Tech. Rep., 2006.

A Study on the Estrus Detection System of the Sow Using the Wireless Sensor Network

Hoseok Jeong[1], **Hyun Yoe**[2]
[12]Dept. of Information and Communication Engineering, Sunchon National University,
Suncheon, Jeollanam-do, Republic of Korea
hsjeong@sunchon.ac.kr, yhyun@sunchon.ac.kr

Abstract - *This paper proposed WSN-based estrus monitoring system for stalled sows, which detects estrus of a sows in real time using wireless sensor network and notifies optimum time of insemination to PC and smart device of a user. The need for detect the estrus cycle has been increased in pig industry as it has the direct effect on productivity and earnings. If failed to detect in timely manner, it may cause a huge loss. The proposed system is one that detects estrus by measuring sow's activity in real time with an accelerometer sensor to transmits it to the server, which analyzes the information received and then informs whether or not to be in heat to PC and smart devices of users. This system is expected to improve rate of return for livestock farms by detecting the exact optimum time of insemination of a sows and minimizing the number of non-production days.*

Keywords: Wireless Sensor Networks, Ubiquitous, Sow, Estrus, Agriculture

1 Introduction

The WSN is a technology that deploys sensor nodes with computing and wireless communication capabilities in a diversity of application environments, forms networks autonomously, and then utilizes information collected from the sensor nodes for the purpose of wireless monitoring and controlling etc[1,2,3]. This WSN technology is a core one to realize a ubiquitous society, and is applied to every field of our life, that is, a variety of industries including distribution, logistics, construction, transportation, defense and medicine etc. to implement advancement of productivity, safety and human living standard[4,5]. Recently, the WSN technology is also applied to the agriculture field to improve the working environment and to increase its productivity[6].

Domestic hog industry is recently having hardships caused by increased production costs, varied animal diseases, unsophisticated management technology, etc. The technology detecting estrus of sows among these pig-breeding management technologies is one to decide the right time to fertilize, which is directly connected to the earnings of livestock farms, if the right time to fertilize is missed, it results in an economic loss because the number of non-production days for the sows becomes longer to reduce the productivity[7]. At present, the sensing of estrus is inefficient because it is mainly carried out manually.

This paper proposes a system to sense estrus of sows and decide the right time to fertilize them in the livestock farms experiencing such damages. Based on the fact that the activity of the sow is more increased in heat than in non-heat, an accelerometer sensor is attached in the form of collar on the neck of the sow in the stall, and measures activity to detect its estrus. The existing systems generally use schemes to directly insert a sensor into the body of sows or to exploit a CCTV, however, the former has a disadvantage that produces damages such as injuries on the body of sows and stresses when inserting the sensor, and the latter has a disadvantage that is too expensive[8]. It is expected that could solve disadvantages of the existing system and minimize the number of non-production days to increase earnings of the livestock farms by understanding the right time to fertilize sows to maximize the fertility rate through the proposed system. In addition, it is expected that would have an effect on improving the working condition and reducing the labor force of the livestock farms because they do not need to visit the pig house frequently for checking estrus.

This paper is organized as follows. The related studies of Chap. 2 explain the estrus of sows and the stall that is a breeding space, and the system design of Chap. 3 describes the structure of the proposed system and the procedure to process services. Chap. 4 implements the final system and measures its performance after verifying the system through experiments, and Chap. 5 finishes with a conclusion.

2 Related research

2.1 Estrus of pigs and detection

In general, a weaning sow begins estrus around 4 to 7 days after weaning. As estrus period lasts 3 to 6 days for pigs, which is relatively longer comparing to other animals, it is not very easy to detect the exact optimum time of insemination[9]. Particularly for artificial insemination, much smaller number of sperms and smaller amount of seminal fluid are inserted comparing to natural breeding and the motility or surviving rate of sperms and surviving time in the

[2] Corresponding Author

reproductive organ of a sow are relatively low, so it is necessary to monitor estrus carefully for insemination at optimum time. However, due to difficulty with continued monitoring, a manager generally checks estrus of a sow two times a day through observation with a naked eye. This method requires much labor as well as high level of skills and rich experience, and has limitations in finding exactly when estrus began with only two checks a day and judging optimum time of insemination exactly because estrus begins mostly at dawn[10].

When artificial insemination takes place at optimum time even when detecting estrus, conception rate is low, which causes economic loss. Because of this problem, artificial insemination is conducted 2 to 3 times, but costs and labor involving this procedure are another factor of adding burden to livestock farms[11,12]

The system proposed in this paper decided the time between 26~34 hours after starting estrus reported at present as the right time to fertilize in order to solve such a problem.

2.2 Accelerometer sensor

Accelerometer sensor is for measuring the dynamic force such as the Acceleration, vibration and shock of object by processing output signal and it has a wide area of usage since it can detect motion status in details[13].

For the purpose of verifying the system proposed in this paper, Accelerometer Sensor of HBE-Zigbexll by HANBACK Electronics Ltd. was used.

Figure 1. Accelerometer sensor

3 System design

System of estrus detection of stalled sows employs a method of measuring activities using accelerometer sensor. For a sow reaching estrus time, activity of walking around with unique sound increases. However, the sows in the stall have a characteristic that the sitting/standing up movements are increased and the lying ones are decreased compared to the non-estrus due to the limitation on movements[14]. It is a system that attaches an accelerometer sensor on the neck of sows in the form of collar based on this characteristic, measures the activity, sends this value to the server to decide whether or not to be in heat, and then informs to PC and smart devices of users in real time. In addition, after detecting the estrus, it decides the right time to fertilize, and informs it, so it could fertilize at the exact time.

3.1 System structure

The proposed estrus detection system consists of physical layer, middle layer and application layer, as shown in Figure 2.

The physical layer consists of accelerometer sensor and sensor node for collecting sow activity.

The sensor manager stores the information collected through the accelerometer sensor in the livestock management server database through storable format processing, measurable unit conversion and update inquiry of processed data.

The management server plays the role of storing in each table the data collected through the accelerometer sensor, as well as the sow activity data collected through the standard values for status notification. And management server database store sow activity data and sow identification information.

The management server is located between the user and database, and periodically notifies the user the data stored in the database. It automatically controls the corresponding estrus notification upon comparing the estrus standard values stores in the table and the status notification table, or comparatively analyzes the existing sow estrus information stored in the database and the measured sow estrus information to notify the producer in real-time of any values that exceed or fall short of the standard values through web and SMS notification services.

The application layer consists of application services that support various platforms such as laptop, web, PAD and smart phone and provides to users livestock estrus detection information service, sow optimum time of insemination service.

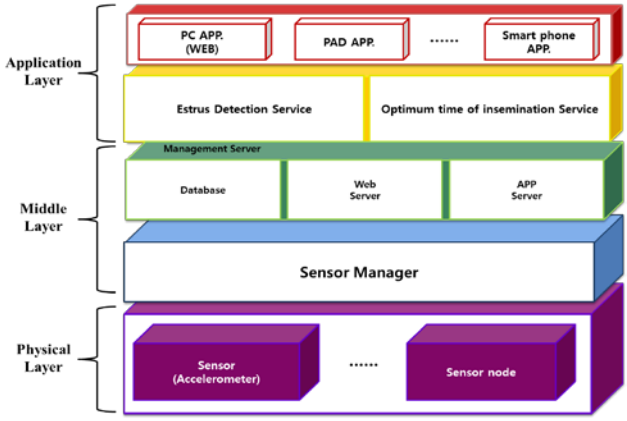


Figure 2. System structure

3.2 System configuration

The entire architecture of system of estrus detection of stalled sows based WSN is shown in figure 3.

This system consists of sensor of detecting estrus, server of storing and analyzing data transmitted from sensor and finally PC and smart device available to a user anywhere. Sensor used for monitoring judges whether or not a sow began estrus is accelerometer sensor. The sensor monitors activities of a sow, and stores and manages collected information in database. Server analyzes collected information, and notifies conditions of a sow to a user in real time after judging whether or not estrus began when the amount of activity exceeds the normal value.

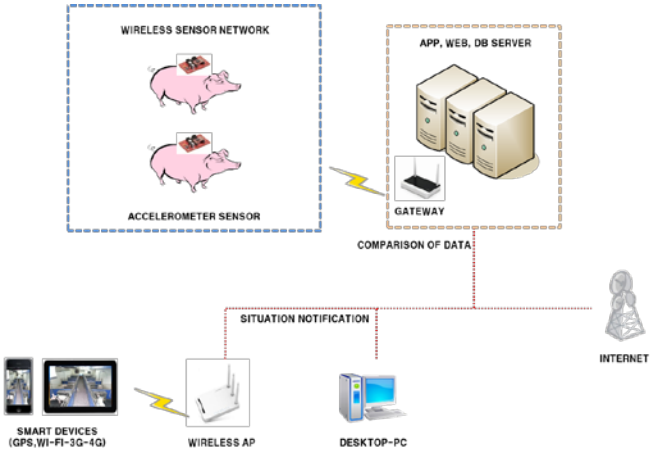


Figure 3. Proposed WSN based Estrus Detection System Configuration

3.3 System service

The proposed system provides information about estrus of a sow, and its operation is shown in figure 4.

Information measured through accelerometer sensor is transmitted to sensor manager, which stores processed data in database after processing transmitted data into format and converting units.

Management server requests sensor data of database on a regular basis, and compares transmitted data with data standard range of information; and when detecting estrus beyond the range, it analyzes information which took place to an individual sow and notifies it to a user.

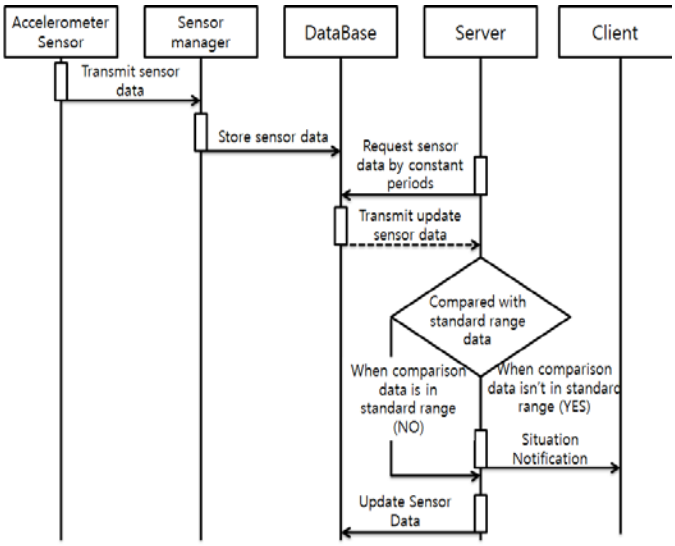


Figure 4. Service Process of Proposed WSN based Estrus Detection System

4 Implementation and result

The prototype model was tested by the method that persons wear accelerometer sensors fabricated in the form of collar to detect their movements. The reference scope was limited to measure activities by comparing with it, however, there were problems that the sensor in the form of collar dangled and there were noises. In order to solve these problems, it made the sensor in the form of collar could be adjust its size when applying it actually to the sow, values of the accelerometer sensor were passed through a LPF(Low Pass Filter) to remove noises and the required signals are obtained, so that the existing problems could be solved and the estrus of sows could be detected.

4.1 Prototype

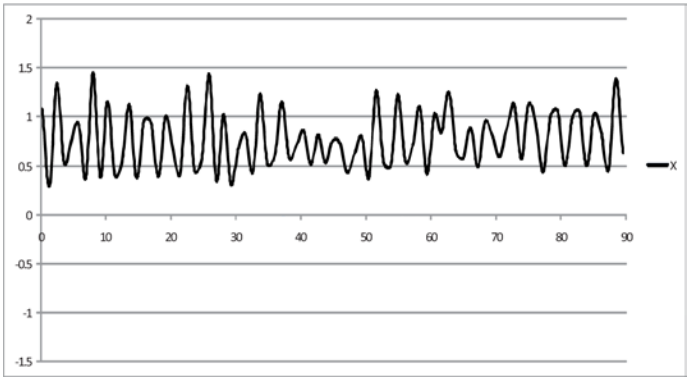


Figure 5. Accelerometer Result Values Graph of Test

Figure 5 is a graph showing acceleration result values of a test for the verification of system. As this system detects movement through accelerometer sensor attached to the neck of a sow, height variation value was used. This paper set a

standard range of data and detected estrus by comparing the amount of measured activity with the standard range. Through this experiment, this paper verified estrus detection system using accelerometer sensor.

The initial prototype of the WSN-based system to detect estrus of sows bred in the stall was fabricated with an accelerometer sensor in the form of collar, and its GUI was designed as figure 6. Comparing the graph of acceleration result values with the reference scope determined by counting movements during one minute, the estrus was detected.

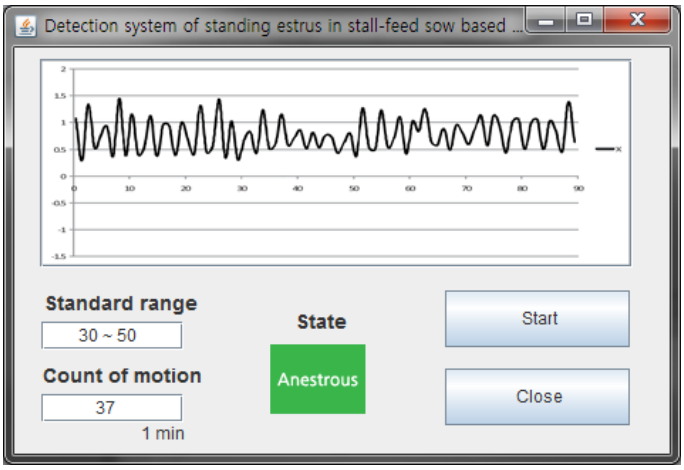


Figure 6. Estrus Detection System Prototype GUI

4.2 The Actual effect

Figure 7 is an image that applied actually to a sow bred in the stall. In order to solve the problem arisen in the prototype, when applying it actually to the sow, it made the size of the sensor in the form of collar could be adjusted, and the accelerometer sensor was intactly used.

Figure 7. Applied actually to a sow

Figure 8 is the GUI providing to users, where ① is the result of measuring activities by the accelerometer sensor, which is the result representing the hourly variation of activities. ② represents the sow's individual classification ID, its activity, the reference scope value of activities and whether or not to be in heat. ③ is the menu to set the reference scope of activities. ④ is the menu to inform the right time to fertilize by calculating time after detecting the estrus.

For the prototype, movements were measured during one minute, but the activities were measured for each time according to the sow's activity when implementing actually.

From the implementation result as above, it could verify the proposed WSN based estrus detection system, and it would like to advance it by developing continuously.

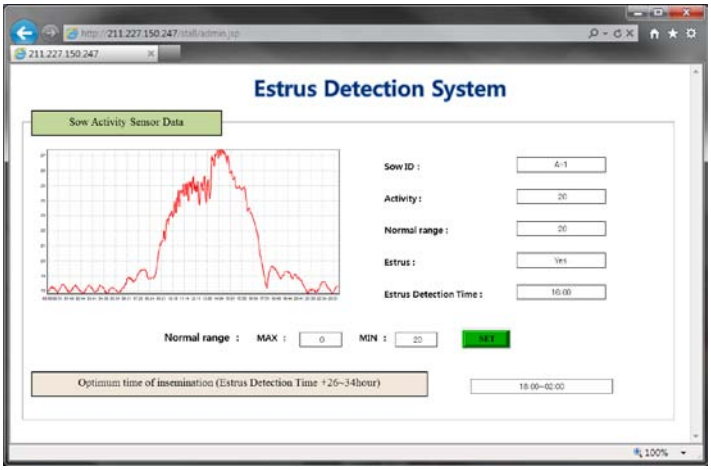


Figure 8. Estrus Detection System GUI

5 Conclusions

To maximize income, Livestock farms should increase reproduction rate by minimizing the number of non-production days for a sow. The system proposed in this paper has a purpose that detects estrus by measuring the sow's activity in real time with an accelerometer sensor, inform immediately the fact that the estrus is arisen and the calculated right time to fertilize to users, so that they could quickly deal with it. The existing method, which persons directly check estrus with their own eyes, is labor-intensive, and has a disadvantage that it has a lower chance of becoming pregnant even if detecting the estrus to make implantation and has much errors. The proposed system could solve problems of the existing method, improve the rate of detecting estrus of sows, and reduce labor force and production cost by deciding the right time to fertilize to increase the chance of artificial fertilization. In addition, it could get out of temporal confinement for detecting estrus to improve the quality of user's life and the working condition.

6 Acknowledgment

"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"

7 References

[1] Min-Nyun Kim, "A Study on Alarm System of a right Fecundation time for a Sow Using Ubiquitous Sensor Network(USN)", Journal of the Korea Academia-Industrial cooperation Society, Vol. 11, No. 1, pp. 92-95, 2010.

[2] Chul-Sik Pyo, Jong-Seok Cha, "Prospect of RFID/USN Technology Development for the Next generation", Journal of the Korean Institute of Communication Science(Infomation and Communication), pp. 7-13, August 2007.

[3] Seok-soo Kim, Gilcheol Park, Kyungsuk Lee, Sunho Kim, “Ubiquitous Military Supplies Model based on Sensor Network”, International Journal of Multimedia and Ubiquitous Engineering, 1, 1, 2006.

[4] Minseong Ju, Seoksoo Kim, “Logistic Services Using RFID and Mobile Sensor Network”, International Journal of Multimedia and Ubiquitous Engineering, 1, 2, 2006.

[5] Paul Golding, Vanesa Tennant Shin, “Evaluation of a Radio Frequency Identification (RFID) Library System: Preliminary Results”, International Journal of Multimedia and Ubiquitous Engineering, 3, 1, 2008.

[6] Jeong-Hwan Hwang, Hyun Yoe, “A Study on the Design of Animal Activity Monitoring System Using Wireless Sensor Networks”, Journal of the Korea Information and Communications Society (2011)

[7] K.Romer, F. Mattern, “The Design Space of Wireless Sensor Network”, IEEE Wireless Communications, 11, 6, 2004.

[8] Evans, L., Britt, J., Kirkbride, C. and Levis, D. Pork Industry Handbook, “Troubleshooting Swine Reproduction Failure”, Purdue University Cooperative Extension Service, 2001.

[9] Perez, J. M., Mornet, P. and Reat, A., “Leporc et son elevage”, Maloine, Paris, France, 1986, pp. 575.

[10] Dong-Joo Kim, Seong-Chan Yeon, Hong-Hee Chang, “Development of a Device for Estimaing the Optimal Artificial Insemination Time of Individually Stalled Sows Using Image Processing”, J. Anim. Sci. & Technol Korea, 2007.

[11] Diehl, J. R., Day, B. N. and Flowers, W., Pork Industry Handbook, “Estrus or Heat Detection”, Purdue University Cooperative Extension Service, 2001.

[12] Jang-Hee Lee, Soon-Hwa Baek, Seung-Ho Yon, “Auto Dispatch Device of Parturition Beginning Signal by Temperature and a Load Sensor at Ubiquitous Circumstance in Pig Industry”, Reprod Dev Biol, 33, 3, 2009.

[13] ″Accelerometor sensor”, http://terms.naver.com/entry.nh

n?docId=1264597&mobile& categoryId=200000576.

[14] Freson, L., Godrie, S., Bos, N., Jourquin, J. and Geers, R., “Validation of an infra-red sensor for oestrus detection of individually housed sows”, Computers and Electronics in Agriculture, 1998.

Novel Model for Vehicle's Traffic Monitoring using Wireless Sensor Networks between Major Cities in South Africa

[1]Munienge Mbodila and [2]Obeten Ekabua
Department of Computer Science, North-West University, Mmabatho, Mafikeng, South Africa
{[1]24087467, [2]obeten.ekabua}@nwu.ac.za

Abstract *- With the growing number of vehicles and users, monitoring road and traffic within cities is becoming a huge research challenge. With the urban scale enlargement coupled with the exponential growth in the number of vehicles, South Africa (SA) is not an exception. Consequently, congestion and pollution (i.e. noise and air) have become the order of the day. Road congestion and traffic-related pollution are well-known for huge negative socio-economic impact on several economics worldwide. For over a decade now, the number of cars on SA roads has increased tremendously and the road transport profile is characterized by its sizeable and total dependence on cars particularly in the highly developed urban areas alongside cycling, and other public transport. This has brought about increasing congestion in public roads which poses a serious problem not only for SA, but many countries of the world and has to be contained. Several solution methods have been proposed requiring dedicated hardware such as GPS devices and accelerometers in vehicles or camera on roadside and near traffic signals. Most other works in literature concentrated on lane system and orderly traffic, which is common in developing world and in some cases, the traffic is highly chaotic and unpredictable. The situation in SA cities like Johannesburg and Pretoria are not different. All there methods are costly and human require much effort. Therefore, in this paper, we propose a novel model that is cost effective, requires less human intervention, but uses wireless sensor networks application to monitor traffic in major SA cities.*

Keywords: Wireless Sensor Networks, GPS, Traffic, Congestion, Nodes

1 Introduction

Traffic vehicle monitoring in South Africa (SA) is becoming more and more vital due to urban scale enlargement coupled with the exponential growth in the number of vehicles. With this development, congestion and pollution (i.e. noise and air) have been the order of the day. Road congestion and traffic-related pollution are well-known for huge negative socio-economic impact on several economics worldwide. Over the last 10 years the number of cars on SA roads has increased tremendously by almost 30% and road authorities are struggling to contain the effects of the growing congestion that resulted [1]. The road transport profile is characterized by its sizeable and total dependence on cars particularly in the highly developed urban areas alongside cycling, and other public transport. Increasing congestion level in public road networks is a growing problem not only applicable to SA but also in many countries of the world and has to be contained. The growth of urban areas in SA has been affected by the increased of traffic flow in most of the roads. As road networks usage increases, traffic congestion increases also characterized mainly by slower speeds, longer trip times, and increased vehicular queuing occurs. In order to keep the situation under control, there are several monitoring systems that exist. Traffic monitoring of vehicles is a complex issue because it requires real-time monitoring, however, the data processed by the monitoring system are huge causing high throughput computation. With the advances in the technology of micro-electromechanical system (MEMS), developments in wireless communications and wireless sensor networks (WSNs) have also emerged [2].

Given the expected growths in urban areas traffic, the scale and complication of the traffic infrastructure will continue to rise gradually in time and in distributed geographical areas in SA. To guarantee vehicles monitoring efficiency, safety, and security in the presence of such growth, and avoiding pollution, it is critical to develop a system that can adapt to enlargements while guaranteeing reliable in urban roads in SA. With the development of WSNs many cities around the world have developed variety of technologies and systems better manage and control their roads network. For instance, currently the majority of urban roads in SA is controlled by *i-Traffic system* which is an integrated system of CCTV cameras linked by fiber optic cable to a central control Centre [1]. At the control center, human operators are in charge for continuously monitoring and analyzing a huge amount of data from video cameras system on the roads. The human decision makers must indicate the correct approach by analyzing the CCTV information, and then inform the ground officer and/or remotely configure traffic control equipment using the communication infrastructure. Unfortunately, this system for traffic monitoring has several

problems in it due to human decisions involvements. Therefore, in this paper, our objective is to overcome some of the limitations that exist in the current traffic monitoring system. To this end, we are proposing a novel system for vehicle traffic monitoring that will use WSNs technology to monitor traffic congestions between two major cities roads in SA.

The paper is organized as follows; section 1 is the introduction, section 2 gives the background information of study, section 3 gives the research proposed approach, section 4 describe the system components, while sections 5 and 6 gives the architecture of the system and the paper conclusion respectively.

2 Background Information

In a typical application, a WSN is scattered in a region where it is meant to collect data through its sensor nodes. WSN are being used in industrial process monitoring and control [3,4], machine health monitoring [5], environment and habitat monitoring, healthcare applications, home automation and traffic control [6,7]. However, one of the major challenges that vehicle control and traffic management applications are facing knowing the position, the lane and speed of the vehicles on the road network in real-time basis [4]. WSN received significant attention in the last decade and successful research put them in the forefront to answer this challenge [4]. WSNs technology can help in the infrastructure development of our novel system for vehicle traffic monitoring. They are multi-hop which autonomously form a network which include a large number of nodes integrating information collection, data processing and wireless communications in order to perceive, collect and process information from objects in the region and allow observers to know when, where and what the incident will occur. A typical WSN node is composed of power, data acquisition unit DAU, data processing unit DPU, data sending and receiving unit DRSC [8]. Each hardware unit has a specific task in the system as shown in Fig.1.

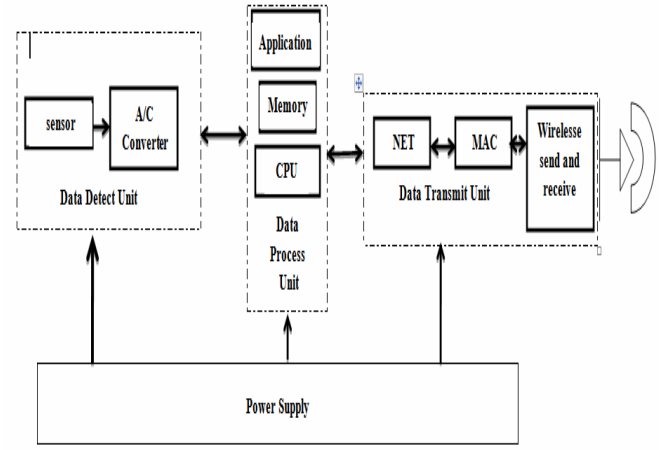


Fig. 1: General structure of a node [8]

2.1 Wireless sensor network application

The development of WSNs was inspired by military applications such as battlefield surveillance and target tracking. Some usual applications of WSNs are monitoring, medical care, Intelligent Transportation and Traffic Management, Parking Monitoring, and Automatic Traffic Control. These are discussed as follows:

1) *Medic Care:* William Walker et al, has mentioned about health care monitoring [9]. In their discussion sensors measure the vital signal of patient such as blood pressure, oxygen rate, heart pulses and according to desecration of their threshold value it sends the message to nurse or doctor.

2) *Agriculture Monitoring:* A system was proposed about agriculture monitoring by N. Medrano [10], where by sensor senses the wide cultivar area in real time and gives the exact answers to variations in the cultivar condition, which improves the product quality.

3) *Structural Monitoring:* In civil engineering and construction, Vivek Katiyar has mentioned that civil structure like bridges, building, and water reservation can be monitor with the help of wireless sensor network. It was said that WSN is more helpful for checking condition of some serious structure at regular interval of time [11].

4) *Intelligent Transportation:* WSNs is extensively being used in nowadays in the area of transportation, where mostly use of automatic traffic control systems, efficient multi storage parking location building identification and many other system. BI Yan-Zhong et al in their argument [12] has stated that wireless sensor network is very suitable in multi storage parking building where sensor nodes are deployed at each parking space. Whichever parking space is free, appropriate sensor sends the message to control centre which will guide the vehicle to that direction.

Finally, Malik Tubaishat also has discussed a system for reducing the traffic by real time monitoring of vehicles using wireless sensor network [13]. In the discussion the concentration of the traffic is been measured, all the sensor nodes on different signals co-ordinate with each other and dynamically changes the duration of green signals. That makes help in reducing traffic in peak hours. Chen Wenjie et al in [14] have also talk about real time dynamic traffic control system using wireless sensor networks. It is obvious that the unique feature of WSNs can assist in building diverse application for efficient vehicle traffic monitoring.

2.2 Logical structure of wireless sensor network

Since the features show that the applications of wireless sensor networks for traffic monitoring have no space

constraints, more flexible distribution, mobile convenience and quick reaction [8] but its architecture is different from diverse application. Regardless of the architecture of wireless sensor network its several parts are logically the same as shown in Figure 2. [15].

The physical hardware layer includes network infrastructure, sensors and other hardware related to wireless sensor networks. It is responsible for modulating, sending and receive data. Data link layer provides communication between physical layer and network layer and establish a data link between adjacent nodes, send the frame organized by a certain format to provide reliable information transmission mechanism for the network layer. The network layer deal with routing, data transfer and other issues between sensor nodes and between sensor and observer, including from the physical connection to the exclusive agreements of WSNs applied to each layer. Application layer include the specific application to meet the user's need, such as traffic flow forecasting.

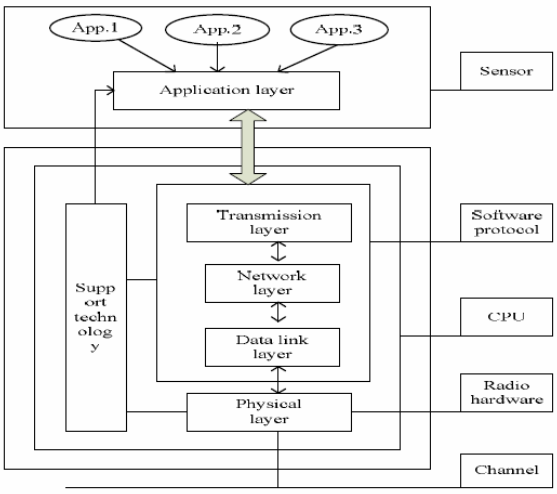


Fig 2. Logical structure of wireless sensor network [8]

3 Proposed Model Approach

Traffic vehicle monitoring in SA is becoming more and more vital due to urban scale enlargement coupled with the exponential growth in the number of vehicles. The continuous increase in the congestion level on public roads, especially at rush hours, is a critical problem not only in SA but in many countries and is becoming a major concern for transportation specialist and decision makers [15]. Currently, the majority of urban roads in SA are control by i-Traffic system which is an integrated system of CCTV cameras linked by fiber optic cable to a central control Centre. At the control center, human operators are in charge for continuously monitoring and analyzing a huge amount of data from video cameras system on the roads. This system for traffic monitoring has several problems arising from it and its required dedicated communication infrastructures that are expensive. With the development of WSNs many cities around the world have developed variety of technologies and systems to better manage and control their roads network. In this paper we review different literature on WSN for vehicles traffic monitoring system, evaluate the existing vehicle traffic monitoring system in SA and in other countries and finally use the knowledge obtained to develop a novel model for vehicles traffic monitoring using WSN that is cost-effective to overcome the challenges of the current system. The hardware components of the proposed system and their functionality together with the proposed architecture for the system are discussed in details in the following section.

4 Proposed System Components

The hardware components of the proposed system are described on the basis of their roles and responsibility regarding to the system and their internal data and communication models in the whole system.

4.1 GPS

Global Positioning System (GPS) is use to establish a position at any point on the globe and determine the location of any object (vehicle, building, person, areas, roads etc.); further more record the position at regular intervals in order to create a track file or log of activities. GPS coverage in SA is wide and most urban roads are mapped to the GPS system. In this project we use GPS system in order to compute possible shortest route close to the congestion areas to help the traffic officer in TMC regenerate message to the vehicle users on the roads.

4.2 Sensors

A sensor is a tiny device with wireless communication capabilities that responds to a physical stimulus (such as het, sound, magnetic, pressure etc.)[1]. Sensors are mostly used to change a physical parameter such as blood pressure, room temperature, motion, wind speed or vibration into a signal that can be measured electrically and convert the physical parameter to an electrical equivalent it is easily inputted into a computer or a microprocessor for manipulating, analysing and displaying [8]. In this research proposal we make uses of sensors, placed at many points throughout the road network (intersection or places where congestion is extremely occur) to sense congestion by counting and measuring the speed of passing vehicles. The sensors are place on the road side in group (Zone1 up to zone4) in a way that if a vehicle cannot be sense by one zone it will be able sense by the other zone.

4.3 RFID scanners

Radio frequency identification (RFID) system works for identification of items or objects wirelessly. Sometimes it only identifies item category or type but it is capable of

identify items or objects uniquely. Schwieren1 and Vossen [21] said RFID also enable data storage for remote items or objects trough remotely access items information. In general most of RFID system consists following components*: RFID tags*, this is use as a unique identifier. *RFID Antennas*, these are first point of contact for the tags reading. *RFID Scanner / readers*, this is usually consists of a radio frequency module, a control unit and coupling element to interrogate the tags via radio frequency communication. *RFID middleware*, this is a kind of software that stands between the scanner / reader network and the application software to assist processing data generated by the reader network. Middleware has the ability of detecting the movement of RFID tags as they pass the read range of one to another [4, 5, and 6]. In our proposed system novel system for traffic monitoring, the scanner/reader should be design like the e-toll placed at distance of 1km from each used junction and the tag should be placed on the top of vehicle for easy wireless scanning of the tags.

4.4 Mobile services

This is access by the traffic officer in duty; to generate randomly message to the drivers using the roads. Basically our proposed system has an infrastructure that consists of wireless sensor network, RFID system, GPS and mobile services. Basically the Sensor nodes deployed alongside of the roads where possible congestion occurs will sense an obstacle along the roads. Assuming that there is congestion or a problem on the road, sensor nodes will then send the sensed data to Traffic Monitoring Centre. The traffic monitoring centre will then generate a message using mobile service for appropriate routing determine by the GPS system so that the drivers can take appropriate routing action to avoid congestion or problem on the roads. This message will be send to all the vehicles scanned in 1km distance where the RFID system was installed, and then drivers who receive this message will take suitable action to escape congestion on the specific roads.

4.5 Traffic monitoring center

This is the main traffic office where all the roads are monitored and decisions are made for the state of the roads. The traffic monitoring centre is operated by a traffic officer who is in charge of continuously monitoring and analysing data from the entire system. The traffic monitoring centre (TMC) will decide the correct approach by analysing the congestion information from the WSN system every second and decided if action is to be taken or not. But in the case of congestion the traffic officer at the TMC will immediately generate a message (result of the GPS communication) and send it to the appropriate entity identify by RFID system on the road using their unique number plate. Then vehicle drivers on the ground will be inform of alternative roads to take using information received from the TMC.

5 System Architecture

Wireless sensor networks is an information monitoring and transmitting network that can be applied to any type of traffic flow monitoring and forecasting system [18]. In this project, we have propose a novel model for vehicle traffic monitoring based on WSN that can be appropriate to any types of urban city setting in South Africa by means of investigating the current status of vehicle traffic monitoring networks in the entire world in general and South Africa in particular; and combining wireless sensor network technology, RFID system and GPS technology to communing to the road users.

The proposed architecture shown in figure1 consists of wireless sensor network, GPS, RFID scanner, and mobile services. The sensor node installed on the road's junction will sense for congestion. Sensors will send information using (congestion images) gateway to the traffic monitoring center (TMC) if observed within route. The control center will simultaneously communicate with GPS and RFID scanner. Communication with the GPS is for possible shortest route computation, while the RFID scanner immediately scans the RFID tag deployed on every moving vehicle from a distance of 1km. The output of such scan is unique number (vehicle number plate in our case) which is use by the traffic monitoring center to communicate with specific vehicle's driver in that specific congestion area. Immediately the scanned unique number is received, the traffic control center will generate a message (result of the GPS communication) and send it to the appropriate entity on the road to use the closed roads with no congestion.

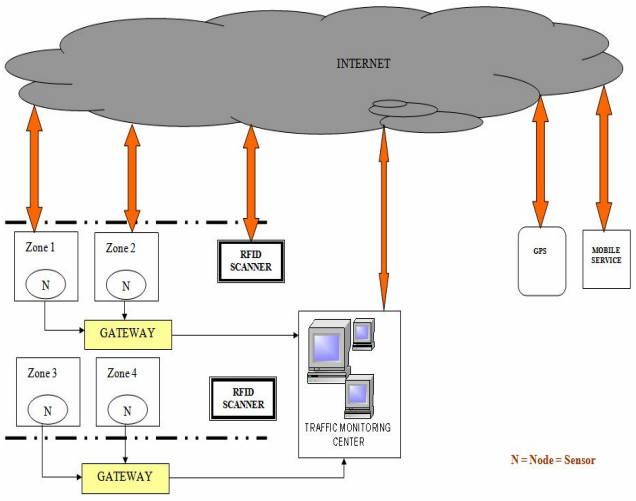


Fig. 3: The proposed system architecture

6 Conclusions

In this paper, we have presented the design of a novel system for traffic monitoring using wireless sensor network. The main purpose of the system is to monitor vehicle traffic

between two cities in South Africa. The system architecture of our novel system consists of wireless sensor network, GPS, RFID scanner, and mobile services. The sensor node installed on the road's junction will sense for congestion. This system is cost effective and can be use in any type of city.

7 References

[1] B. Barbagli, L. Bencini, I. Margini, G. Manes, A. Manes, "A Real-Time-Traffic Monitoring based on wireless sensor networks technologies", *Wireless Communications and Mobile Computing Conference (IWCMC),* pp. 820-825, 2011.

[2] K. Vivek and el, "An Intelligent Transportation Systems Architecture using Wireless Sensor Networks" international journal of computer application, Vol.14-No2, Jan 2010.

[3] D. J. Cook, and S.K. Das, "Wireless Sensor Networks" *Smart Environments: Technologies Protocols, and Applications*, *Wiley Publications*, Oct. 2004.

[4] T. Haenselmann. (2006) Sensor networks. GFDL Wireless Sensor Network textbook. Available

[5] A. Tiwari, P. Ballal and F.L. Lewis, "Energy-efficient wireless sensor network design and Implementation for condition-based maintenance", *ACM Transactions on Sensor Networks*, Vol. 3, No. 1, pp. 1-23, Mar. 2007.

[6] R. Kay and F. Mattern, "The Design Space of Wireless Sensor Networks", *IEEE Wireless Communications* Vol. 11, No. 6, pp. 54–61, Dec. 2004

[7] S. Hadim and N. Mohamed, "Middleware: middleware challenges and approaches for wireless sensor networks," *Distributed Systems Online, IEEE*, Vol.7, No.3, pp. 1-23, Mar. 2006.

[8] B. Xu, P. Hong et al, "Research on Traffic Monitoring Network and its Traffic Flow Forecast and Congestion Control Model Based on Wireless Sensor Network", International *Conference on Measuring Technology and Mechatronics Automation, IEEE*, pp. 142-147, 2009
International Journal of Computer Applications, Vol.14-No2, Jan. 2010

[9] H. Abhiman, P. Todd, W. William and B. Dinesh, "Self Powered Wireless Sensor Networks for Remote Patient Monitoring in Hospitals", *Sensors*, Vol.6, pp. 1102-1117, 2006.

[10] J. Hwang, C. Shin, H. Yoe, "Study on an Agricultural Environment Monitoring Server System using Wireless Sensor Networks", *Sensors*, Vol. 10(12), pp. 11189-11211, Dec. 2010.

[11] K. Vivek et al, "Recent advance and future trends in wireless sensor networks", *International Journal of applied engineering research, Dindigul*, volume 1, No 3, 2010.

[12] Y.Z Bi, L.M Sun, H.S Zhu, T.X Yan, Z.J Luo, "A Parking Management System Based on Wireless Sensor Network", *ACTA AUTOMATICA SINICA* Vol. 32, No. 6, Nov. 2006.

[13] T. Malik, S. Yi and S. Hongchi, "Adaptive Traffic Light Control with Wireless Sensor Networks", *Consumer Communications and Networking Conference (CCNC)*, 4th *IEEE*, pp. 187- 91, 2007.

[14] W. Chen, C. Lifeng, C. Zhanglong, T. Shiliang, "A Realtime Dynamic Traffic Control System Based on Wireless Sensor Network," *International Conference on Parallel Processing Workshops (ICPPW)*, pp.258-264, 2005.

[15] Shruthi, KR & Vinodha, K (2012), priority based traffic lights controller using wireless sensor networks, *IJESS* ISSN: 2231-5969, vol-1 Iss-4, 2012.

[16] Song Dong, Han Dong, Ma Xuesen, Luoqian, *"Design of an Intelligent Bus-upervised System Based on Wireless Sensor Networks", Journal of Hefei University of Technology*, Vol. 31 No. 1 Jan. 2008, pp.21-23.

[17] Zhu Liang, Fan Yonghua, *"Embedded Network Monitoring System of Wireless Temperature Sensor"*, Microcomputer Applications, Vol.28, No.2, Feb. 2007, pp.145-149.

[18] Xiao Laisheng, Peng Xiaohong et al, *"Research on Traffic Monitoring Network and its Traffic Flow Forecast and Congestion Control Model Based on Wireless Sensor Network", 2009* International Conference on Measuring Technology and Mechatronics Automation, *2009 IEEE computer society,*

[19] Jeffrey Considine, Feifei Li, George Kollios, and John Byers, *Approximate Aggregation Techniques for Sensor Databases Computer Science Dept., Boston University*

[20] Dr. Harshal Arolkar, Kashyap Dhamecha, Darshan Patel, *"architecture for Accident Monitoring in BRTS corridors using Wireless Sensor Networks", Vol.2 Issue I ISSN (online): 2230-8849,* January 2012

[21] Schwieren1, J. & Vossen, G, 92009), *"A Design and Development Methodology for Mobile RFID Application based on the ID-Services Middleware Architecture"*, Tenth International Conference on Mobile Data Management; Systems, Service and Middleware, IEEE Computer Society.

Dynamic TDMA Sense Window Prediction for Wireless Sensor Networks

Charith Gunasekara and Ken Ferens
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB, Canada
Email: Charith_Gunasekara@umanitoba.ca, Ken.Ferens@ad.umanitoba.ca

***Abstract*—In this paper, we develop a TDMA sleep scheduling technique for Wireless Sensor Networks using a queuing theoretic approach. We consider the scenario where TDMA listening by wireless sensor cluster head wastes energy when the nodes send data in low rates. To save energy consumption we propose a technique of a traffic aware TDMA sensing window which allows the cluster head antenna to go to sleep mode after listening for a portion of the TDMA time slot if there are no data traffic arrivals. We model the traffic patterns from each node using a discrete time Markov chain and the next arrival probability is calculated based on the patterns of past traffic arrivals. This allows us to accurately predict co-related data traffic behaviors in wireless sensor networks. Hence we define the TDMA sense window size proportional to the predicted probability of traffic arrivals.**

Keywords: Wireless Sensor Networks, TDMA Sense.

I. Introduction

Wireless sensor nodes are autonomous self organized communication nodes used to sense different environmental conditions. There is a central base station which monitors and controls these nodes. This central base station gathers sensing data through the nodes. It is not energy efficient for every node to directly communicate with with this central station. Therefore the nodes are organized in to different clusters and the cluster head is responsible for collecting data from every node in the range and then forward the collected data to the monitoring station. This clustering technique is being adopted by all major algorithms used in wireless sensor networks.

Intra cluster communication can be done using TDMA technique to have an ideal, collision free communication between the cluster head and the nodes. Each node is given a chance to transmit data to the cluster head during a TDMA cycle. The cluster head antenna has to listen through out each TDMA slot to gather data from each node. This listening is energy consuming and if the nodes sends data in a low rate because sensor nodes may not have data to transmit to the cluster head all of the time. When no data is sent from a sensor node to the cluster head during the TDMA time slot, the cluster head will be wasting energy listening and waiting for sensor data.

As a solution to this, the cluster head can dynamically calculate the average data arrival rate from each node and reschedule the TDMA frame size accordingly but this technique adds overhead of resynchronizing all nodes to a new TDMA schedule frequently. The cluster head can listen to a portion of the TDMA time slot and see if the node is attempting to send any data if not it can turn off the antenna to the rest of the TDMA time slot to save energy consumption. This method is called TDMA sense and it can save energy consumption by a considerable amount. However finding this TDMA sense window size is challenging. If the sensing window is too small, there is a high chance that a node actually attempts to transmit data during the latter part of the TDMA slot but due to the unavailability of cluster head's attention, the node has to delay it's data until it's next TDMA slot. If the sensing window size is too large while the node has no new data to send, the cluster head will be unnecessarily wasting energy by turning on the antenna. One obvious way to determine the size of this TDMA sense window is to let the cluster head to monitor data arrivals from each node and decide the TDMA sense window sizes proportional to the average data arrival probability of each node. But it is highly likely that the data generated by wireless sensor nodes are correlated and the odds of sending a data packet during a TDMA slot is depending on the past data traffic patterns.

In this paper we present an efficient and novel technique for the cluster head to determine the data arrival probability by analyzing past traffic patterns. We propose a technique based on N- order Markov chain to dynamically train the cluster head to keep track of the data traffic patterns generated by each node hence calculate the data arrival probability. Based on the dynamically predicted data arrival probabilities , the cluster head adjudges the TDMA sense window size.

The rest of the paper is organized as follows. Section II describes the related work done in literature and contributions of this paper. In section III we present our system model and describe how we model the traffic behaviors in the network using an N-Order discrete time Markov chain (DTMC) and then in section III we demonstrate simulation results obtained by the proposed system model. Finally, section IV concludes the paper.

II. Related Work

A major constraint in wireless sensors nodes is battery life. Depending on the place they are deployed it will be very hard to replace the batteries. Therefore energy consumption in wireless sensor nodes has been intensively studied in literature.

Operation	Time	Power Consumption
Initialize radio	350 μs	18 mW
Turn on radio	1.5 ms	3 mW
Switch to TX/TX state	250 μs	45 mW
Receive 1 byte	416 μs	45 mW
Transmit 1 byte	416 μs	60 mW
Active	-	32 mW
Sleep	-	90 μW

TABLE I: Power Consumption of Mica2 sensor

Most algorithms proposed for wireless sensor networks try to minimize the energy consumption in wireless nodes. Energy models for different wireless nodes have been studies in [1] [2] [3] which shows the power consumption rates for different sensor nodes while the antenna is in different modes. Table I is some interesting data for Mica2 sensor node. It shows how the power consumption is dramatically reduced in the antenna sleep mode. Once the antenna is turned off, it consumes additional energy for turning on the radio and initialization of the antenna. The time and energy overhead of transitioning between active and sleep states may consume energy so these state transitions should be carefully done.

Time Division Multiple Access (TDMA) is a good way of reducing the energy consumption in sensor nodes. which allows sensors to turn off antenna until the allocated time slot. TDMA scheduling for intra-cluster communication in wireless sensor networks was introduced in [4]. The cluster head always has to turn on the antenna to listen data from each nodes in its rage, so it is not very beneficial for the cluster head regarding the energy consumption. The authors in [5] proposed to change the TDMA frame size according to the data arrival rate of the nodes. But this technique requires the cluster head to update the schedule with the nodes which cause additional over heads and energy consumption.

A TDMA sense technique was proposed in [6]. During the sensing window, 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. This technique is an interesting concept to save energy but if the listening window size is not carefully selected it may not be very effective and also it can cause unnecessary delays for data packets.

In this paper we propose a dynamic method to train the cluster head to calculate the data arrival probability based on past traffic patterns. Hence dynamically change the TDMA sense window size proportional to the predicted data arrival probability. We use N-Order Markov chain to keep track of the past traffic arrived from each node and then dynamically update the patterns as new data arrives from node. Usually Markov chain dependent on only one past event [7] but in this case we use a higher order Markov model which models the system depending on the past N events. This allows us to efficiently find a relationship among different data traffic patterns. In the simulation section we show that our prediction scheme performs well compared to fixed windows size scheme which defines the TDMA sense window sizes proportional to the average data arrival probability from each node.

III. System Model

We consider a cluster head which communicate with the nodes in its range using TDMA time scheduling. TDMA time frame is chosen such that each node can transmit one data packet. The cluster head listens to data coming from each node and find traffic patterns. To do this the cluster head calculate the probability of having an arrival after each traffic pattern. A data packet arrival during a TDMA cycle is denoted by a binary word. Binary 1 represents a data arrival and binary 0 represents no arrival. The cluster head keeps track of data arrivals from past N TDMA cycles from each node. i.e. the cluster head has a N-bit binary word to represent the past traffic pattern for each node. While doing this the cluster head calculates the probability of having an arrival in the next time slot after each traffic pattern using the real data. We use an N-order DTMC to keep track of these calculated probabilities. Here each state is denoted by an N bit binary word representing the traffic pattern. Fig. 1 shows the graphical representation of an N-Order Markov chain for $N = 3$. It shows how the system goes from one state (traffic pattern of past N TDMA slots of the node) to another according to the data arrivals during the next time unit. A data arrival in the next time slot is denoted by 1 and 0 is to denote no data arrival. There are 2^N number of states in the Markov chain, from each state $S_k \in S$ it can go to one of two states depending whether there is a data packet arrival or not during the next TDMA cycle. Here S is the total state space of the Markov chain.

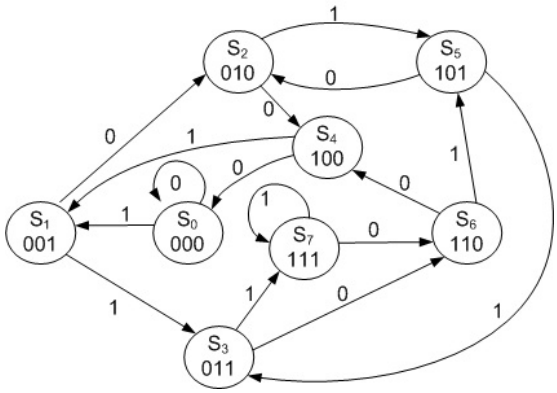


Fig. 1: N-Order Markov Chain (N=3)

A. Training the N-Order Markov Chain

These state transition probabilities can be found by training the system for real data for a reasonable amount of time and these probabilities derive a transition matrix for this N-Order DTMC.

For the purpose of training the cluster head stores an M bit word $B = b_1 b_2 \cdots b_M$ for each sensor node defining whether there was a data transmission by the sensor node to the cluster head or not. The cluster head uses past $M(>> N)$ TDMA cycles to update the Markov chain.

$$P = \begin{array}{c} \\ S_0 \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \end{array} \begin{array}{c} \begin{array}{cccccccc} S_0 & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 \end{array} \\ \begin{pmatrix} P_{(0|S_0)} & P_{(1|S_0)} & & & & & & \\ & & P_{(0|S_1)} & P_{(1|S_2)} & & & & \\ & & & & P_{(0|S_2)} & P_{(1|S_2)} & & \\ & & & & & & P_{(0|S_3)} & P_{(1|S_3)} \\ P_{(0|S_4)} & P_{(1|S_4)} & & & & & & \\ & & P_{(0|S_5)} & P_{(1|S_5)} & & & & \\ & & & & P_{(0|S_6)} & P_{(1|S_6)} & & \\ & & & & & & P_{(0|S_7)} & P_{(1|S_7)} \end{pmatrix} \end{array} \quad (1)$$

Algorithm 1 Cluster Head Training Algorithm to calculate Markov chain probabilities

1: Set $Count_{S_k} = 0$, $Count_{(1|S_k)} = 0$, $\forall S_k \in S$
2: **for** $i = 1$ to $M - N + 1$ **do**
3: Find state $S_k \in S$ defined by $(b_i b_{i+1} \cdots b_{i+N-1})$
4: $Count_{S_k} = Count_{S_k} + 1$
5: **if** $B(i + N) = 1$ **then**
6: $Count_{(1|S_k)} = Count_{(1|S_k)} + 1$
7: **end if**
8: **end for**
9: **for** $\forall S_k \in S$ **do**
10: $P_{(1|S_k)} = \frac{Count_{(1|S_k)}}{Count_{S_k}}$
11: **end for**

B. TDMA Sense Listening Window

When the cluster head listens to data coming from its children nodes, it turns the antenna to listening mode during a portion of the TDMA time slot and see if there is a packet arrival. If there is no packet arrival it turns antenna to idle mode to save energy as shown in Figure 2. This TDMA sense listening window length τ is increased according to the the probability $p = P_{(1|S_k)}$ of packet arrival during the next time slot given that the system is in state S_k

With the knowledge of traffic prediction we can define this TDMA sense listening window τ proportional to the data arrival probability as follows,

$$\tau = \begin{cases} pT & \text{if } pT < t_s + t_w \\ T & \text{otherwise} \end{cases} \quad (2)$$

Here t_s and t_w are the time taken to switch antenna from listening state to idle and idle to listening respectively.

If the sleeping time is smaller than the two state transition times then the antenna avoid going in the sleep state.

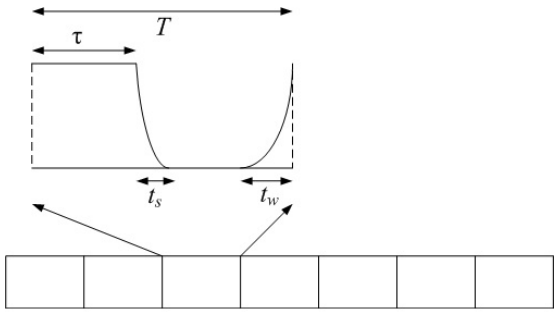


Fig. 2: TDMA sense time slot structure

If there is a packet arrival during this TDMA sense window at time $\tau_a < \tau$, then the cluster does not switch the antenna to sleep mode, otherwise the radio is listening during the whole sensing window τ and turns it to idle mode.

Now we can calculate the probability of a packet delayed by one TDMA cycle because it arrived during cluster head sleep period during the allocated time slot for that node.

$$P_{delay} = \frac{(T - \tau - t_s - t_w)}{T} p \quad (3)$$

We can find the total energy consumption during each TDMA time slot as follows,

$$E = E_{trans} + E_{pkt} + \tau' W_l + (T - \tau' - t_{trans}) W_{idle} \quad (4)$$

where,

$$E_{trans} = \begin{cases} W_s t_s + W_w t_w & \text{if } pT < t_s + t_w \text{ AND no data} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$t_{trans} = \begin{cases} t_s + t_w & \text{if } pT < t_s + t_w \text{ AND no data arrival} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$\tau' = \begin{cases} \tau & \text{if } pT < t_s + t_w \text{ AND no data arrival} \\ \tau_a & \text{data arrival at } \tau_a (\leq \tau) \\ T & \text{otherwise} \end{cases} \quad (7)$$

t_s - The transition between listening to idle state
t_w - The transition between idle to listening state

W_s - Average power consumption during the transition from listening to sleep state

W_w - Average power consumption during the transition from idle to listening state

E_{pkt} - The average energy consumption during a data packet transmission

W_{idle} - The power consumption when the receiver is in the idle state

W_l - The power consumption when the receiver is in listening state

IV. Numerical Results

Based on our proposed system model we simulate the performance of the cluster head by its energy consumption and the probability of a packet being delayed when it arrives during antenna sleep period as shown in Fig. 3. and 4. We use the real data from Mica2 sensor node given in Table I for as the parameters and a 10-Order Markov chain to train the cluster head. The sensor nodes are simulated to generate data from a correlated source with average burst size of 10 packets. We assumed that each sensor node sends a data packet with average size of 100 bytes. The graphs clearly show how the proposed model reduces the energy consumption compared to fixed TDMA sense window scheme. Both scheme shows the same energy consumption at very low and very high arrival rates because in these cases the TDMA sense window is either too small or too large. When the data arrival rate is very high the prediction scheme does not perform very well as it may need a longer code word to train the system more accurately. We can see that this prediction scheme is very suitable for wireless sensor nodes generating correlated data patterns with medium or low rate traffic. It reduces the total energy consumption by the cluster head upto 17% and packet delay probability upto 81% for medium data arrival rates.

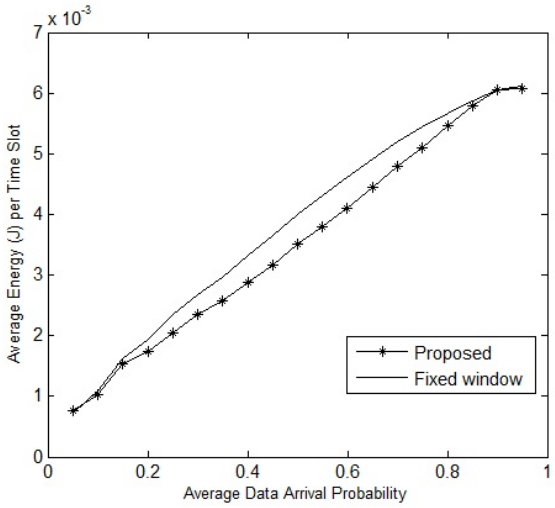


Fig. 3: Cluster head energy consumption per TDMA time slot

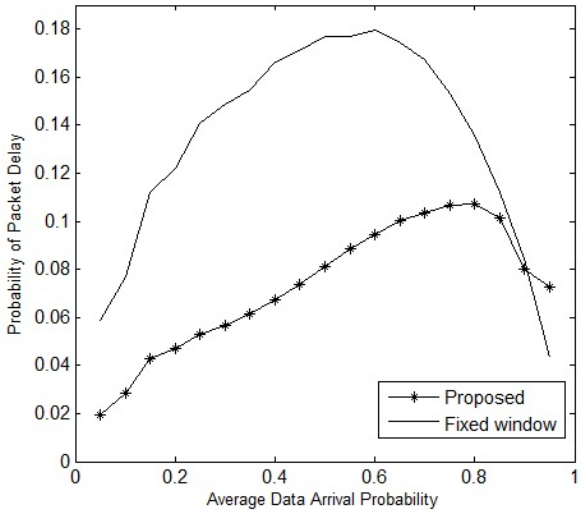


Fig. 4: Probability of packet delay

V. Conclusion

We presented a novel technique for TDMA sense window prediction which can dynamically change the listening window size based on traffic pattern predictions. We proposed a method to train the cluster head to keep track of past data arrival patterns from sensor nodes. From the simulations we showed how our proposed scheme performed better in terms of energy consumption and packet delay probability compared to fixed sensing window scheme which is normally calculated based on average arrival rate from sensor nodes.

References

[1] Calle, M.; Kabara, J., "Measuring Energy Consumption in Wireless Sensor Networks Using GSP," *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on* , pp.1,5, 11-14 Sept. 2006

[2] J. Polastre, J. Hill, D. Culler, "Versatile Low Power Media Access for Wireless Sensor Networks", SenSys1704, Baltimore, Maryland, USA, November 3175, 2004

[3] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci. "A Survey on Sensor Networks", IEEE Communications Magazine, August 2002.

[4] 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, pp.100-105, 2000

[5] 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, pp. 547-551, 3-4 Aug. 2008

[6] S. Cui and K. Ferens, "Energy Efficient Clustering Algorithms for Wireless Sensor Networks," *International Conference on Wireless Networks* , Las Vegas, NV, USA, July 2011

[7] A. A. Markov. "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons, 1971.

[8] A. Raftery and S. Tavare, "Estimation and Modeling Repeated. Patterns in High Order Markov Chains with the Mixture Transition Distribution Model", Journal of Applied Statistics, 43, No. 1, pp. 179-199, 1994

SESSION

MOBILE COMPUTING + MOBILE AGENTS + AD-HOC NETWORKS + MANET

Chair(s)

TBA

Dynamic Cost with Multi-Criteria Decision Method (MCDM) for Load Balancing in Heterogeneous Mobile Networks

Bassem El Zant and Maurice Gagnaire
Institut Mines-Telecom -Telecom ParisTech– LTCI – UMR 5141CNRS
Networks and Computer Sciences Department
23, avenue d'Italie, 75013 Paris, FRANCE

Abstract—*Wireless networks are designed to operate independently without any cooperation. With the many existing inconveniences of 3G networks, operators are persuaded to switch to heterogeneous access networks. This reduces the costs that operators should pay in order to move to 4G and thus make use of the existing networks infrastructures, to come up with heterogeneous environment where multiple wireless technologies coexist. This paper shows how networks users will be able to select their networks through a user-friendly interface where criteria such as the offered throughput and cost are listed. Networks operators should make continuous modifications to the networks listed criteria such as the offered throughput and cost in order to maintain all the available networks working properly. We simulate the intervention using a dynamic cost that the operators can perform on the networks; whereby, the performance parameters, and the operators' profits are taken into consideration along with the users' satisfaction.*

Keywords*:* Multi-criteria decision method (MCDM), Heterogeneous Mobile Networks, Satisfaction-based decision method, Dynamic cost.

1. Introduction

Wireless networks are designed to operate independently. With the explosion of traffic, telecom operators are confronted with the problem of mobile infrastructure saturation. The use of new technologies (Wireless MAN-Advanced which is a development of WiMAX, and LTE-Advanced) enables the operators to integrate different radio technologies already deployed; such as pooling of resources of WiFi, mobile WiMAX, the LTE and HSPA+. Besides, taking into consideration the 3G disadvantages whereby the deployment has proven to be costly, and the higher costs to be paid in order to migrate to 4G networks or to increase the density of the existing 3G networks, operators are increasingly convinced by deploying Heterogeneous Wireless Access Networks (HWAN). In the heterogeneous networks, multiple wireless technologies coexist, and the radio resource management is coordinated. In such networks, mobile users can connect to different radio access technologies. To optimize the system performance, network operators aim to balance loads, as much as possible, in its various radio access networks. In heterogeneous wireless networks, the main challenge is to keep connections among the different networks such as WiFi, WiMax, and WLAN. The 4th generation of wireless (NGWN/4 G) networks is expected to present heterogeneity in terms of wireless technologies and services. The mainly advantage of the mobile networks 3G (UMTS and 1xEV-/ DV) is their global coverage. However, the weaknesses of 3G lie in their bandwidth capacity and operating costs. The WLAN technology such as IEEE 802.11 offers higher bandwidth with low operating costs, although it covers a relatively short range. In addition, technological advances in the evolution of mobile devices made possible the support of different Radio Access Technologies. This raised much interest for integration and interoperability of 3G wireless networks and wireless local networks to take advantage of their respective potentials. In this paper, we start by defining the heterogeneous mobile networks and their benefits before presenting the existing methods used to access network. Then, we present our MCDM with the dynamic cost for load balancing in heterogeneous mobile networks.

In this paper, network selection will be made by the user based on the provided criteria: the throughput and the cost. However to optimize the overall performance, networks operators should intervene. They play on guarantees of QoS (the offered throughput), and economic incentives (the cost) to guide the user's final network selection in order to achieve the best performance in the system. To alleviate networks, the operators offer a lower throughput or a higher cost. However, to attract new arrivals, they provide higher throughput, or lower cost. It is sought to study the result of the intervention and the optimal strategies of intervention of the operators. In this paper, the intervention will be limited by the use of dynamic cost in the system based on a binary logic with two thresholds. Different methods that improve the quality of services and try to provide solution to the above problems have been suggested [1, 2, 12 and 13]. In [1] we can see how the handover technique is used to redirect the mobile user's service network from current network to a new network or one Base Station (BS) to another one or from one Access Point (AP) to another one using the same or different

technologies in order to reduce the processing delay in the overlapping area. Handover network type [11] has horizontal and vertical handover. The homogenous wireless network performs horizontal handover, if there are two BSs using the same access technology. This type of mechanism use signal strength measurements for surrounding BSs to trigger and to perform the handover decision. The authors in [4] have done a comparison among SAW, Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), Grey Relational Analysis (GRA) and Multiplicative Exponent Weighting (MEW) for vertical handoff decision. The vertical handoff decision algorithm for heterogeneous wireless network has been discussed in [3]. The author formulated the problem as Markov decision process. And the vertical handoff decision is formulated as fuzzy multiple attribute decision making (MADM) in [5]. A vertical handoff decision scheme DVHD uses the MADM method to avoid the processing delay in [6], when the goal in [7] of the authors was to reduce the overload and the processing delay in the mobile terminal. A novel vertical handoff decision scheme to avoid the processing delay and power consumption has been proposed in this paper. A novel distributed vertical handoff decision scheme using the SAW method with a distributed manner to avoid the drawbacks has been suggested in [8] when the authors in [9] define using the emerging IEEE 802.21 standard a Media Independent Handover (MIH) functions as transport service in order to offer a vertical handoff decision with a minimum of processing delay. A four-step integrated strategy for MADM-based network selection has been proposed in [10]. In this paper, the strategies are classified into five main categories: function-based, user-centric, multiple attribute decision, fuzzy Logic and Neural Networks based, and context-aware strategies. The paper is organized as follows. Section2 discusses the network selection process, while section 3 presents the selection of network architecture. Section 4 is dedicated to the new Multi criteria decision method MCDM, and 5 for the key performance indicator KPI. Section 6 presents the operator strategies, and section 7 is dedicated to simulation description. Section 8 discusses the results, and section 9 presents our conclusion and the future work.

2. Network selection process

Before opening a new session or making a handover, the mobile must evaluate different alternatives and select the best network and the best class of service. For this multi-criteria decision, we consider the QoS requirements, radio conditions, the cost and the user's preferences. The different decisions attributes are as following: the minimum guaranteed throughput (d_{min}), the maximum throughput (d_{max}) and cost (C). Figure 1 shows the hierarchical representation of the criteria. During the first stage, the general criteria must take into consideration the user's preferences. The user may prefer paying more for a better quality of service or he prefers to save money and get low quality. The user therefore assigns weights to QoS incentives and the cost according to his preference. However the secondary criteria depend on the type of application, for example the minimum throughput is critical for a Constant Bit Rate (CBR) application when this did not make sense for a best effort application. As following three distinct types of applications:

1. Inelastic applications with constant throughput. Therefore, the maximum throughput will not be considered and has no importance so its weight will be null. d_{min} and C will be taken into consideration.
2. Streaming applications: used for real time services with variable throughput (ex: Video service using Mpeg4). Three parameters will be considered for those applications: d_{max}, d_{min} and C.
3. Elastic applications: used for data transfer services such as the transfer of files, email, and web traffic services. For these applications just d_{max} and C will be considered, when d_{min} will be ignored because these applications do not require guarantees in QoS.

To evaluate the different alternatives of a session and select the best choice, a Multi-Criteria Decision Method (MCDM) will be adopted. Indeed, it defines an utility function that depends on the standard weights of various criteria. Several utility functions have been proposed, the Simple Additive Weighting (SAW) and the Multiplicative Exponent Weighting (MEW), but these methods do not consider the current needs of the session. For example, when a user using CBR application is willing to pay, the best alternative will be severed without taking into account the throughput required for the session, and thus an overqualified alternative will be severed, while there is another less expensive alternative meeting the requirements of the session.

3. Selection of network architecture

We suggest a hybrid approach taking into account the preferences of both, operator and user. Indeed, the policies of the operator are implicitly integrated in the system information. This information will be sent to all mobile terminals. The mobile terminal will decode the information, assess different options and then choose the best network. The operator offers three classes of services (Premium, Regular, and Economic). Financial and QoS considerations will determine the selection of the best alternative. An alternative is defined by a combination of available access technology (WiMax, WiFi, 3G, etc.) and classes of services such as Premium, Regular, and Economic. When a new session will be opened or where a hand over is made, the mobile terminal will receive information from the system which includes the monetary cost and the QoS incentives, decodes the incentives of the operator, evaluates the different alternatives and then chooses the best network with the service class.

The monetary cost: we suggest a cost according to the QoS. As there are different classes of services (Premium, Regular, and Economic) in different networks, the network operators provide different throughputs for each. Indeed, a Premium

session will be more priority, will get a better quality of service and will be more expensive than the regular and the economic. Mobile's users will be charged according to their priority (class of service). The prices will change dynamically in real time according to the conditions of the networks, the radio resources are used efficiently, and the performance of the system will be improved. As it's more complex to implement a dynamic price, we consider a static price with the different classes of service which will be fixed and does not change with the network load conditions. The flat pricing strategies are not used because they will cause a waste of resources and will force light users to subsidize heavy users and prevent the deployment of internet service quality. In addition, the use of flat price will result in the congestion of a computer network and will degrade the performance of the system .Accordingly, as we will provide a guaranteed QoS, we propose a model based on the used volume, and then the sessions will be billed according to the amount of traffic that they transmit. The price of traffic per unit will depend on the radio resource access and the class service. In conclusion, the monetary cost is for each unit of traffic and in our case is defined per Kbyte.

QoS incentives: different QoS parameters should be considered based on the application's requests. They specify the minimum number (n_{min}) and the maximum (n_{max}) units of radio resource localized for a session. These sub-parameters depend on radio resources and the service class. The total traffic for a specific RAT is hidden. Different n_{min} and n_{max} are generated for the different classes of services reflecting the strategies of the operator. These sub-parameters don't necessarily reflect the conditions of the network but rather the operator's wishes to serve the different sessions of different classes. Based on the signal noise ratio(SNR) report, the mobile terminals will adopt the modulation type and the forward error correction (FEC) for the encoding. This is why the number of bits per radio resource unit(RRU) and the minimum and maximum throughputs depend on radio's conditions. Indeed, during the evaluation of different alternatives, the mobile terminal combines its radio conditions (which differed from rat to another) with the reported QoS Sub-parameters then determines the minimum and maximum expected throughput. E.g. for OFDM-based technologies the minimum throughput expected will be:

$$d_{min}=\frac{n_{min}\times N_u\times K\times R_C}{SI} \quad (1)$$

With n_{min}: the minimum guaranteed of OFDM reserved symbol, N_u: the number of carrier used for data transmission, K: the number of bits per symbol module that vary with the modulation, Rc: FEC report, and SI: scheduling interval.

To make it more simple and homogeneous the QoS incentives for different radio technologies, we will express the Qos sub-parameters in terms of minimum guaranteed throughput d_{min} and the maximum throughput d_{max} (instead of n_{min} and n_{max}). To evaluate the different alternatives, the mobile terminal will determine the expected throughput that represent the result of multiplication of minimum throughput (resp. maximum) of the class of service in the alternative with the gain of modulation g_M and g_C coding gain.

$$d^{userk}_{\min oumax}= d^{\text{service classe i}}_{\min oumax} * g_M * g_C \quad (2)$$

4. New Multi-Criteria Decision Method

The new method presented in this paper is a Multi-Criteria Decision Method (MCDM). It defines the alternative as a combination between a network and one of the classes of service Premium, Regular or Economic. The alternatives will be evaluated according to their monetary cost, minimum throughput that they guarantee and the maximum throughput they offer based on the user's satisfaction. We define then a function of satisfaction for each type of session (inelastic, streaming and elastic) and user profile. The HWAN (or NGWAN) allow the efficient use of available radio resources. They can serve more customers and this will generate more profit. Mobile users can connect simultaneously or not to the different access technologies that meet their needs in terms of QoS or cost.

As cited before, our method is a hybrid method shared between network and users. We define an environment that integrates the operator's objectives and the user's preferences based on the user's satisfactions. Our method selects the best alternative based on the expected user satisfaction. The utility function is defined as the weighted sum of partial satisfaction functions. The function of partial satisfaction ($s_{c,\,p}$) depends on the decision criterion (c) and the profile of the user (p). There are two types of users: those who are willing to pay for best performance and those who prefer to save. As mentioned before, there are three different types of applications, so we will have six users' profiles.

The function of satisfaction expected S (a_i) for the alternative a_i is given by:

$$S(a_i) = w_{dmin,\,p} * s_{dmin,\,p} + w_{dmax,\,p} * s_{dmax,\,p} + w_{cost,\,p} * s_{cost,\,p} \quad (3)$$

Where ($w_{dmin,\,p}$, $w_{dmax,\,p}$, $w_{cost,\,p}$) represent the static weight vector of profile p, and ($s_{dmin,\,p}$, $s_{dmax,\,p}$, $s_{cost,\,p}$) represent the functions of partial satisfaction of profile p.

The function of satisfaction of throughputs depends on the QoS needs of the session. Inelastic applications are characterized by a fixed throughput, R_f. The QoS requirements for these applications are strict and inflexible and therefore the function of satisfaction of the minimum throughput ensures is defined by the following formula (see figure 1).

$$S_{d_{min,p}} = \begin{cases} 0 & si\ d_{min}(a_i) < R_f \\ 1 & si\ d_{min}(a_i) \geq R_f \end{cases} \quad (4)$$

Figure 1: Form of the function of satisfaction for an inelastic session

Where d_{min} (a_i) represents the minimum rate guaranteed by the alternative a_i. In this case the function of maximum satisfaction will not be considered.
Streaming sessions require a minimum rate but also a maximum throughput as it is real-time applications. Their function of satisfaction is in the form of sigmoid and is defined the following formula (see figure 2).

$$S_{d_{min,p}} = 1\text{-}\exp\frac{-\alpha(\frac{d_{min}(a_i)}{R_{av}})^2}{\beta+\frac{d_{min}(a_i)}{R_{av}}} \quad (5)$$

$$S_{d_{max,p}} = 1\text{-}\exp\frac{-\alpha(\frac{d_{max}(a_i)}{R_{av}})^2}{\beta+\frac{d_{max}(a_i)}{R_{av}}} \quad (6)$$

(5) and (6) present the function of satisfaction for the streaming sessions.
Where α, β are positive constants that determine the shape of the sigmoid, $d_{max}(a_i)$ and $d_{min}(a_i)$ are the maximum and minimum throughput guaranteed by the alternative a_i and average throughput R_{av}.
For elastic sessions, the function of satisfaction is a concave function defined by the formula given in figure 2.

$$S_{d_{max,p}} = 1\text{-}\exp -\frac{d_{max}(a_i)}{R_c} \quad (7)$$

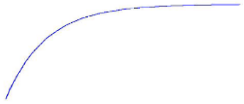


Figure 2: Form of the function of satisfaction of the elastic sessions

Where R_c is the throughput of comfort.
The satisfaction's function of the monetary cost depends on the user's tolerance, it is modeled by the following formula (see figure 3).

$$S_{cost,p} = \exp -\frac{cost(a_i)^2}{\lambda_p} \quad (8)$$

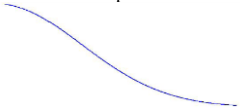


Figure 3: Form of the function of satisfaction of the monetary cost

Where cost (ai) is the monetary cost of the alternative a_i and λ_p a positive constant that depends on the profile p. More the user is tolerated in terms of cost, more λ_p is greater.

5. Key Performance Indicators (KPIs)

Different KPIs are defined to assess the user satisfaction and operators gain. To assess the user's satisfaction, we take into consideration the applications' types. So, for the streaming and inelastic applications, we consider the throughput, the delay and the drop probability. And for the elastic applications, we consider the throughput and performance test d/ d_c where d is the effective throughput received per the user and d_c represents the throughput of comfort. It then sets the user's satisfaction by:

$$S = w_{QoS}*S_{QoS} + w_{cost}*S_{cost} \quad (9)$$

Where w_{QoS} and w_{cost} represent the respective weights assigned to the QoS criteria and monetary cost, and S_{QoS} and S_{cost} the respective user's satisfaction. The functions of satisfaction are identical to those defined for the calculation of the best alternative and satisfaction will be measured at the actual rate received by the user. To assess the operator's gain we calculate the average revenue per user for consumption per Kbyte.

6. Operator strategies

The operator strategies are implicitly embedded in the system's information. This will influence the decision process. The operator will inform the user of the minimum and maximum throughput of each network access with the corresponding monetary cost and the user is who choose the most suitable access technologies among different ones. We can say that the operator will affect and guide the user when making the final decision. In our work, the QoS incentives will be fixed when the cost will vary dynamically to optimize short and long term goals. In both cases the monetary cost and the different incentives for QoS will reflect the operator strategies and contribute in the final user's decision. The operator will hide the total capacity of the system and will only provide to the users information about the incentives of QoS and presents the different guaranteed throughputs offered with the cost. On the other hand, when the operator is ready to reserve, in a RAT j, a band for the session of the service class i, excellent throughput will be suggested, and then the class i will attract new coming sessions to the radio access technology RAT j. So to avoid new sessions, the operator can offer excellent throughput or low cost which will push the user to pass to one RAT. In conclusion, the dynamic incentives of QoS variation or cost variation will allow the operator to more or less attract new users to a class in a specific RAT and then the operator will contribute in the final user's decision.

7. Simulation

Before presenting the results obtained per our method, it's necessary to describe the simulation environment, the discrete events system, and the Dynamic Cost.

7.1 Simulation Environment

In our simulation, we first consider three similar Radio Access Technologies (R1, R2, and R3), which mean they all have same service classes, same Cost and QoS incentives for all classes: Premium, Regular and Economy. The goal is to give transparent offer to the user who is not interested in the technology but in the offer in terms of Cost and QoS incentives. Therefore, it is possible that there is no load balancing:

- In the worst case, R1 can be filled first, then R2 and R3.Therefore the operator's intervention is necessary for load balancing in the network.
- For battery consumption or security reasons, users may prefer specific network.

7.2 Discrete Events System

A discrete event system is a system described by discrete state variables, i.e. changes occur on the occurrence of a set of states. We have different types of possible events during the lifetime of the system, thus we must describe the operating logic between events (determine state changes for each event and the events that result). In our system, we define three main events: session arrival (A), session departure (D) and the end of a frame (FT) see figure 4.

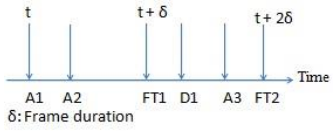


Figure 4: Discrete Events System

If the event is an arrival we must expect the departure, mark the session as active and expect the new arrival. If it's a departure, we must free the resources and mark the session as terminated, and if it is an end of frame, there must be an allocation of resources. Matlab has been used to make the simulation, and the intervention of the operator under different strategies has been programmed. As following the different approach simulated: The operator can intervene either by adjusting the cost or the QoS incentives. The intervention by adjusting the cost using a binary logic will be examined.

7.3 Dynamic Cost: Binary logic

For each network, we defined two thresholds S1 and S_2 with S_1= 0.5*TotalCapacity and S_2= 0.8*TotalCapacity where Total Capacity is the full capacity of a network. When dminTotal represents the network charge (i.e the total amount of reserved resources for d_{min}) exceeds the first threshold in a network, the costs that were initially a,b,c for Premium, Regular and Economic offers respectively become a+x, b+x and c+x. Also, when dminTotal exceeds the second threshold the costs become a+y, b+y and c+y. So when the first network's costs increase for the first time, the second network will be selected automatically by new sessions. Then when the second network's costs increase, the third network will be selected automatically by new sessions until it increases its own costs. We have exactly the same approach for the second threshold. No operator's intervention for low charge (dminTotal<S_1) nor for high charge (dminTotal>S_2).

$$\text{Cost}= \begin{cases} C_{init} & \text{if dminTotal}<S_1 \\ C_{init}+x & \text{if S1<dminTotal<S2} \\ C_{init}+y & \text{if dminTotal>S} \end{cases} \quad (10)$$

8. Results

In this section, we present the results of our simulation in which we compare two systems: with and without intervention (green and blue line respectively) of the operator. N_{max} is the total number of sessions in the system.

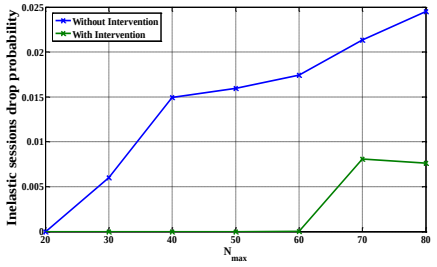


Figure 5: Inelastic session drop probability

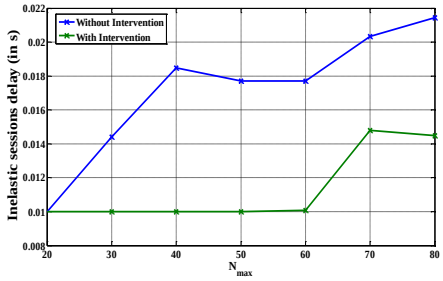


Figure 6: Inelastic session delay

We notice that the intervention using the dynamic cost will offer better performances for both inelastic and partially elastic sessions. These performances are shown by the delay (Figure 6 and 8) and by the drop probability (Figure 5 and 7). The fact that different networks are filled in a continuous way with intervention of the operator will give sessions a better load balancing of charges between different networks. Therefore, sessions will have a higher throughput and thus a lower delay and drop probability.

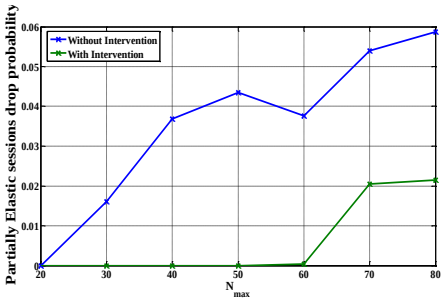


Figure 7: Partially elastic session drop probability

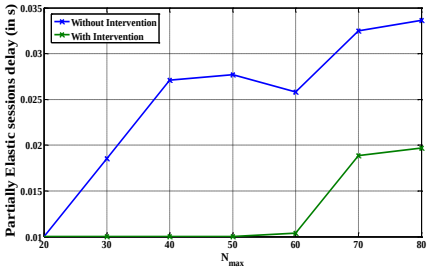


Figure 8: Partially elastic session delay

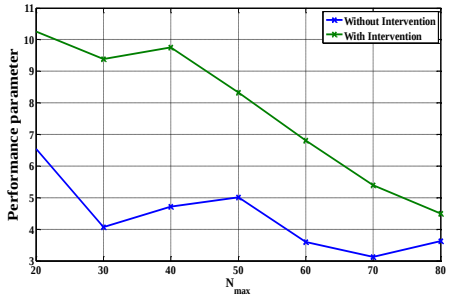


Figure 9: Performance parameter for elastic session

Same results are shown for the performance parameter of elastic sessions with best performance with intervention of the operator using dynamic cost (Figure 9).

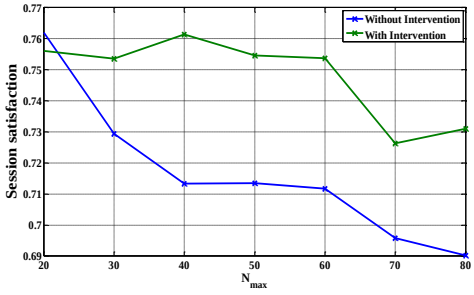


Figure 10: Session satisfaction

Figure 10 shows that the session satisfaction is higher with intervention using the dynamic cost.

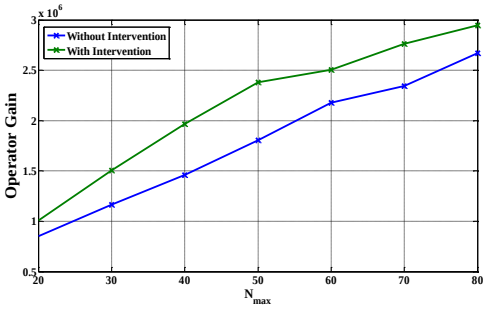


Figure 11: Operator gain

Figure 11 shows that the operator's gain is higher with the operator intervention because the operator will increase the costs as a strategy to achieve a better load balancing in the system.

9. Conclusion and Future Work

This paper presents a new Multi-Criteria Decision Method (MCDM) with dynamic cost for balancing loads in heterogeneous mobile networks. Matlab has been used to make the simulation of our method and the different existing scenarios. The intervention of the operator under different strategies has been programmed. We notice a big number of improvements in the whole system due to the operator's intervention. This intervention guides the users' final decision of the RAT. Thus, the users will be divided between all operators' RAT which give the operator a better channel utilization. The operators will have a higher gain using the dynamic cost intervention strategy in the system. On the other hand, this division between RAT reduces the number of users in each RAT, so users will find a better performance in different RAT due to available throughput used by operators to improve the system performance such as throughput, delay, drop probability, performance parameter and the sessions satisfaction. As a future work, a comparison between different intervention strategies will be done to get the best performances. We will study this intervention using a fuzzy logic for the cost and the intervention by adjusting QoS incentives using binary and fuzzy logic will be done. On the other hand, these methods could be implemented with SON (Self Organization Networks), which is a new technology presented in the LTE standard for auto-configuration, auto-optimization and auto-exploitation of cellular networks equipment for mobile telephony.

10. References

[1] Tripathi, N.D.J., Reed, H..VanLandinoham, H.F.: Handoff In Cellular System. IEEE Personal Communications, 49, 2276--2285 (2000)

[2] Gustafsson E., Jonsson, A., Research, E.: Always Best Connected. IEEE Wireless Communications,10, No. 1, 49--55 (2003)

[3] Steven-Navarro, E., Wong, V.W.S., Lin, Y.: A Vertical Handoff Decision Algorithm For Heterogeneous Wireless Networks. Wireless Communications and Networking Conference, IEEE, Kowloon 2007, pp. 3199--3204.

[4] steven-Navarro, E., Wong, V.W.S.: Comparison between vertical handoff decision algorithms for heterogeneous wireless network.Vehicular Technology Conference, IEEE 63rd ,Melbourne, Vic. , 947--951 (2006)

[5] Zhang, W.: Handover Decision Using Fuzzy MADM In Heterogeneous Networks. Wireless Communications and Networking Conference, IEEE, Vol.2, 653-658 (2004)

[6] Tawil, R., Pujolle, G., Salazar, O.: A Vertical Handoff Decision Schemes In Heterogeneous Wireless Systems. Vehicular Technology Conference,VTC Spring 2008IEEE, Singapore, 2626--2630 (2008)

[7] Tawil, R., Demerjain, J., Pujolle, G., Salazar, O.: Processing-Delay Reduction During The Vertical Handoff Decision In Heterogeneous Wireless System. International Conference on Computer Systems and Applications, AICCSA IEEE/ACS,381--385 (2008)

[8] Savitha, K., Chandrasekar, C.: Vertical Handover decision schemes using SAW and WPM for Network selection in Heterogeneous Wireless Networks. Global Journal of Computer Science and TechnologyVolume 11 Issue 9 Version 1.0 May (2011)

[9] Tawil, R., Pujolle, G., Salazar, O.: Vertical Handoff Decision Schemes For The Next Generation Wireless Networks. Wireless Communications and Networking Conference, 2789--2792 (2008)

[10] Tawil, R., Pujolle, G., Demerjain, J.: Distributed Handoff Decision Scheme Using MIH Function For The Fourth Generation Wireless Networks, 3rd International Conference on Information and Communication Technologies: From Theory to Applications, 1--6 (2008)

[11] Wang, L., Binet, D.: MADM- Based Network Selection In Heterogeneous Wireless Networks: A Simulation Study. 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology, 559--564 (2009)

[12] Nasser, N., Hasswa, A., Hassanein, H.: Handoff In Fourth Generation Heterogeneous Networks. Communications Magazine, IEEE , 44, 96--103 (2006)

[13] Martinez- Morales, J.D., rico, V.P., Steven, E.: Performance comparison between MADM algorithms for vertical handoff in 4G networks, 7th International Conference on Electrical Engineering Computing Science and Automatic Control (CCE), 309—314, (2010)

The iTrust Local Reputation System for Mobile Ad-Hoc Networks

Wei Dai, L. E. Moser, P. M. Melliar-Smith, I. Michel Lombera, Y. T. Chuang
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106
weidai@umail.ucsb.edu, {pmms, moser, imichel, ytchuang}@ece.ucsb.edu

***Abstract*—The iTrust search and retrieval network serves as a trustworthy medium for the distribution of information, that addresses the problems of censorship and filtering of information. To combat subversive behavior of nodes that might undermine the trustworthiness of iTrust, a reputation system is needed. The iTrust reputation system presented in this paper detects and blacklists malicious nodes. It minimizes the expectation of cooperation between nodes through local reputations based solely on direct observations of the nodes. Simulation results demonstrate that local neighborhoods provide better malicious node detection and blacklisting than does the entire network, which is particularly appropriate for mobile ad-hoc networks.**

Keywords—search and retrieval; mobile ad-hoc network; peer-to-peer network; reputation management; iTrust

I. Introduction

Mobile ad-hoc networks (MANETs) are intrinsically dependent on cooperation and collaboration among nodes. MANETs do not rely on a static network infrastructure, but they do rely on several assumptions [18]. Due to the lack of infrastructure and other limiting factors, such as transmission range, the nodes in a MANET develop symbiotic relationships. Such relationships assume that all of the nodes are equally trustworthy and have the same objectives. Such assumptions about the nodes are not appropriate for the iTrust search and retrieval network [1], [14], [15], which aims to ensure freedom from censorship and filtering of information, even in the presence of malicious nodes.

A MANET requires cooperation among the nodes in the network to function properly. Without the fulfillment of this requirement, packets would not be forwarded, routes would not be established, and the network would not function properly. Despite the importance of cooperation among the nodes in a MANET, it is not guaranteed. Consequently, a reputation system is needed. However, the addition of a reputation system, in which reports of misbehavior are collected and redistributed, treads dangerously close to encroaching on the fundamental principle of iTrust, which is to provide a distributed, uncensorable, reliable, trustworthy system with no central authority.

The iTrust reputation system for MANETs, presented in this paper, is based on local reputations and neighborhoods, and uses direct observations of the nodes to detect malicious neighbors, with as few interactions between the nodes as possible. It avoids reliance on information from other nodes, while maintaining a method of detecting the misbehavior of malicious nodes. The iTrust reputation system is designed specifically for iTrust operating over MANETs.

In designing the reputation system for iTrust MANETs, we investigated the merits of utilizing a local neighborhood for each node. Simulation results provide increased insight into the rationale behind using local neighborhoods for iTrust. They reveal a distinct relationship between neighborhood size and the number of transmissions required to detect malicious nodes. Essentially, with smaller numbers of transmissions, local neighborhoods consistently yield a higher proportion of malicious nodes detected and blacklisted, compared to the entire network with more transmissions. This finding is particularly important for MANETs, as it is necessary to eliminate malicious nodes as quickly as possible with as few interactions and transmissions as possible, in order to reduce the costs associated with a reputation system.

The rest of this paper is organized as follows. Section 2 presents an overview of the iTrust search and retrieval network. Section 3 describes the iTrust local reputation system, and the details of its three modules. Section 4 provides an evaluation of the iTrust reputation system, and insight into the use of neighborhoods. Section 5 discusses other reputation systems, and their relationship to the iTrust reputation system. Section 6 concludes the paper and presents future work.

II. The iTrust Search and Retrieval Network

The iTrust search and retrieval network [1], [14], [15] addresses potential problems with centralized search and retrieval systems that are subject to censorship, filtering, and suppression of information. Moreover, the iTrust network is intended to be robust against malicious nodes. To achieve these objectives, the iTrust system adopts a probabilistic, distributed, and decentralized approach.

The nodes that participate in an iTrust network are referred to as the *participating nodes* (Figure 1). Some of the participating nodes, the *source nodes*, produce information, and make that information available to other participating nodes (Figure 2). The source nodes also produce metadata that describes their information, and distribute the metadata, along with the address of the information, to randomly chosen nodes in the iTrust network. Other participating nodes, the *requesting nodes*, request and retrieve information. The requesting nodes generate requests (queries) that contain keywords, and distribute their requests to randomly chosen nodes in the iTrust network (Figure 3). Nodes that receive a request compare the keywords in the request with the metadata they hold. If a node finds a match, which we call an *encounter*, the matching node returns the address of the associated information to the requesting node (Figure 4). The requesting node then uses the address to retrieve the information from the source node. A *match* between the keywords in a request received by a node and the metadata held by a node can be an exact match or a partial match, or can correspond to synonyms.

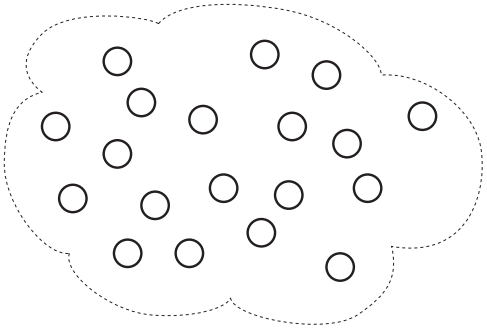


Fig. 1. An iTrust network with participating nodes.

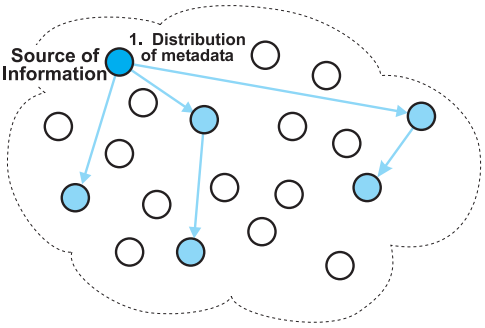


Fig. 2. A source node distributes metadata, describing its information, to randomly chosen nodes in the network.

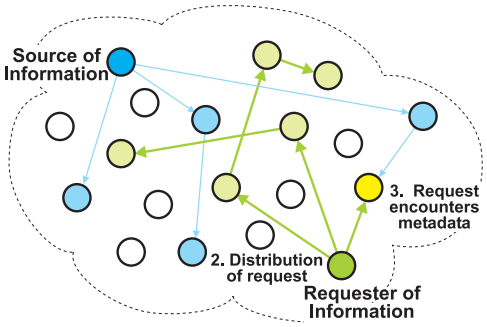


Fig. 3. A requesting node distributes its request to randomly chosen nodes in the network. One of the nodes has both the metadata and the request and, thus, an encounter occurs.

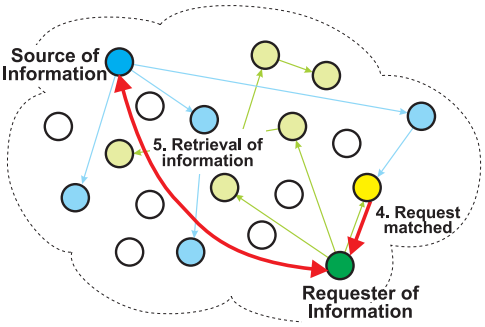


Fig. 4. A node matches the metadata and the request and reports the match to the requesting node. The requesting node then retrieves the information from the source node.

The iTrust search and retrieval system is based on the hypergeometric distribution [8], which is given in terms of the following variables:

- n: The number of participating nodes
- x: The proportion of the n participating nodes that are operational, *i.e.*, $1 - x$ is the proportion of non-operational or malicious nodes
- m: The number of participating nodes to which the metadata are distributed
- r: The number of participating nodes to which the requests are distributed
- k: The number of participating nodes that report matches to a requesting node.

In iTrust, the probability $P(k \geq 1)$ that a request yields one or more matches is given by:

$$P(k \geq 1) = 1 - \frac{n - mx}{n} \frac{n - 1 - mx}{n - 1} \cdots \frac{n - r + 1 - mx}{n - r + 1} \quad (1)$$

for $n \geq mx + r$. If $mx + r > n$, then $P(k \geq 1) = 1$. In [14], we showed that, if $m = r = 2\lceil\sqrt{n}\rceil$, then the probability that a request yields one or more matches is $P(k \geq 1) \geq 1 - e^{-4} \sim 0.9817$. We use this result and Equation (1) in our evaluation of the iTrust reputation system given in Section IV.

III. The iTrust Local Reputation System

The iTrust local reputation system for MANETs monitors packet forwarding, and watches for non-operational nodes and nodes that do not respond to requests (queries). A local reputation system reduces overheads and the dependence among nodes. It also reduces the amount of storage required, because only information about one-hop neighboring nodes needs to be recorded. In contrast, a global reputation system would result in higher overheads, and also a higher expectation of cooperation among nodes [1].

The iTrust reputation system is based on a *local neighborhood* of each node, consisting of the nodes within one hop of the node, and a neighborhood watch mechanism that monitors the interactions of the neighboring nodes. The iTrust reputation system maintains *reputation ratings* of the nodes. A node uses only direct observations to update the reputation ratings of its neighboring nodes. Consequently, the reputation ratings of different nodes might not be consistent. This design choice limits the expectation of cooperation among nodes, thus reducing the opportunities for malicious behavior.

The two primary types of malicious behavior that the iTrust reputation system addresses are:

- A node does not send responses to requests when it has a match.
- A node sends requests and responds to requests, but does not forward messages.

Thus, the iTrust reputation system primarily serves to ensure that nodes send messages as expected; it does not address other threats such as Sybil attacks.

In the extreme case in which a node becomes isolated due to the lack of any well-behaved neighbors, the node needs to move to another location where well-behaved nodes are present.

The two main principles under which the iTrust reputation system operates are:

- Intermittent behavior is not punished as much or as rapidly as consistently bad behavior, because intermittent bad behavior is more difficult to detect.

- Efforts are directed towards observing the behavior of nodes within one hop. Malicious behavior that occurs beyond that range is the responsibility of other nodes.

Each node in the MANET maintains a *local reputation table* that consists of a list of nodes within its local neighborhood. Whenever an interaction with another node occurs, the node increases or decreases the reputation rating of that other node. This mechanism addresses malicious behavior.

The iTrust reputation system consists of three modules that interact with each other. These three modules are the Neighborhood Module, the Reputation Rating Module, and the Monitoring Module, which are illustrated in Figure 5 and are described below.

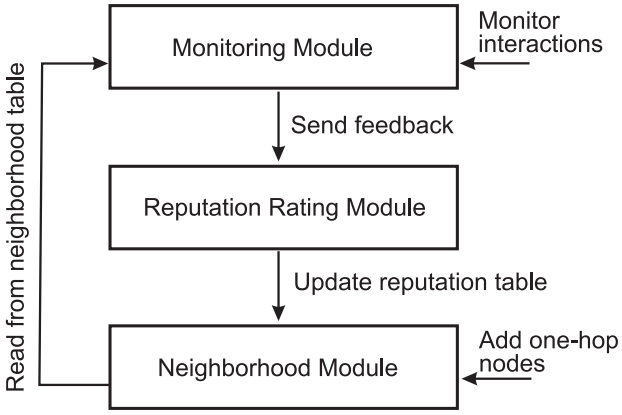


Fig. 5. The three modules of the iTrust reputation system and their interactions.

A. Neighborhood Module

The Neighborhood Module at a node maintains the local neighborhood of the node and the reputation table for the neighborhood. All of the nodes within one hop of the node, together with their reputation ratings, are represented in the reputation table. Each time a new node is within one hop of the node, the Neighborhood Module adds an entry for the node to the reputation table. A new node starts with a neutral reputation rating of zero. The reputation rating of a node depends on positive and negative interactions with neighboring nodes, as determined by the Reputation Rating Module.

B. Reputation Rating Module

The Reputation Rating Module at a node performs the calculations required to update the node's reputation table. It relies heavily on the Monitoring Module to supply feedback, so that it can decide whether to increase or decrease a node's reputation rating. The Reputation Rating Module is responsible for blacklisting and graylisting nodes.

Blacklisting involves recording malicious nodes in the reputation table. Whenever the reputation rating of a node falls below a certain threshold, the node is blacklisted. A blacklisted node is effectively permanently removed from the neighborhood and from the reputation table maintained by the Neighborhood Module. As a precautionary measure, there is the option of graylisting a node.

Graylisting is a second-chance mechanism that provides a modicum of leniency in an otherwise unforgiving system. Essentially, a node on the graylist is given a second chance before it is blacklisted. A node on the graylist functions as a normal node with the ability to send and receive messages and, to a certain extent, redeems itself through "good" behavior. If a node is graylisted twice, it is put on the blacklist.

C. Monitoring Module

The Monitoring Module at a node provides first-hand observations of the behaviors of the nodes within the node's neighborhood. The Monitoring Module provides feedback to the Reputation Rating Module about the good and bad behaviors of neighboring nodes.

For an iTrust MANET, nodes are expected to distribute messages to nodes in the network. Whenever a node interacts with another node by sending a request or a response, it listens to the node's transmissions to check that it is sending messages appropriately. If a node appears to be unresponsive or forwards messages improperly, the Monitoring Module provides feedback to the Reputation Rating Module regarding the negative behavior of the node. The Reputation Rating Module then decreases the reputation rating of the node accordingly.

When a node joins the network, it is given the neutral reputation rating of 0. The reputation rating of a node never exceeds 0, which is done to prevent a malicious node from building a positive reputation rating over time and then committing a series of malicious acts.

If a node exhibits malicious behavior, the Reputation Rating Module decreases the reputation rating of the offending node by -2. If a node exhibits good behavior, the Reputation Rating Module increases the reputation rating of the node by +1. Decreasing a maliciously behaving node's reputation by -2 allows the system to reward good behavior by +1, thus preventing a node from entering cycles of bad and good behavior to dupe a neutral reputation. We selected -2 and +1, but other values of the reputation rating may be used, as long as the ratio of punishments to rewards is 2:1. The difference in the values of ratings attributed to bad and good behavior allows the system to implement the second-chance mechanism described below.

Depending on the level of strictness desired, the user can place a threshold of -2 or -4 on blacklisting a node. With the threshold of -2, a node's first offense results in its being blacklisted. With the threshold of -4, a node is allowed two offenses before it is blacklisted; in this case, a node's first offense results in its being graylisted. These thresholds provide for immediate blacklisting and a second-chance mechanism using graylisting before the node is blacklisted.

With blacklisting for one offense (-2 threshold), a maliciously behaving node is allowed no leniency. If the node behaves maliciously, the reputation system at a node immediately adds the node to the blacklist and permanently bans the node from the node's local neighborhood. In this case, the node's reputation rating is 0 or -2. If the reputation rating is 0, the node sends metadata, requests and responses as usual. If the node's reputation rating falls to -2, the node is blacklisted and is effectively permanently removed from the network.

With blacklisting for two offenses (-4 threshold), the reputation system does not blacklist the node immediately; rather, it decreases the reputation rating of the offending node, and places the node on the graylist. If the node behaves maliciously a second time, the system then places the node on the blacklist, and permanently bans the node from the node's local neighborhood. In this case, the node's reputation rating is 0, -1, -2, -3 or -4. If the node's reputation rating is 0 or -1,

the node sends metadata, requests and responses as usual. If the node's reputation rating falls to -2, the node is graylisted (or blacklisted if it was previously graylisted). If the node's reputation rating falls to -3 or -4, the node is blacklisted.

In our evaluation of the iTrust reputation system, we investigate the performance of the system with blacklisting for one offense and blacklisting for two offenses.

IV. Evaluation

To evaluate the iTrust reputation system, we performed a simulation. In the simulation, we assume that good behavior can be distinguished from malicious behavior. The experimental setup comprises a network of 1000 nodes, where each node has a neighborhood of 150 nodes. We investigate the advantages and disadvantages of a 150 node neighborhood vs. a 1000 node network, in maintaining the reputations of the nodes and in detecting and blacklisting malicious nodes.

As a baseline, first we investigate the behavior of iTrust in finding a match, for the 150 node neighborhood and the 1000 node network. We let m be the number of nodes to which the metadata are distributed, and r be the number of nodes to which the requests are distributed, in the 150 node neighborhood. Similarly, we let M be the number of nodes to which the metadata are distributed, and R be the number of nodes to which the requests are distributed, in the 1000 node network. We investigate the match probability $P(k \geq 1)$ for several values of m and r in the 150 node neighborhood and several values of M and R in the 1000 node network.

A. Results without Malicious Nodes

In the first experiment, we set the number M of nodes to which a node in the 1000 node network distributes the metadata to $M = 64$ nodes. Thus, on average, only $m = 9 \sim (64/1000) \times 150$ nodes in the 150 node neighborhood receive the metadata. We set the number R of nodes to which a node in the 1000 node network distributes its requests to $R = 64$ nodes. Likewise, we set the number of nodes to which a node in the 150 node neighborhood distributes its requests to $r = 64$ nodes in the 150 node neighborhood. As shown in Figure 6, the match probability $P(k \geq 1)$ is slightly higher for the 150 node neighborhood than for the 1000 node network.

We then performed two more experiments in which we change the value of r in the 150 node neighborhood. In both experiments, we retain $m = 9$ in the neighborhood, because $m = 9$ represents the proportion of nodes in the 150 node neighborhood that receive the metadata distributed by a node in the 1000 node network.

Thus, in the second experiment, we set $r = 9$ for the 150 node neighborhood, where the value of m is still $m = 9$. As shown in Figure 7, the match probabilities for the 150 node neighborhood are significantly worse than the match probabilities for the 1000 node network with $M = 64$, $R = 64$. We conclude that, to utilize a smaller local neighborhood, it is necessary to increase the number r of nodes in the local neighborhood to which a request is distributed.

Consequently, in the third experiment, we select a value of r for the 150 node neighborhood between 9 and 64. In particular, we choose $r = 24 \sim 2\sqrt{n}$, where $n = 150$, the number of nodes in the local neighborhood. The value of m is still $m = 9$. In Figure 8, it can be seen that the 1000 node network slightly outperforms the smaller 150 node neighborhood, but $r = 24$ is an improvement over $r = 9$ for the 150 node neighborhood. However, there still might be reason to choose the smaller local neighborhood for detecting and blacklisting malicious nodes, which we investigate below.

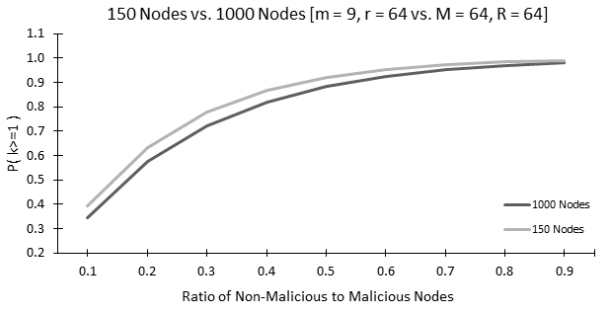


Fig. 6. Probability of one or more matches for the 150 node neighborhood with $m = 9$, $r = 64$ vs. probability of one or more matches for the 1000 node network with $M = 64$, $R = 64$.

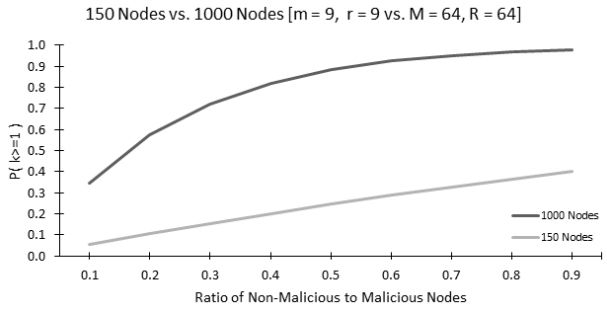


Fig. 7. Probability of one or more matches for the 150 node local neighborhood with $m = 9$, $r = 9$ vs. probability of one or more matches for the 1000 node network with $M = 64$, $R = 64$.

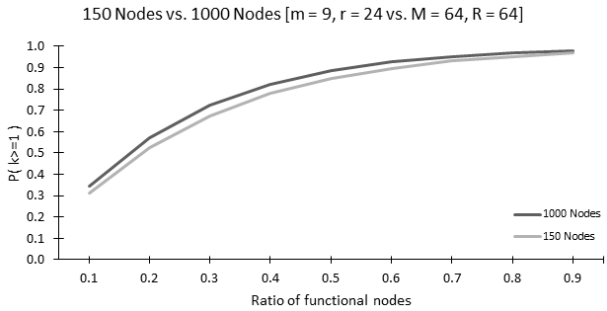


Fig. 8. Probability of one or more matches for the 150 node neighborhood with $m = 9$, $r = 24$ vs. probability of one or more matches for the 1000 node network with $M = 64$, $R = 64$.

B. Results with Malicious Nodes

We now consider malicious nodes within the iTrust network, by having the simulator randomly flag nodes as malicious. For these experiments, we use a proportion of 0.2 malicious nodes, or 200 malicious nodes in the 1000 node network. Thus, the 150 node neighborhood contains, on average, $30 = (200/1000) \times 150$ malicious nodes.

If a requesting node's message is mishandled by a malicious node, the requesting node detects the malicious node, and either blacklists the offending node immediately, or allows it to have a second chance by placing it on the graylist. We

investigate both possibilities, and run the simulation for 10, 100, 1000, and 10000 requests.

The results are presented in the following tables and graphs. For example, when Requests equals 10, there are 10 sets of metadata, each of which is distributed to m nodes in the local neighborhood and to M nodes in the entire network, and there are 10 requests, each of which is distributed to r nodes in the local neighborhood and to R nodes in the entire network. The Proportion Blacklisted is calculated as Blacklisted/30 for the 150 node neighborhood and as Blacklisted/200 for the 1000 node network.

In the first experiment, we investigate the use of blacklisting for two offenses in the 150 node neighborhood with $m = 9$, $r = 64$, and in the 1000 node network with $M = 64$, $R = 64$. As shown in Table I, the proportion of malicious nodes blacklisted varies with the number of requests and the number of nodes. With 10000 requests, both the 150 node neighborhood and the 1000 node network are successful in blacklisting a similar proportion of malicious nodes, 0.90 and 0.91, respectively. However, when the number of requests is 1000, we see significant differences in the proportion of malicious nodes blacklisted. The 150 node neighborhood is still successful in blacklisting 0.87 of the malicious nodes, but the 1000 node network is able to detect merely 0.13 of the malicious nodes. For 100 requests, the difference is even greater. The 150 node neighborhood recognizes 0.70 of the malicious nodes, but the 1000 node network detects none. In terms of the number of requests it takes for malicious nodes to be blacklisted for two offenses, a smaller local neighborhood performs better than a larger network, as is evident in Figure 9.

Nodes	Distribution	Requests	Blacklisted	Remaining	Proportion Blacklisted
150	$m = 9$	10	3	27	0.10
	$r = 64$	100	21	9	0.70
		1000	26	4	**0.87**
		10000	27	3	**0.90**
1000	$M = 64$	10	0	200	0.00
	$R = 64$	100	0	200	0.00
		1000	25	175	**0.13**
		10000	182	18	**0.91**

TABLE I. THE 150 NODE NEIGHBORHOOD WITH $m = 9$, $r = 64$ VS, THE 1000 NODE NETWORK WITH $M = 64$, $R = 64$, WITH BLACKLISTING FOR TWO OFFENSES.

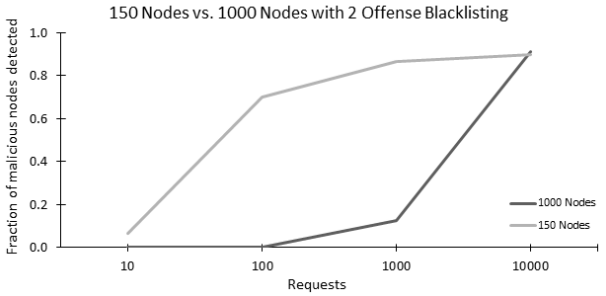


Fig. 9. Proportion of malicious nodes blacklisted for two offenses for various numbers of requests for the 150 node neighborhood with $m = 9$, $r = 64$ vs. the 1000 node network with $M = 64$, $R = 64$.

In the next experiment, we investigate the use of blacklisting for one offense. The values of the parameters are the same as those in the previous experiment. In Table II, we see that with blacklisting for one offense, the 150 node neighborhood still outperforms the 1000 node network for 10000 requests. At 1000 requests, a drop in the proportion blacklisted occurs, similar to the previous case with blacklisting for two offenses. While less severe, the difference between the 150 node neighborhood and the 1000 node network is still significant, with 0.93 vs. 0.34 of the malicious nodes blacklisted. Overall, we see an increase in the proportion of malicious nodes blacklisted when compared with blacklisting for two offenses. However, the proportion blacklisted might be inflated by the detection of false positives. Blacklisting for one offense is more severe, and does not take other factors into account, such as temporary loss of connectivity. Doing so leads to more non-malicious nodes being blacklisted, making blacklisting for two offenses a more reasonable choice, despite the marginal improvement in the proportion blacklisted, compared to blacklisting for one offense. As Figure 10 shows, the difference between the 150 node neighborhood and the 1000 node network is still significant, even with blacklisting for one offense.

Nodes	Distribution	Requests	Blacklisted	Remaining	Proportion Blacklisted
150	$m = 9$	10	6	24	0.20
	$r = 64$	100	24	6	0.80
		1000	28	2	**0.93**
		10000	26	4	0.87
1000	$M = 64$	10	3	197	0.02
	$R = 64$	100	10	190	0.05
		1000	68	132	**0.34**
		10000	183	17	0.92

TABLE II. THE 150 NODE NEIGHBORHOOD WITH $m = 9$, $r = 64$ VS. THE 1000 NODE NETWORK WITH $M = 64$, $R = 64$, WITH BLACKLISTING FOR ONE OFFENSE.

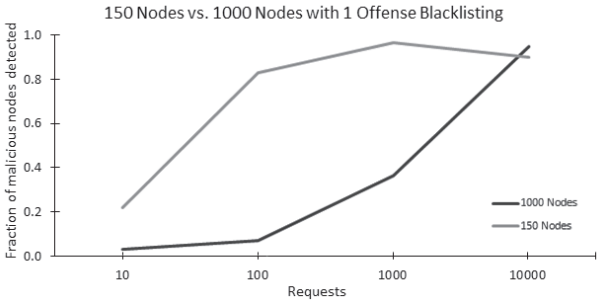


Fig. 10. Proportion of malicious nodes blacklisted for one offense for various numbers of requests for the 150 node neighborhood with $m = 9$, $r = 64$ vs. the 1000 node network with $M = 64$, $R = 64$.

The final experiment that we performed investigates the 150 node neighborhood with $r = 24 \sim 2\sqrt{150}$ requests distributed, compared to the 1000 node network, with blacklisting for two offenses and blacklisting for one offense, as shown in Table III and Table IV, respectively. Whereas a 150 node neighborhood leads to slightly lower match probabilities than does a 1000 node network, in terms of finding malicious nodes, the 150 node neighborhood is able to catch more malicious nodes, for a given number of requests, with both blacklisting for two offenses and blacklisting for one offense.

Table V aggregates the results presented previously, to enable comparisons of the performance of blacklisting for two offenses and blacklisting for one offense, for the 150 node neighborhood and the 1000 node network. Again we consider $r = 24 \sim 2\sqrt{150}$, $m = 9$ for the 150 node neighborhood,

Nodes	Distribution	Requests	Blacklisted	Remaining	Proportion Blacklisted
150	$m = 9$	10	0	30	0.00
	$r = 24$	100	8	22	0.27
		1000	25	5	0.83
		10000	29	1	0.97
1000	$M = 64$	10	0	200	0.00
	$R = 64$	100	0	200	0.00
		1000	25	175	0.13
		10000	182	18	0.91

TABLE III. THE 150 NODE NEIGHBORHOOD WITH $m = 9$, $r = 24 \sim 2\sqrt{150}$ VS. THE 1000 NODE NETWORK WITH $M = 64$, $R = 64$, WITH BLACKLISTING FOR TWO OFFENSES.

Nodes	Distribution	Requests	Blacklisted	Remaining	Proportion Blacklisted
150	$m = 9$	10	3	27	0.10
	$r = 24$	100	21	9	0.70
		1000	27	3	0.90
		10000	29	1	0.97
1000	$M = 64$	10	3	197	0.02
	$R = 64$	100	10	190	0.05
		1000	68	132	0.34
		10000	183	17	0.92

TABLE IV. THE 150 NODE NEIGHBORHOOD WITH $m = 9$, $r = 24 \sim 2\sqrt{150}$ VS. THE 1000 NODE NETWORK WITH $M = 64$, $R = 64$, WITH BLACKLISTING FOR ONE OFFENSE.

and $M = 64$, $R = 64$ for the 1000 node network. As the table shows, for 10000 requests, the proportions of malicious nodes blacklisted are all greater than 0.9. Moreover, for 1000 requests, the proportions of malicious nodes blacklisted for the 1000 network are substantially less than the proportions of malicious nodes blacklisted for the 150 node neighborhood, with blacklisting for both two offenses and one offense.

	150 Nodes $m = 9$ $r = 24$ 2 offenses	1000 Nodes $M = 64$ $R = 64$ 2 offenses	150 Nodes $m = 9$ $r = 24$ 1 offense	1000 Nodes $M = 64$ $R = 64$ 1 offense
Requests	Proportion Blacklisted			
10	0.00	0.00	0.10	0.02
100	0.27	0.00	0.70	0.05
1000	**0.83**	**0.13**	**0.90**	**0.34**
10000	0.97	0.91	0.97	0.92

TABLE V. PROPORTION OF MALICIOUS NODES BLACKLISTED AS A FUNCTION OF THE NUMBER OF REQUESTS FOR THE 150 NODE NEIGHBORHOOD WITH $m = 9$, $r = 24 \sim 2\sqrt{150}$ VS. THE 1000 NODE NETWORK WITH $M = 64$, $R = 64$, WITH BLACKLISTING FOR TWO OFFENSES AND BLACKLISTING FOR ONE OFFENSE.

One could also investigate blacklisting for three offenses. However, we would expect an even greater decrease in the proportion of malicious nodes blacklisted, particularly for the 1000 node network.

These simulations demonstrate the effectiveness of smaller local neighborhoods in an iTrust MANET. Whereas a traditional reputation system requires repeated interactions between nodes to build a reputation table, iTrust seeks to reduce the number of interactions with its smaller local neighborhoods.

V. RELATED WORK

Cho *et al.* [6] present a survey of trust management for MANETs. They discuss classifications, potential attacks, performance metrics, and in particular a trust metric that combines the notion of trust from social networks with quality-of-service. In [5], Cho *et al.* investigate selfish behavior in packet forwarding within MANETs. Their analysis balances altruism, *i.e.*, forwarding packets for the public good, against selfish individual welfare, *i.e.*., not forwarding packets to conserve battery power; however, it does not consider malicious behavior. Such an analysis might be interesting for the iTrust reputation system, but might be vulnerable to malice.

Damiani *et al.* [7] enumerate a range of malicious behaviors that can distort the reporting of nodes' behaviors and the evaluation of nodes' reputation ratings in the Gnutella peer-to-peer network [9]. Their approach to collecting reputation information is based on gathering reports from large numbers of nodes, and on gathering reports for both the resources and the nodes that provide access to those resources. Their global approach does not fit in with iTrust's local neighborhoods that aim to limit the expectation of cooperation among nodes.

Buchegger and Le Boudec [4] investigate a Bayesian approach to evaluate a node's reputation from second-hand reports obtained from other nodes, which is contrary to iTrust's aim to limit the interactions between nodes. To protect a node's reputation against malicious reports, reputation reports that are inconsistent with the node's current reputation are rejected, which might result in the failure to adjust a node's reputation in the presence of subtle malicious attacks. Extending this work, Mundinger and Le Boudec [17] employ an interesting mean-field approach. Such an approach is effective at masking uncorrelated noise, but might not be able to handle correlated misinformation in coordinated malicious attacks.

Guo *et al.* [10] take a different approach to monitoring packet forwarding in wireless ad-hoc networks. They exploit fuzzy sets using mathematical analysis based on Grey theory to detect inconsistent and potentially malicious behavior. We are investigating whether such an approach can be extended to the rather more complex behavior of iTrust nodes.

Zhou and Hwang [20] present a distributed reputation system that places more weight on nodes considered to be the most reputable. Doing so can result in a system that is dominated by a small number of nodes which, in turn, can result in subtle malicious attacks. To address this issue, Jesi *et al.* [13] aim to detect hub attacks. Because hubs concentrate power over reputations, routing, *etc.* into relatively few nodes, hub attacks can distort or disrupt the behavior of the system.

The Collaborative Reputation (CORE) system for MANETs, developed by Michiardi *et al.* [16], uses a collaborative monitoring technique and reputation mechanism, where reputation is based on a node's ability to cooperate with other nodes. Nodes with good reputations are granted the use of resources, whereas nodes with bad reputations are gradually filtered out. Their watchdog mechanism is similar to the neighborhood watch mechanism of the iTrust reputation system; however, their mechanism is less well protected against malicious manipulation of reputation information.

Jelasity *et al.* [12] maintain logs of all outgoing and incoming messages, with signed messages to preclude forgeries. Periodically, the nodes exchange logs, which allows them to check the behavior of other nodes and to detect various kinds of malicious behavior. Such a strategy is less effective in MANETs, where neighborhoods can change quite quickly as the nodes move around. The iTrust reputation system uses only first-hand observations to monitor the behavior of the neighboring nodes.

Ruohomaa *et al.* [19] present a peer-to-peer reputation system in which nodes distribute their reputation rankings to other nodes. In their system, potential interactions are described

by a collaboration contract. Whether such interactions satisfy the contract is verified using non-repudiation receipts, thus preventing reputations from being distorted by misinformation. Adding such a mechanism to the iTrust reputation system would be quite expensive.

Hu [11] presents a reputation system that resists malicious attacks. Nodes develop reputations of their neighbors from observations of their neighbors' behaviors. To make it difficult for a malicious node to subvert the reputations, their system does not communicate the reputations to other nodes. The iTrust reputation system adopts a similar strategy.

The Cooperation Of Nodes: Fairness In Dynamic Ad-hoc Networks (CONFIDANT) protocol, proposed by Buchegger and Le Boudec [3], attempts to detect and isolate uncooperative nodes. The nodes use passive observations of packets forwarded within a one-hop neighborhood. To prevent dissemination of false reputation ratings, the system incorporates a trust rating for each node. First-hand information is stored locally and disseminated to the neighbors, but reputation and trust ratings are not shared. Similarly, the iTrust reputation system does not share reputation information among the nodes.

The Observation-based Cooperation Enforcement in Ad-hoc Networks (OCEAN) system, developed by S. Bansal and M. Baker [2], recognizes that reporting a node's behavior to other nodes renders the system vulnerable to malicious reports. It focuses on first-hand observations of other nodes' behaviors, exploiting the ability of nodes in wireless ad-hoc networks to listen to the transmissions of neighboring nodes. Simulations demonstrate that OCEAN works quite well, even though ratings are based only on monitoring neighboring nodes. Likewise, the iTrust reputation system adopts a local neighborhood strategy.

VI. Conclusions and Future Work

The iTrust local reputation system for MANETs detects malicious nodes and puts such nodes on a blacklist or a graylist. For 10000 requests, the results for a 150 node neighborhood and a 1000 node network are similar in detecting and blacklisting malicious nodes. In contrast, for fewer requests, the 150 node neighborhood yields superior results to the 1000 node network, with most of the malicious nodes being blacklisted. In a MANET, such as that of iTrust, having a large number of repeated interactions with the same nodes is rare and, thus, relying on a large number of requests to detect and blacklist malicious nodes is inappropriate. The smaller local neighborhood provides a means to eliminate the need for a large number of interactions between nodes.

The current reputation system for iTrust mitigates the effects of subversive nodes with respect to messages. However, misbehaving nodes are still capable of disseminating bad data. In a future version of iTrust, we plan to incorporate a mechanism that monitors the message content and rates the information at the end user, perhaps with the user's help. As a result, iTrust will be able to rate the source nodes, based on the content those nodes distribute. This addition will improve the overall robustness of iTrust against malicious behavior.

References

[1] C. M. Badger, L. E. Moser, P. M. Melliar-Smith, I. Michel Lombera and Y. T. Chuang, "Declustering the iTrust search and retrieval network to increase trustworthiness," in *Proc. 8th International Conference on Web Information Systems and Technologies*, Porto, Portugal, Apr. 2012, pp. 312–322.

[2] S. Bansal and M. Baker, "Observation-based cooperation enforcement in ad hoc networks," Tech. Rep. NI/0307012, Stanford Univ., 2003.

[3] S. Buchegger and J. Y. Le Boudec, "Performance analysis of the CONFIDANT protocol (Cooperation Of Nodes: Fairness In Dynamic Ad-hoc NeTworks)," in *Proc. IEEE/ACM Symposium on Mobile Ad Hoc Networking and Computing*, Lausanne, Switzerland, June 2002, pp. 226–236.

[4] S. Buchegger and J. Y. Le Boudec, "A robust reputation system for P2P and mobile ad-hoc networks," in *Proc. Second Workshop on the Economics of Peer-to-Peer Systems*, Cambridge, MA, June 2004.

[5] J. H. Cho, A. Swami and I. R. Chen, "Modeling and analysis of trust management protocols: Altruism versus selfishness in MANETs," in *Proc. IFIP Conference on Trust Management*, Morioka, Japan, June 2010, pp. 141–156.

[6] J. H. Cho, A. Swami and I. R. Chen, "A survey of trust management for mobile ad hoc networks," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 4, 2011, pp. 562–583.

[7] E. Damiani, D. di Vimercati, S. Paraboschi, P. Samarati and F. Violante, "A reputation-based approach for choosing reliable resources in peer-to-peer networks," in *Proc. 9th ACM Conference on Computer and Communications Security*, Washington, DC, Nov. 2002, pp. 207–216.

[8] W. Feller, *An Introduction to Probability Theory and Its Applications I*, John Wiley & Sons, New York, NY, 1968.

[9] Gnutella, 2000, http://gnutella.wego.com/

[10] J. Guo, A. Marshall and B. Zhou, "A trust management framework for detecting malicious and selfish behavior in ad-hoc wireless networks using fuzzy sets and Grey theory," in *Proc. IFIP Conference on Trust Management*, Copenhagen, Denmark, June 2011, pp. 277–289.

[11] J. Hu and M. Burmester, "Cooperation in mobile ad hoc networks," in *Guide to Wireless Ad Hoc Networks*, Springer, London, England, 2009, pp. 43–57.

[12] M. Jelasity, A. Montresor and O. Babaoglu, "Detection and removal of malicious peers in gossip-based protocols," in *Proc. 2nd Bertinoro Workshop on Future Directions in Distributed Computing*, Bertinoro, Italy, June 2004, pp. 23–25.

[13] G. P. Jesi, D. Hales and M. van Steen, "Identifying malicious peers before it's too late: A decentralized secure peer sampling service," in *Proc. First International Conference on Self-Adaptive and Self-Organizing Systems*, Bologna, Italy, July 2007, pp. 237–246.

[14] P. M. Melliar-Smith, L. E. Moser, I. Michel Lombera and Y. T. Chuang, iTrust: Trustworthy information publication, search and retrieval, in *Proc. 13th International Conference on Distributed Computing and Networking*, Hong Kong, China, Jan. 2012, pp. 351–366.

[15] I. Michel Lombera, L. E. Moser, P. M. Melliar-Smith and Y. T. Chuang, "Mobile ad-hoc search and retrieval in the iTrust over Wi-Fi Direct network," in *Proc. Ninth International Conference on Wireless and Mobile Communications*, Nice, France, July 2013.

[16] P. Michiardi and R. Molva, "Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks," in *Proc. IFIP Communication and Multimedia Security Conference*, Canterbury, England, Sep. 2012, pp. 107–121.

[17] J. Mundinger and J. Y. Le Boudec, "Analysis of a reputation system for mobile ad-hoc networks with liars," *Performance Evaluation*, vol. 65, no. 3–4, Mar. 2008, pp. 212-226.

[18] E. M. Royer and C. K. Toh, "A review of current routing protocols for ad-hoc mobile wireless networks," *IEEE Personal Communications Magazine*, Apr. 1999, pp. 46–55.

[19] S. Ruohomaa, P. Kaur and L. Kutvonen, "From subjective reputation to verifiable experiences – Augmenting peer-control mechanisms for open service ecosystems," in *Proc. IFIP Conference on Trust Management*, Surat, India, May 2012, pp. 142–157.

[20] R. Zhou and K. Hwang, "PowerTrust: A robust and scalable reputation system for trusted peer-to-peer computing, *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 4, Apr. 2007, pp. 460–473.

Supporting Secure Scalable End-To-End QoS In 4G Mobile Wireless Networks

Odhiambo Marcel O, Muchenje Best
ohangmo@unisa.ac.za, 46730133@mylife.unisa.ac.za
Department of Electrical and Mining Engineering, University of South Africa (UNISA),
P.O. Box 392, UNISA - 0003, South Africa.

Abstract: *With the convergence of the Internet and wireless communications, mobile wireless networks and data services are undergoing tremendous evolutionary growth that has seen the development of fourth generation (4G) mobile wireless access technologies based on an all-IP platform. However, major challenges in the development of such heterogeneous network infrastructure such as quality of service (QoS) provisioning and network security services for mobile users' communication flows, among others still exists. In this paper an integrated architectural view and methodology for QoS and security support in 4G mobile wireless networks, which integrates QoS signaling with secure enhanced evolved packet system authentication and key agreement (SE-EPS AKA protocol) is presented. The success of 4G mobile wireless networks depends on the prudent deployment of homogeneously designed, high-speed, secure, multiservice IP-centric integrated multimedia, voice and data networks.*

Keywords: Terms- 4G mobile wireless networks, Security enhanced Evolved Packet System Authentication and Key Agreement (SE-EPS AKA), quality of service (QoS)

1. Introduction

The introduction of fourth generation (4G) mobile wireless networking has brought about a number of interesting but also scaring challenges, chief among them is the integration of quality of service (QoS) and network security in an environment now heavily proliferated with computing devices with diverse computing capabilities, which poses a great risk when in the wrong hands. This is further compounded by business models pursued by different telecom services. The use of IPv6 protocol as a convergence layer has immensely eased the support of seamless mobility and QoS across heterogeneous networking environments, provision of content-rich multimedia and value-added services in such multi-provider heterogeneous network environments demands a common signalling framework for session negotiation, network resources reservation, session and QoS negotiation, and most importantly, integrating QoS and network security services in the signalling framework.

This paper focuses on developing a seamless integration of QoS and network security services for heterogeneous 4G mobile wireless networks. And as such the QoS sub-system conceptualised here is based on the use of QoS brokers (the mobility management entities, MME) that manage network resources and performance admission control for user equipment (UE) data flows. The proposed architecture give rise to three scenarios for session setup and (re)negotiation, differing on the entity that issues requests to QoS broker, namely (a) user equipment (EU) itself, (b) services proxies within the framework and (c) modules in the network access routers, that are able to do application signalling parsing and modification.

2. Overview of the QoS Services Architecture

Figure 1 is an outline of a 4G mobile wireless network. It illustrates the architecture of the evolved packet core (EPC). The radio-access network (RAN) and the evolved packet core (EPC) are also referred as the evolved packet system (EPS). Detailed explanation for the functionalities of the various entities of this network architecture is given in the literature [1].

The main design aim of 4G mobile wireless networks is the support of seamless UE mobility under a unified heterogeneous architecture that accommodates scalable and incremental development of new advanced applications and services. Thus, the IPv6 protocol, used as the convergence layer of 4G systems, is used natively to support mobility. The IPv6 creates an abstraction layer that conceals technology-specific application environments. Extension enhancements added to IPv6 in 4G mobile environments completely provides seamless mobility with fast handoffs. The correlation between mobility and QoS is outlined in [2].

In the proposed 4G network model several network domains, each with a host of access networks supporting disparate wireless technologies, are interconnected to each other via a core network, thus allowing different network operators to internetwork in a common environment. Special arrangements amongst operators have to be in place to allow integration of services and applications across different network domains. Figure 2 illustrates the proposed network architecture.

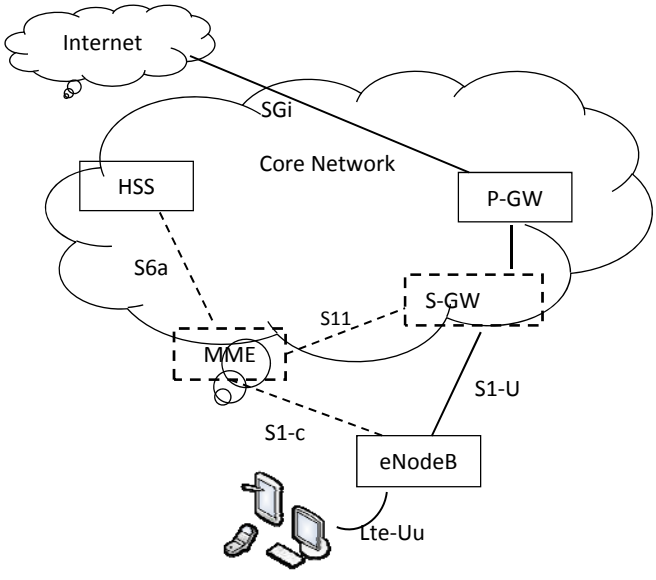


Figure 1: Core Network Architecture [1]

The MMEs in the access network perform admission control for data flows and inter- and intra-domain handoffs, and manage network resources, configuring the

access routers, in policy decision or enforcement relationships. In addition, the MMEs help optimise network resources by performing load balancing for the users and sessions among the available networks through the use of network initiated handoffs. QoS support in the network core is based on DiffServ for scalability reasons, thus enabling aggregated inter-domain network segments.

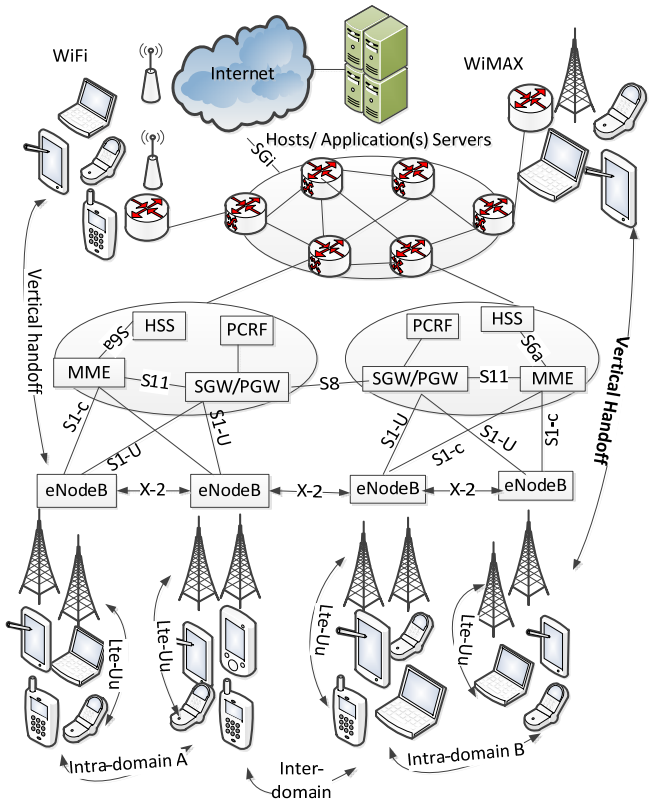


Figure 2: Network architecture [2]

The aggregated information is propagated to the access network MME where it is used for admission control in order to achieve end-to-end QoS for data flow. The integration of InterServ and DiffServ allows per-flow and per-aggregate processing of data in two-layer hierarchy architecture of which the end result is providing fine-grained QoS control while keeping the scalability properties of per-aggregated core resource management decoupled from per-session signalling. The service provision platform (SPP), within the network core, enable the running of the services and applications, through the multimedia service platform (MMSP), which consists of a broker, and proxy servers responsible for the provision and control of multimedia services and is also capable of mapping application level QoS configurations to network resource requirements and performing QoS requests for data flows. This architecture has a large degree of flexibility in QoS signalling, enabling the use of diversity of QoS access signalling scenarios that will fulfil the needs for different applications and different business cases for a diverse range of network operators and access services providers.

Unification of the scenarios is achieved by centralisation of the admission and handoff control at the access network MMEs. The SPP contains a core network MME which is responsible for resource management in the network core. Policies for resource management are defined by the policy-based network management system (PBNMS) and are sent to the core network MME where they are cached in a local database for use. The central monitoring system (CMS) collects statistics and other network usage data from network monitoring entities and feeds the PBNMS and MMEs with this information for proper network resources management [2].

During the user registration on the network the access MME retrieves a subset of the user's profile from the Authentication, Authorisation, Accounting, Auditing and Charging (AAAAC) system, which is part of the HSS and PCRF, in the EPC. This is meant to improve efficiency and scalability. This subset, called the network view of the user profile, contains information on the set of network services such classes of services, bandwidth parameters etc. as outlined in the service level agreement specifications for each user. Similarly, a service view of the user profile, containing information on the higher level services available to the user such as voice calls, video telephony, and the respective codecs, that is retrieved by the MMSP to control multimedia services [2].

3. QoS Signalling Scenarios Testing and Validation

This section outlines different QoS signalling scenarios during multimedia call initiation between two UEs. According to this proposal the MMSP, ARM and the UE are able to issue QoS requests (several signalling protocols such as SIP can be used). Figure 3 illustrates a simplified scenario in which UE1 initiates a multimedia session with another terminal UE2, where the two UEs are in different network domains.

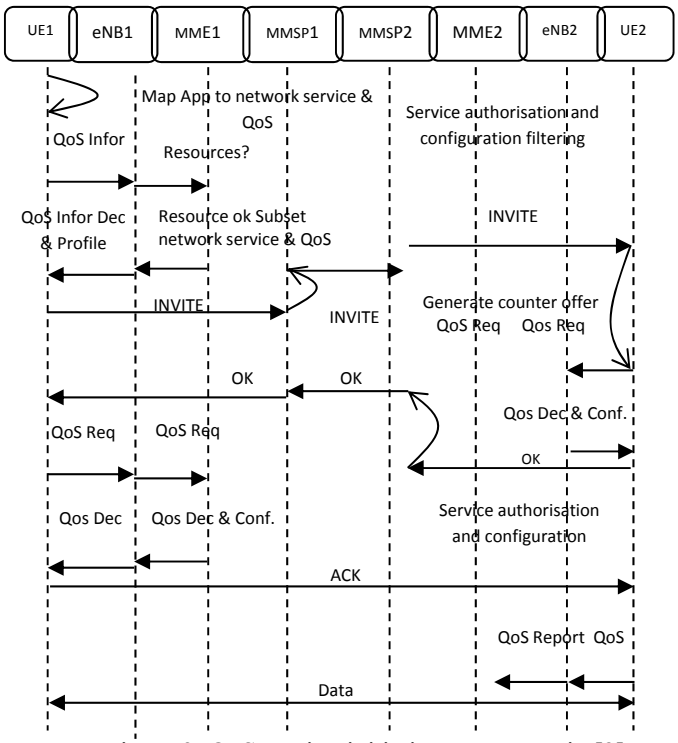


Figure 3: QoS session initiation, UE scenario [2]

The user equipment UE1, with the help of its resident QoS client, maps application needs to the networks and QoS requirements and sends requests to its serving access

network QoS broker MME1 via a QoS attendant in the access router eNB1. The QoS signalling between the QoS client and the attendant is implemented as an extension to resource reservation protocol (RSVP). The MME1 respond with information on the available resources according to user profile and network status. If allowed by the MME1, the UE1 sends an INVITE message indicating the initial QoS parameters to UE2. Upon receiving the INVITE message MMSP1 performs service authorisation (with the help of the SEEP aka protocol), filtering out services and applications not allowed in the service level agreement specification outlined for the user UE1. Once authorised, the INVITE is forwarded to UE2. UE2 matches the QoS parameters in the INVITE messages to its own, requests MME2 for available network resources, and generates a counter-offer. Upon receiving this message the MMSP2 filters the services to those authorised. When the response message arrives at UE1, it selects the service to use, informs MME1 to configure the access router accordingly with the required bandwidth and queues available space for the data flows and service classes, and sends the acknowledge message (ACK) containing the final configurations that will be used. This message triggers the sending of QoS reports to MME2 confirming the QoS configuration parameters in the access routers. Applications that make use of out-of-band signalling (signalling that make use of some form of a separate dedicated communication channel) may also be made QoS aware by coding them to invoke this procedure [3].

The second scenario involves the MMSP. The UEs do not perform QoS requests: in this instance they perform some form of SIP (session initiation protocol (SIP) used in signaling communications protocol for controlling multimedia sessions such as voice and video calls over the Internet Protocol networks) signalling through the use of extended proxy servers which are capable of parsing QoS configuration parameters, mapping them to networks resource requirements and contacting the MMEs to perform QoS requests. These proxies also enforce policies set out in the service level agreements configured by the operators as per user service needs and requirements (as reflected by the respective service view of user profile) [2].

4. Security Services Protocol Verification

The following section describes a computational framework for proving the SEEP AKA using a cryptographic verification tool, the CryptoVerif tool. The specification of the CryptoVerif tool is translated into OCaml [5] to produce the implementation of the SEEP AKA protocol. OCaml is a functional language, which also facilitates the compilation because the CryptoVerif specification uses oracles that can be immediately translated into functions [3].

Proving the security protocol alone is not sufficient. The specification of the protocol may be correct, but the implementation can carry some errors as explained in [4]. There are several ways of obtaining a secure implementation of a security protocol, one of which is writing the specification first, proving it correct, and then generating and implementing it. Thus, according to [4] the general belief is to start by designing the protocol, formalise it, prove it secure formally and then finally implementing it. This is the methodology adopted in this research paper.

In order to generate the SEEP AKA implementation a compiler that takes a CryptoVerif specification and returns an implementation in OCaml is pursued [3].

Figure 4 illustrates an overview of the approach used to obtain a proved implementation of a cryptographic protocol. Two distinct steps are observed. First, a written specification of the CryptoVerif protocol is obtained. This specification contains a list of security assumptions on the cryptographic primitives. This specification guarantees the desired security properties, for example, the secrecy, authentication, authorisation, auditing etc., in the computational model by using the CryptoVerif tool [4].

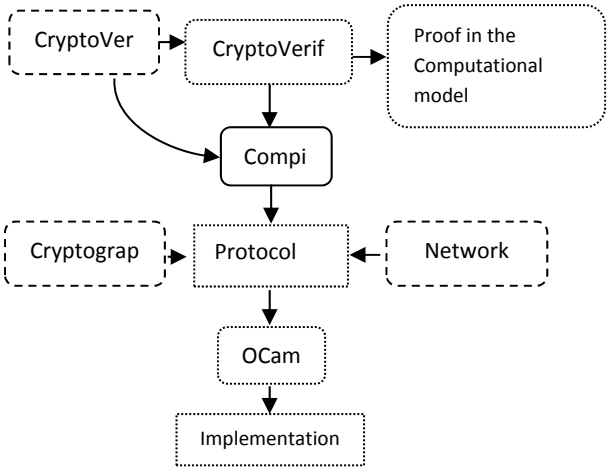


Figure 4: Overview of the approach [5]

Second, the compiler transforms the specification into protocol code. To build the implementation, the following codes are generated:

(i) The code corresponding to the exchange of messages across the network, which uses the results given by the functions in the protocol code. This code can be considered as a part of the adversary, and so it is not required to prove this part of the code.

(ii) The code corresponding to the cryptographic primitives. This is used by the protocol code, and thus must be proved manually that the primitives satisfy the security assumptions made in the specification file [5].

Then, the OCaml compiler is used on these codes to implement the protocol, from which a single protocol specification is obtained as both proof that the protocol is secure in the computational model and in executable implementation of the protocol.

The protocol implementation derived is used to formulate a security framework. The solution so proposed, on the user's side, comprises of the user equipment (UE), whose design is based on some form of trusted mobile platform (TMP) [6], and a biometric reader (BR) as shown in Figure 5. The access network, the home network environment, the service provider and the user equipment manufacturers host some form of public certificate issued by their own trusted authority which in turn should have connectivity to 4G mobile networks to

enhance secure flow of information between the user equipment and the various 4G network entities.

Nomenclatures used for the proposed security scheme are donated as follows:

ID_X - X's identity
SK_X - private key
PK_X - public key
$Cert_X$ - digital certificate
Sig_X - digital signature
H(x) - a secure hash function and E(k, x) represents encrypting content x with key k.

The authentication schemes begins by having a user password (PW) and a universal subscriber identification module (USIM) that is capable of checking the integrity and validity of the mobile platform, and also have the capability of storing authentication parameters that includes the user's biometric template (F_U, usually eye iris, facial identity or fingerprints), SK_{User}, $Cert_{User}$, $Cert_{HE}$, x, y and z. The authenticating parameters x, y and z are computed by user's HE as follows before the home environment (HE) issues the USIM card to the user. n is a secure module of RSA signature algorithm given as.

$x=H(F_U||PW)$, $y=x\oplus H(PW)$, $z=S\oplus H(F_U\oplus PW)$, and then $S = H(ID_{user}||\ PW\ ||\ F_U)^{SKHE} \bmod n$.

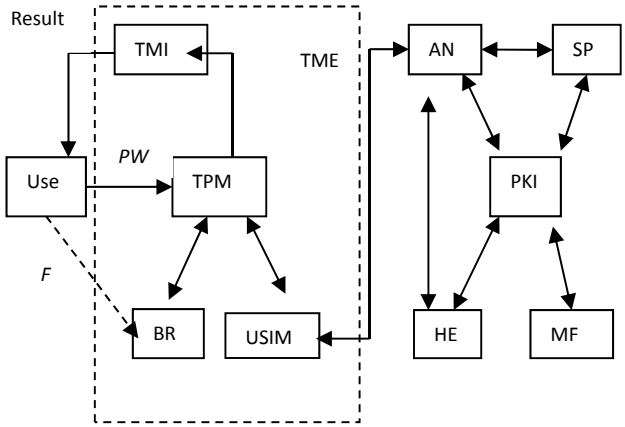


Figure 5: Security framework based on TMP and PKI [6]

The UE stores the symmetric key, SK_{TPM}, the biometric template, K_{Fu}, shared and $Cert_{TPM}$, and as well as integrity metrics of other components in the UE. The HE saves user's security credentials (ID_{User}, S, $Cert_{User}$) securely in its database [6].

Since 4G mobile networks architecture are based on an all-IP network platform authentication is perceived as a service performed at higher protocol layers regardless of the underlying technology. To accomplish mutual authentication between the UE and the 4G mobile network, two phases of authentication are proposed, namely local authentication where the UE checks the integrity and validity of the authentication parameters input by the user (password and biometric information) during initial boot up of the UE, and the remote mutual authentication between the UE, serving network (SN) and the HE when UE initially requests to be attached to the network [6].

5. Local Mutual Authentication

Local authentication procedure can be described in eight steps m_1 to m_8 as outlined as follows.

First, the USIM generates a token r_1 and sends an integrity check request D_1 with (r_1, ID_{USIM}) as m_1 to the TPM.

m_1: USIM →TPM: r1, ID_{USIM}, D_1

On receipt of m_1, the TPM issues a token r_2 and sends an integrity check request D_2 with (r_2, ID_{TPM}) to the BR.

m_2: TPM → BR: r_2, ID_{TPM}, D_2

Upon receiving m_2, BR encrypts its integrity metric D_3 with (r_2, ID_{TPM}) using K_{Fu} and responds with MAC_{BR} to the TPM.

m_3: BR → TPM: MAC_{BR}
$MAC_{BR} = E(K_{Fu}, r2||ID_{TPM}||D_3)$ (1)

Using the integrity metrics of BR and that of other components pre-stored in its internal database, the TPM checks whether the received MAC_{BR} is valid and also the integrity of other components of the TMP needed to perform the authentication operation are correct. Then the TPM generates a token r_3 and signs its own integrity metric D_4 with (r_1, r_3, ID_{USIM}). The TPM forwards the token r_3, $Cert_{TPM}$ and Sig_{TPM} to the USIM.

m_4: TPM → USIM: r_3, $Cert_{TPM}$, Sig_{TPM}
$Sig_{TPM} = E[SK_{TPM}, r_1||r_3||ID_{USIM}||D_4]$ (2)

The USIM then issues a token r_4 and calculates (C_1, Sig_{User}) as in equations 3 and 4. The SN's public-key parameters can be gained with the help of PK_{BP}. Then USIM sends them with (r_1, r_3) to SN to verify Sig_{TPM}. Where IDC_{User} is a unique identity of user's certificate and TS is a timestamp [6].

m_5: USIM → SN: r1, r_3, C_1, Sig_{TPM}, Sig_{User}, $Cert_{TPM}$.
$C_1 = E(PK_{AN}, ID_{User}||IDC_{User}||r_4||TS)$ (3)
$Sig_{User} = E(SK_{User}, IDC_{User}||ID_{TPM}||r_1||r_3||TS)$ (4)

The serving network SN decrypts C_1, checks TS if it is acceptable and turns to PKI to gain valid $Cert_{User}$ according to (ID_{User}, IDC_{User}). After verifying the validity of ($Cert_{TPM}$, Sig_{TPM}, Sig_{User}), SN pre-authenticates user. Then SN buffers (ID_{User}, $Cert_{User}$, $Cert_{TPM}$, r_4) temporarily and responds to USIM after checking result D_5 on TMP followed by MAC_{AN}.

m_6: SN → USIM: D_5, MAC_{AN} .
$MAC_{AN} = E(r_4, r3||ID_{User}||ID_{TPM}||D5)$. (5)

If the received MAC_{AN} is correct and (Sig_{TPM}, $Cert_{TPM}$) pair is valid according to D_5, then both TPM and SN are identified by USIM. As shown in Figure 6, USIM generates a token r_5 and sends (C_2, C_3) computed as in equations 6 and 8. The biometric comparison software (CS) is also encrypted in C_3 and sent from USIM to TPM. After USIM shows the current state of platform is

trustworthy via TMI, the user is allowed to input his password and the BR to TPM. The captured biometric template (F_u') is encrypted in C_4, and including K_{Fu} which is then sent to the TPM [5].

m_7: USIM → TPM: C2, C3. BR → TPM: C_4.

$C_2=E(PK_{TPM}, r_5||y||ID_{USIM})$, (6)

$K_{ST}=H[(r_5 \oplus r_3)||x||ID_{TPM}]$, (7)

$C_3=E(K_{ST}, r_5||ID_{TPM}||F_U||CS)$. (8)

$C_4=E[K_{BT}, ID_{BR}||ID_{TPM}||r_2||F_U']$. (9)

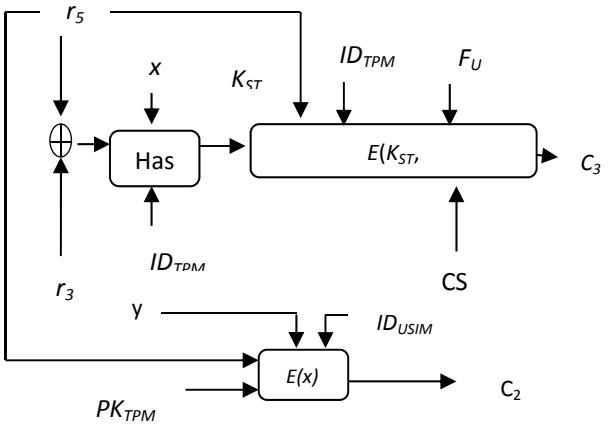


Figure 6: Data encapsulation algorithm in USIM [6]

Once the password PW is input by user and (r_5, y) pair is decrypted from the received C_2, the TMP, first, calculates

$x=y \oplus H(PW)$ (10)

and K_{ST} as in equation 7. Using equation 8 the TPM then decrypts C_3 with K_{ST} and recovers (r_5, ID_{TPM}, BT, CS). If the (r_5, ID_{TPM}) contained in C_3 are both correct, the TPM checks whether equation 11 holds.

$H(BT||PW) = ID_x$ (11)

If equation 11 holds, USIM is identified by TPM. Then TPM decrypts C_4 sent by BR and checks (ID_{BR}, ID_{TPM}, r_2) if they are all correct. TPM makes a comparison between the F_U and F_U' in use in CS to determine to what degree they match. If the match is achieved successfully, the user is authenticated by TPM.

Then, the TPM computes $H(F_U \oplus PW)$ and transfers C_5 computed as equitation 10 to USIM, where D_6 is the authentication result of the user. If (ID_{USIM}, r_5) contained in C_5 are both correct and, then the user is valid according to D_6. Both user and TPM are now identified by USIM.

m_8: TPM → USIM: C_5.

$C_5=E(K_{ST}, ID_{USIM}||r_5||H(F_U \oplus PW)||D_6)$. (12)

6. Remote Mutual Authentication

The SE-EPS AKA (mutual authentication) follows a seven step process as detailed in Figure 7, employing the encryption keys generated during phase 1 of the mutual authentication process. The UE initiates the network access Attach Request, first by using the HSS public PK_H to encrypt the (international mobile subscriber identity) IMSI and get the A (where A = $\{IMSI\}_{PKH}$,) and ID_{HSS} pair, which is then subsequently forwarded to the MME during the access request process. Upon receiving the access request of the UE, The MME uses the public PK_H to encrypt its own network identity (SNID), and then derive information B. The encrypted data A and B, regarded as the authentication data request, is sent to the HSS. Upon receiving the authentication data request from the MME, the HSS decrypts A and B to get the IMSI and SNID using its own public SK_H. The IMSI and SNID are validated by comparing them to the information stored in the HSS database. Once the verification process is over the HSS generates the random array RAND(1,...., n) and the authentication vector AV(1,....n).

As outlined in Figure 7, the SE-EPS AKA protocol calculates the following parameters:

$$K_{ASME} = s10_K\,(f3_K\,(RAND), f4_K\,(RAND), SNID)$$

$$XRES = RAND \oplus SNID \quad and;$$

$$AV = RAND \,||\, SNID \,||\, K_{ASME} \,||\, XRES$$

This information is used by the HSS to calculate the encryption data C=$\{AV(1,....,n), IMSI\}_{PK\,H}$ and sends to the MME as the response [5].

The MME then decrypts C to derive the authentication vectors AV(1,...,n) and IMSI. Amongst the authentication vectors AV(1,....,n) the MME chooses only one authentication vector AV(i) which has not been used before and extracts the random number RAND(i) and SNID found in the database. Exclusively the MME allocates the cipher key identifier $KSI_{ASME}(i)$ to $K_{ASME}(i)$ of the authentication vector AV(i) and utilize the IMSI and the algorithm shared by the MME and the UE to create S-TMSI used for access once more. After completing the one-time authentication and cipher key negotiation the UE and the MME both store the corresponding relation between $KSI_{ASME}(i)$ and $K_{ASME}(i)$. If access is required once again the UE and the SN will take into account the $K_{ASME}(i)$ and in so doing confidential communication can be established without initiating the authentication process again. Finally the MME encrypts the RAND(i), S-TMSI (securely generated temporary mobile subscriber identity), $K_{ASME}(i)$ and SNID by public key of the UE to calculate data D, which is then subsequently sent in subscriber authentication request to the UE. Thus eventually the operation MME → UE: D = $\{RAND(i), SNID, KSI_{ASME}(i), S\_TMSI\}_{PKU}$ is completed in the process [6].

Once the UE has received the subscriber authentication request from the MME it decrypts D using the public key SK_U to recover RAND(i), S-TMSI and the SNID. The UE compares S-TMSI derived from the decryption of D to the one it has calculated to realize the authentication to HSS. If there is no consistency, it means the HSS is not valid and the process is terminated. In case of consistency being observed, the UE computes:

$RES(i) = RAND(i) \oplus SNID$ and $K_{ASME}(i) = sIO_K(f3_K(RAND(i)), f4_k(RAND(i), SNID)$ and RES(i)

is considered the response to the subscriber authentication request sent to the MME [7].

The MME compares the RES(i) received to the XRES(i) of the authentication vector AV(1,...,n). If these two agree, the subscriber is valid. For any subsequent local communication the MME and the UE will consider the $K_{ASME}(i)$ as the intermediate cipher key with which to create the encryption cipher key (CK(i)) and integrity cipher key (IK(i)), or else the while process in halted [7].

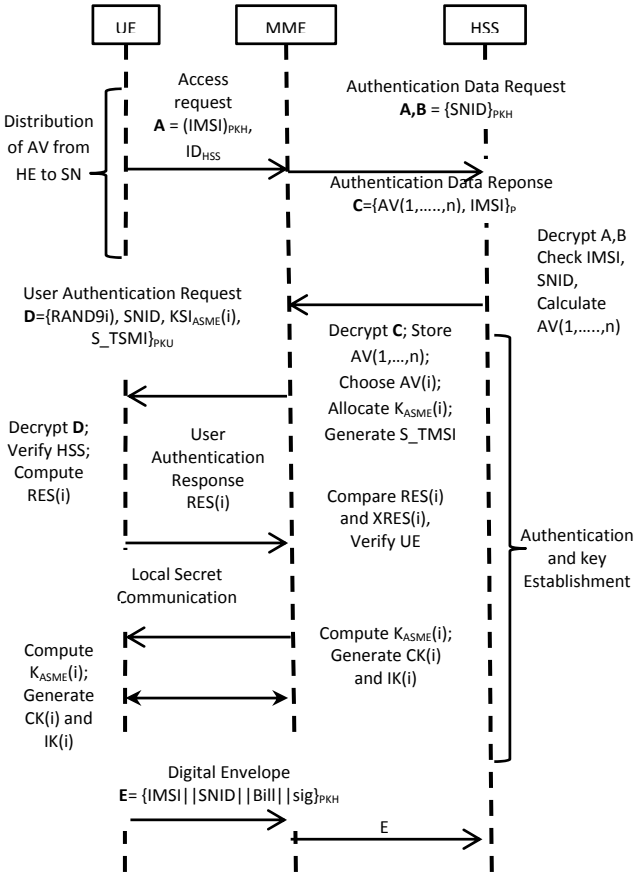


Figure 7: The SE-EPS AKA process [7]

Finally, the MME and the UE store the corresponding relation between the S-TMSI and (IMSI, AV(1,...,n), $KSI_{ASME}(i)$, $K_{ASME}(i)$, CK(i), IK(i)) in their internal databases. After the subscriber and the service network complete the transaction the UE can utilise its own cipher key SK_U to sign the IMSI, SNID and the business information bill for creating charging evidence {IMSI || SNID || bill || sig}. Furthermore, in order to prevent leakage of IMSI and SNID, the public key PK_H is used to create the digital envelope E in which information is transferred to HSS via MME and can be used as evidence for presence and business participation of MME and subscriber as well as creation of related charging relation [6].

7. Implementation and Simulation Results

The simulation model developed is designed to test end-to-end QoS services in mobile wireless networks taking into account the three signaling scenarios outlined earlier. To achieve this session signaling delay and system response to congestion situations due to multiple calls and services requested at the same time are tested as a way of trying to define the QoS parameters of 4G mobile wireless networks. Situations simulated involve (a) UE1, the mobile caller, in the home domain or roaming, (b) UE2, the call recipient, in the home domain or roaming, (c) UE1 and UE2 involved inter- and intra-domain exchange of information and (d) both UE1 and UE2 attached to different wireless access technologies both in inter- and intra-domain scenarios [2].

The 4G network thus developed and designed consists of a pure but basic 4G network architecture interconnected to WiMAX and WiFi networks by a purely IPv6 backbone network, highlighting the presence of inter- and intra-domain interconnections.

The SEEP AKA protocol would be verified using the CryptoVerif tool in conjunction with OCaml.

8. Future Studies and Recommendations

The paper is a reflection of work in progress in which a model for a heterogeneous 4G mobile wireless network is simulated to test and verify the feasibility of integrating QoS and network securing signaling using ns-3 [8], a discrete-event network simulator. Developing a 4G core network with clearly defined network entities to allow an almost real industry-like live network environment that can seamless simulate nearly all network scenarios should be the thrust for future work. Such a network model will make it possible to test new services, especially QoS and security-related issues in order to cope with the ever-changing security threats of the ICT landscape. Current low cost and open-source simulation tools and models should be enhanced and developed.

9. Bibliography

[1] Dahlman, E., Parkvall, S., & Sköld, J. 4G LTE/LTE-Advanced for Mobile Broadband. Academic Press, 2011.

[2] Rui, P. & Sargnto, S. QoS and Session Signaling in a 4Gnetwork. http://www.researchgate.net/publication/4244345_QoS_and_session_signaling_in_a_4G_network

[3] http://caml.inria.fr/

[4] http://www.cryptoverif.ens.fr/

[5] Cadé D. & Blanchet B. From Computationally-proved Protocol Specifications to Implementations, In 7th International Conference on Availability, Reliability and Security (AReS 2012), pages 65-74, Prague, Czech Republic, August 2012. IEEE.

[6] Zheng, Y., He, D., Yu, W. & Tang, X. Trusted Computing-Based Security Architecture For 4G Mobile Networks. Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies. Computer Society, IEEE, 2005.

[7] Xiehua, L. & Yongjun, W.: Security Enhanced Authentication and Key Agreement Protocol in Next Generation Mobile Network, International Journal of Advancements in Computing Technology (IJACT) Vol. 4, No.3, February 2012.

[8] http://www.nsnam.org/z

Impact of Compression and Encryption Methods for Software Architectures in Data-Oriented Mobile Applications

Md. Ashrafur Rahaman and Michael Bauer
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada

Abstract *- Mobile applications have evolved to become extensions of enterprise applications and systems that often interact with remote data sources and databases. Two software architectural approaches have emerged as the primary ways to develop mobile applications accessing remote data sources: client-agent-server and client-intercept-server. Both make use of middleware with the main difference being in the way agent components are used. Our previous compared the performance of these architectural approaches under differing sizes of data retrieved and with and without caching at the middleware. In this paper we compare these architectures in similar data scenarios but looking at the impact of data compression and encryption on the performance of the two approaches. Statistical analysis shows that the client-agent-approach generally performs better, though not in all circumstances. The results of this research provide useful guidelines for developing mobile applications needing to connect to remote databases.*

Keywords: mobile applications, software architectures, enterprise data sources, performance.

1 Introduction

With the popularization of mobile computing and the emergence of a variety of mobile devices, developers face a number of challenges besides concerns of limited power: the heterogeneity of devices [1], the occasional intermittent communication in a mobile environment, and device limitations, such as: limited working memory, limited storage, limited processing power, and small screen. This means that developing large scale mobile applications which can connect to remote data sources or databases through wireless connections with high computational business logic must take into account these limitations. Support for complex or extensive applications in mobile phones which make use of data from remote sources or need remote computing capabilities still require servers and/or middleware.

Popular mobile relational databases, like IBM's DB2 Everywhere 1.0, Oracle Lite, Sybase's SQL etc., work on hand held devices and can provide local data storage for relational data acquired from enterprise relational databases. The main constraints for such databases relate to the size of the devices' memory and size of the program, as handheld devices have memory constraints [2]. Moreover, enterprise databases cannot be replaced by these mobile relational databases.

To support extensive applications in mobile phones that require retrieval of data from remote data sources, middleware is needed which has the capacity to deal with mobile agents and remote database servers, and can improve the transmission of data by implementing caching, pre-fetching, data prioritizing and data compression techniques. Similarly, some mobile applications require data encryption, such as medically-related applications involving patient data. Mobile applications access the middleware through an API to make the communication between mobile agent and middleware transparent.

Various software models have been proposed for addressing the development of mobile applications that interact and utilize enterprise data sources; these include the client-server (C/S), client-agent-server (C/A/S), client-intercept-server (C/I/S), peer to peer, and mobile agent models. Spyrou et al. [3] qualitatively and quantitatively analyzed a set of software models built on the client/server model for mobile agents accessing a Web server. They argued that the C/A/S and C/I/S models were most appropriate and they compared the C/A/S and C/I/S models in the context of browsers and web servers for a wired network. Based on their results, the C/A/S model requires considerably less time than any other client/server model. Though the C/I/S model lacks in performance due to the presence of a client-side agent, it supports compatibility, since it can be built on top of existing applications. According to the researchers, the C/I/S model provides more flexibility than C/A/S and that should have been translated to better performance; their results show that the C/I/S model performs better for heavy-weight clients with large computational power.

Their work focuses on browser-based interactions. They do not consider mobile applications which may need to access data sources within one or more enterprises. The submission of queries and retrieval of results introduces different challenges in the movement of data and related processing, e.g. for security purposes. Our previous research [4] examined both the C/A/S and C/I/S models in the context of mobile applications where there is a need to access data sources or databases that are in remote locations. We compared the performance of the C/A/S and C/I/S software architectures for different size data queries and databases and considered the impact of using middleware caching. We

compared the performance impact in terms of transmission time required for queries. The comparison entailed a mobile application running on a mobile phone, connecting to a middleware server hosted remotely with a database server located on another remote host. Statistical analysis showed that the C/A/S model performed somewhat better, though not in all cases.

In this paper, we continue our comparison of these two architectures by comparing them when data compression and data encryption are involved. In this case, compression and encryption are done at the middleware server and then decompressed or decrypted at the mobile application; we assume that the middleware server is part of the same "enterprise infrastructure" as the server and that the primary concern is communication with the mobile application. Our experiments are performed in real life scenarios instead of a laboratory setup. The experimental results provide guidelines to the mobile application developers about the software models and techniques or combination of technologies to use for QoS of data transmission for mobile applications.

The remainder of this paper is organized as follows: In Section 2, we review some of the existing literature on architectural approaches for supporting mobile applications. In Section 3, we briefly describe the client-agent-server and client-intercept-server software architectures. In Section 4 we describe our implementation and Section 5 presents a summary of the experiments and analyses. Section 6 provides a discussion of the results and some future directions.

2 Middleware for Mobile Applications

A number of different research efforts during the last decade that have focused on middleware and data transmission techniques for mobile applications. The work of Spyrou et al. [3] was discussed above. Capra and Mascolo [5,6] identify requirements for middleware that supports mobility. They suggest the use of a reflexive middleware [7] in a mobile computing environment to solve the problem of losing network connectivity during the movement of the mobile devices. Campbell et al. [8] propose a middleware, named Mobiware, to support QoS in multimedia applications. The platform uses adaptive algorithms to support QoS controlled mobility. Bellavista et al. [9] present a middleware for mobile users, also based on mobile agents, to keep services running regardless of a user's mobility. The middleware provides services to support a virtual environment, virtual terminal, and resource manager.

Chan and Chuang [10] propose MobiPADS, a middleware that uses information from the physical environment to perform self-configuration. It allows dynamic adaptation in both the application and the middleware itself. MobiPADS has two parts: a client at a mobile device attached to the Internet through wireless or cellular networks and a server at the wired network. This extended server-client application is designed to support multiple MobiPADS clients and is responsible for most of the optimization computation. MobiPADS collects metrics about the environment such as bandwidth, latency, and processor usage, and notifies the applications that use those data.

Middleware for multi-client and multi-server mobile applications based on the C/A/S model discussed in [11], where the authors addressed the problems of the heterogeneity of devices in such environments. The proposed middleware provides a communication API to develop applications for mobile environments and allows applications to divide their work amongst server and client sides. The middleware enables applications to exchange messages transparently between the application in mobile device and the middleware, and independently of a specific communication protocol. The researchers did not quantitatively compare the performance of the C/A/S and C/I/S models for their applications.

Caching and pre-fetching to improve a mobile application's data transmission efficiency have been explored by several researchers. Gupta et al. [12] concentrated on the aspects of data management over mobile ad-hoc networks and proposed to estimate the global distribution and then predict and cache the most popular data in the hope of being able to provide it to other devices. Similarly, Yin and Cao [13] propose a cooperative caching scheme for mobile ad hoc networks using caching of popular data so that the availability in mobile ad hoc networks is increased. According to their simulation results, the proposed schemes can significantly improve the performance in terms of query delay and message complexity when compared to other caching schemes. Cheluvaraju, et al. [14] propose anticipatory retrieval of data. Caching is done asynchronously in the background during times of high bandwidth. They propose algorithms to assess the relevance of the data and then prioritize data downloads to cloud storage. The model provides better performance, adapts to varying bandwidth, and pre fetches data from cloud with better accuracy and relatively little overhead.

Data transmission in the wireless network can be vulnerable to security attacks and, thus, ensuring data security is an important concern in some situations. Huang et al [15] introduced a security model based on message digests, encryption and decryption technology to access remote data securely. Researchers implemented this security model in mobile-agent architecture with large number of remote data processing tasks. This research did not explicitly examine the performance impact of using encryption for delivery of data to mobile applications.

Our research focuses specifically on assessing the performance of the C/A/S model and C/I/S model on different sized query results as well as investigating the impact of of compression and encryption. Our main goal is to analyze the performance of the architectures for mobile applications that require access to and retrieval of data from remote enterprise databases.

3 Software Architectures

Mobile applications are normally structured as multi-layered applications consisting of UI, business, and data layers. A developer may choose to develop a thin Web-based

client or a rich client. In the case of a rich client, the business and data services layers are likely to be located on the device itself. On the other hand, the business and data layers will be located on the server for a thin client. Figure 1 illustrates a common rich client mobile application architecture with components grouped by areas of concern [16].

In our research, we focus on the client-agent-server (C/A/S) and client-intercept-server (C/I/S) models because we are retrieving data from remote enterprise databases or data sources and these two models seems to be the best fit for those tasks based on previous research. Moreover, our review of middleware based on mobile browsers says that most of the mobile browser engines were developed using either C/A/S or C/I/S model to retrieve from websites.

The C/A/S architecture is a popular extension of the client-server model, providing a three-tier architecture (Fig. 1). Here, any communication goes through the mobile agent. On one hand, the agent acts as a mobile host. As well, the agent is attached to a remote database or data source and any client's request and server's response is communicated through the agent. In this scenario, a mobile host is associated with as many agents as the services it needs to access. Agents split the interaction between mobile clients and fixed servers into two parts: between the client and the agent, and between the agent and the server.

The C/A/S model has several advantages. It alleviates some of the impact of the limited bandwidth and poor reliability of wireless links by constantly maintaining the client's presence on the network via the agents. The agent splits the interaction between the mobile client and fixed servers into two parts, one between the client and the agent and one between the agent and the server. Data transmission can be optimized in the middleware so the QoS of data transmission improves with lower cost computation in the middleware or agent. A security wrapper in the middleware can provide data security over the wireless network.

Though the client-agent-server model offers number of advantages, it fails to sustain the current computation at the mobile client during periods of disconnection. In addition, the agent can directly optimize only data transmission over the wireless link from the fixed network to the mobile client but not in the opposite direction. In our research, we focus on an enterprise database, which is located at a remote server and connected to the middleware through the Internet.

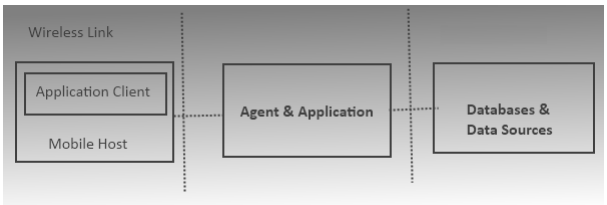


Figure 1: Client-Agent-Server (C/A/S) model with remote database.

The C/I/S model proposes the deployment of an agent that will run at the mobile device along with an agent that will run in the server side or middleware (Fig. 2). This client-side agent intercepts the client's requests and together with the server-side agent performs optimizations to reduce data transmission over the wireless link, improve data availability and sustain the mobile computation. From the point of view of the client, the client-side agent appears as the local server proxy that is co-resident with the client. Since the pair of agents is virtually inserted in the data path between the client and the server, the model is also called C/I/S instead of C/A/S.

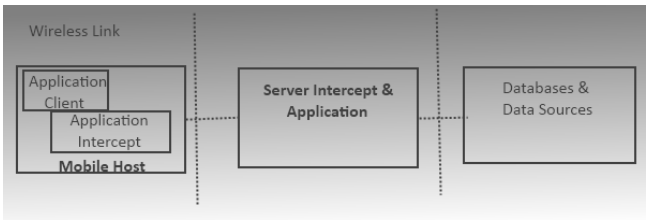


Figure 2: Client-Intercept-Server (C/I/S) model with remote database.

4 Implementation

To evaluate the different models, we developed: two remote databases, a middleware API, a mobile agent API and mobile applications for the two different domains. The middleware API provides a web service which returns data in XML format, so that any mobile application can retrieve the data and use it. The API has the flexibility to retrieve data from databases through the middleware and make use of caching, compression, encryption or any combination of these technologies. The Mobile Agent API was developed using J2ME and provides functionalities for retrieving cache data, decompressing data, and decrypting data coming from the middleware API.

In the data transmission lifecycle, the mobile application sends a request to the middleware; the middleware retrieves data from the cache or executes the query on the remote database and processes the data; after processing the middleware returns data to the mobile application. Finally, the mobile application processes the data returned from the middleware and displays it. Figure 3 presents a high level overview of the data transmission lifecycle.

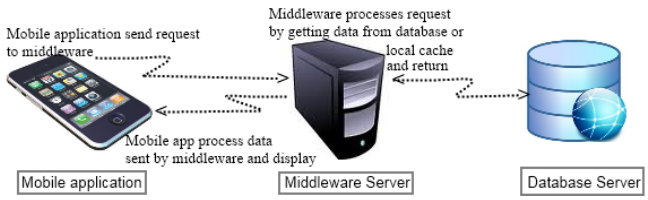


Figure 3: High level overview of the system.

In the C/A/S model, the mobile application sends requests to the middleware and the middleware processes the data based on the requests from the mobile application. If the mobile application requests that data be cached, the middleware searches for the data in the cache. If the data is found, the middleware returns it back to the mobile application. If the data is not in the cache, the middleware retrieves the data from the database, caches it and sends it

back to the mobile client. After getting the data from the middleware, the mobile application processes the XML formatted data from the middleware, and then displays it on the mobile screen.

Figure 4 illustrates the request and response cycle for the C/A/S model. In case of C/I/S, the middleware and the database are same as in the C/A/S model. However, the mobile application does not directly communicate with the middleware or processes returning data from the middleware; rather, it calls the Intercept. Instead of calling web service, the mobile application calls methods of the mobile API which does any local computation and returns data to the application to display it. Figure 5 illustrates the request and response interaction of the C/I/S model.

Databases: We experimented with a database containing medical information of around 10,000 patients (with fictitious names and identifying information) The database contains several tables of hospital information, patient information, patient in hospital information, and the diagnosis results of patients. The database contains approximately ten thousand patient results. Using the patient data we can do different experiments.

Middleware: The middleware retrieves data from the remote databases and returns it back to the mobile application. The middleware is developed using C# programming language in an ASP.NET platform. We experiment with caching, compression and encryption within the middleware to understand their impact on data transmission in a wireless network. We report only on compression and encryption in this paper. The mobile software developer can use the API to specify whether data should be cached or compressed or encrypted or any combination of these techniques by the middleware, just by calling the web services provided by the API. The API provides a generic interface to a middleware platform and the mobile application developers can customize the use of these techniques to meet the need of their application to retrieve data.

Compression: Data compression involves encoding information using fewer bits than the original representation. Compression is useful because it helps reduce the consumption of resources such as data space or transmission capacity. As the bandwidth of wireless network is scarce, it may be advantageous to compress data to get the maximum out of the bandwidth. Mobile applications suffer from limited memory where data compression may help save memory and at the same time improve the speed of data transfer.

Lossless compression is mainly used for spreadsheets, text and executable program compression; on the other hand, lossy compression is mainly used for image, video, and audio compression. Lossless data compression is used in many applications such as the ZIP file format and in the UNIX tool gzip. Lossless compression is used in cases where it is important that the original and the decompressed data be identical, or where deviations from the original data could be deleterious. Typical examples are executable programs, text documents and source code.

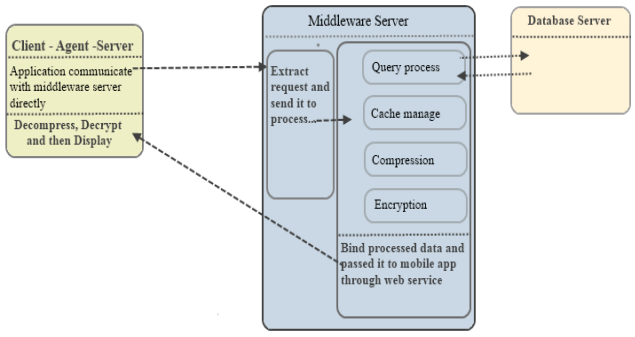


Figure 4: Client-Agent-Server request and response.

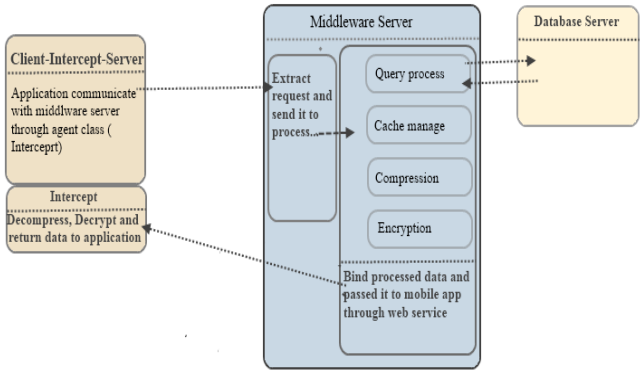


Figure 5: Client-Intercept-Server request and response.

In the middleware, we used gzip compression; it is based on the DEFLATE algorithm, which is a combination of Lempel-Ziv (LZ77) and Huffman coding. Gzip contains a 10-byte header, containing a magic number, a version number and a time stamp; optional extra headers, such as the original file name; a body, containing a DEFLATE-compressed payload; and an 8-byte footer, containing a CRC-32 checksum and the length of the original uncompressed data. To implement gzip compression in the middleware, we used SharpZipLib [17], an ASP.NET compression library that supports Zip files using both stored and deflate compression methods.

Encryption: In the case of enterprise databases, it is often necessary to move data over networks – to other sites or to remote desktop and mobile applications. Data transmission across networks, particularly public networks, creates potential security problems [17]. Given the importance of data security, we have implemented encryption in the middleware so data transmission in the wireless network can become secured.

We encrypt the data to Base64 data format before transmitting it to wireless network. Instead of using a real encryption algorithm for our experiments, we used Base64 encoding to emulate encryption processing; in reality this encoding scheme is much less complex than a real encryption scheme. Base64 is a group of encoding schemes that represent binary data in an ASCII string format by translating it into a radix-64 representation. Base64 encoding schemes are commonly used when there is a need to encode binary data

that needs be stored and transferred over media that are designed to deal with textual data. This is to ensure that the data remains intact without modification during transport. Base64 is commonly used in a number of applications including email via MIME, and storing complex data in XML.

5 Experiments and Results

Our experimental environment consisted of the following components: A remote database server hosting two databases; A middleware server (providing web service); Two mobile applications (C/A/S application and C/I/S application). Table I summarizes the tools used for the development of experimental systems. The following provides an overview of the experimental environment in more detail.

Table I. List of development tools for the experimental environment.

Component	Development Tool
Remote database Server 1	MySQL Database
Remote Database Server 2	MySQL Database
Middleware API	ASP.NET web service
Mobile Application C/A/S	J2ME
Mobile Application C/I/S	J2ME

Database Server: The databases are hosted on a remote server (www.godaddy.com), accessible over the Internet through a public IP address. It runs Linux and MySQL server, version 5.0. The middleware communicates with the remote database through the public IP address through the Internet.

Middleware Server: The middleware was developed using ASP.NET web service and C#. The middleware server runs Internet Information Services (IIS) 7 server. The configuration of the server is: Windows 7 32-bit operating system, Intel core duo 2.3 GHz processor and 4 GB memory. It is accessible through the Internet through a public IP address.

Mobile Application: Our mobile applications were implemented on a Nokia E72 smart phone; the configuration of the mobile phone is: Symbian OS 9.3, 600 MHz CPU, 250 MB Internal memory, and 128 MB RAM. The mobile application connects to the middleware services through wireless (WiFi). Testing was done from a home residence connection which had a 2Mbps dedicated wireless connection.

5.1 Basic Experiments with a Medical Database

We developed a Medical Information Application for the mobile device. Using this application, a user (e.g., doctor, nurse, etc.) can log in and search a patient database by last name or first name or any part of the name. The search result returns a list of patients along with their unique identity number, name, age, and sex. Selecting a patient from the list returned, will cause the application to search the database and return the medical history and previous diagnosis results of the patient. Using the Medical Application, we carried out three experiments:

- Experiment 1: Basic experiment: compare the two architectures for two different "data" scenarios.
- Experiment 2: Same scenarios as basic experiment using compression.
- Experiment 3: Same scenarios as basic experiment using encryption.

For each experiment, the same retrievals and same steps were used:

- Step 1: User logs in using their user name and password; saved in the database along with permissions to check a patient's medical information.
- Step 2: After successful login, the user can search for a patient by first name, last name, or any part of the name, for example: 'Henry', 'John', or 'Jo'.
- Step 3: Based on the search key, data is retrieved from the remote database; basic information is displayed: Patient Identification Number, Name, Age and sex. The total time taken in milliseconds to retrieve data from database is recorded along with total number of records found and the size of the returned data in bytes.

In the basic experiment, the mobile application requests, through the middleware, data from the remote database; the middleware retrieves the data and sends it to the mobile application through the web service in XML format. The mobile application extracts the retrieved data from the XML formatted data and display it on the mobile screen. We execute the experiment for three different scenarios involving different size data results. The scenarios with the data sizes are presented in Table II.

Table II. Scenarios and Data Sizes for Experiments.

Scenario 1	Search key for patient search is 'Henry'	Total patients found: 17, Data Size: ~2KB
Scenario 2	Search key for patient search is 'Li'	Total patients found: 397, Data Size: ~45KB
Scenario 3	Search key for patient Search 'T'	Total patients found: 1850, Data Size: ~230KB

Table III. Experiment 1 Results.

Scenarios	C/A/S (SD[1])	C/I/S (SD)	Comments
Scenario 1	525.67 ms (31.39)	577.67 ms (24.22)	C/A/S little faster than C/I/S but negligible difference.
Scenario 2	856.00 ms (24.64)	1073.00 ms (37.26)	C/A/S faster than C/A/S .
Scenario 3	-----	-----	Failed; unable to complete.

We replicate each experiment three times. The experimental results and analyses are presented in Table III.

[1] SD = Standard Deviation

The application on our test mobile phone could not handle the data of the Scenario 3 (returns around 230KB data from middleware) to display. We compared the performance of the two architectures for each of the scenarios using a t-test. For both Scenario 1 and Scenario 2, there is a significant difference between C/A/S and C/I/S at a 95% confidence level, that is, C/A/S performs better.

Table IV. ANOVA Results for Experiment 1.

Software Architecture (SA)	Scenarios (S)	SA+S	%Error
8.94%	84.18%	3.36%	3.52%

Further analysis using an analysis of variance (ANOVA) was also performed and the percentage of variation explained by the factors is presented in Table IV. The Software Architectures (SA) explained 9% of the variation, the Scenarios (S) 84% and the interaction of these (SA+S) was 3%. The error also explained about 3% of the variation. Clearly, the different scenarios had the greatest impact on the variation in results while the variation attributable to the different architecture was minor.

Summary of Experiment 1: Data size has the greatest impact on transmission time, though C/A/S performs better than C/I/S.

5.2 Compression Experiments

In Experiment 2, the scenarios are the same as Experiment 1, but here, we compress data in the middleware. For this experiment, the results are mixed: the C/A/S model performs better than the C/I/S model for Scenario 1, but the C/I/S model performs better for Scenarios 2 and 3. Note that in this case the use of compression enabled Scenario 3 to work successfully, unlike the case without compression. The results for the compression experiment are shown in Table V.

Table V. Experiment 2 Results.

Scenarios	C/A/S (SD)	C/I/S (SD)	Comments
Scenario 1	478.67 ms (21.08)	545.00 ms (45.92)	C/A/S little faster than C/I/S but negligible difference.
Scenario 2	875.67 ms (41.02)	817.67 ms (39.72)	C/I/S faster than C/A/S; difference is small.
Scenario 3	2365.00 ms (39.92)	2190.67 ms (63.36)	C/I/S faster than C/A/S; difference is small.

As with Experiment 1, a t-test at a 95% confidence level shows a significance difference between the two architectures for Scenario 1. However, for Scenarios 2 and 3, at a 95% level there was no significant difference between the two architectures.

Table VI. ANOVA Results for Experiment 2.

Scenario (S)	Compression (CO)	S+CO	Error
82.38%	7.96%	3.91%	4.57%

A summary of the main results of the ANOVA (Table VI) shows that the percentage of variation explained by the factors is: Scenarios (S) – 82.38%, Compression (Co) – 7.96%, the interaction of these (S+Co) – 3.91%, and the error explained about 4.57%. Clearly the data size (different scenarios) again has the greatest impact on the variation in performance results, though compression has some impact; variation attributable to the different architectures is minor.

Summary of Experiment 2: Not surprisingly, compressing and decompressing the data requires more time. Data size (scenarios) is still the most significant factor.

5.3 Encryption Experiments

In Experiment 3, the scenarios are the same as before, but here, we encrypt data in the middleware. As with Experiment 1, at a 95% level the C/A/S model performs better than the C/I/S model for Scenarios 1 and 2; again, the data transfer in Scenario 3 failed. Encryption tends to increase data size, even for the simplistic method used in our experiments. The data in Scenarios 1 and 2 changed, respectively, from 2Kb to 3Kb and from 45Kb to 60Kb. The results for the encryption experiment are shown in Table VII.

Table VII. Experiment 3 Results.

Scenarios	C/A/S (SD)	C/I/S (SD)	Comments
Scenario 1	669.33 ms (15.95)	806.00 ms (81.07)	C/A/S better but negligible difference
Scenario 2	1134.00 ms (46.81)	1236.00 ms (15.72)	C/A/S better but negligible difference
Scenario 3	-----	-----	Failed; unable to complete.

Table VIII. ANOVA Results for Experiment 3.

Scenario (S)	Encryption (En)	S+En	Error
83.86%	7.96%	1.44%	4.87%

A summary of the main results of the ANOVA (Table VIII) shows that the percentage of variation explained by the factors is: Scenarios (S) – 83.86%, Encryption (En) – 7.96%, the interaction of these (S+En) – 1.44%, and the error explained about 4857%. Clearly the data size (different scenarios) again has the greatest impact on the variation in performance results; though encryption has some impact. The variation attributable to the different architectures is again minor.

Summary of Experiment 3: The C/A/S model performed better; data size (scenarios) is still the most significant factor.

6 Discussions and Directions

We examined the performance of the C/A/S and C/I/S models because they had been identified as the most plausible models for data-oriented mobile applications by previous researchers though primarily in the context of browsers. One

objective of our research was to investigate whether one performed better than the other and for which scenarios and techniques. Based on our research results, we conclude that the C/A/S model generally performs better than the C/I/S model, though not in all situations. This is consistent with previous findings of Spyrou et. al [3] and others. However, an analysis of variance shows that the differences in execution times stems from the different scenarios (size of data retrieved by the application) and caching (when used).

Though the C/A/S application performs better than the C/I/S application in our experiments, the magnitude is still relatively small and based on our experiences in using both architectures for implementation of the application, there are other reasons to consider the C/I/S model:

- In the C/I/S model, the communication is through the *Intercept* API. In our experience, using the API makes mobile programming simpler and may be an advantage when developing application for heterogeneous mobile operating systems.
- Using the C/I/S model provides reusable code, which can help improve the quality of the product. So even if in some circumstances the C/I/S model quantitatively requires a little extra transaction time, it may qualitatively improve the system.

A number of future directions exist based on this work:

- Future work could look at extending the C/I/S model for client to utilize background threading in the intercept when retrieving or processing data from the middleware.
- Our experiments used a Nokia E72 smart phone which runs the Symbian Operating System. It would be useful to evaluate the APIs by porting them to other mobile operating systems and developing mobile applications. There are many new and different mobile devices, and testing the APIs with other types of smart phones tablets, etc would be useful.
- With other devices, it would be useful repeat some of these experiments to see if similar results can be obtained. Newer devices have more capabilities, faster processors, more memory so the absolute times may be faster or data sets larger. It would be interesting to see if the impact in performance of these techniques follows a similar pattern.

References

[1] Rocha B P S, Rezende C G, Loureiro A A R: "Middleware for multi-client and multi-server mobile applications"; 2nd International Symposium on Wireless Pervasive Computing, 2007. ISWPC '07, doi 10.1109/ISWPC.2007.342643.

[2] Swaroop V, Shanker U: "Mobile distributed real time database systems: A research challenges"; 2010 International Conference on Computer and Communication Technology (ICCCT), 2010, pp. 421-424, doi 10.1109/ICCCT.2010.5640495.

[3] Spyrou C, Samaras G, Pitoura E, Evripidou P: "Mobile agents for wireless computing: the convergence of wireless computational models with mobile-agent technologies". 2004, Mobile Networks & Applications.

[4] Rahaman, M. A. and Bauer, M. "A Comparison of Software Architectures for Data-Oriented Mobile Applications". Sixth Joint IFIP/IEEE Wireless and Networking Conference, 2013, pp. (to appear).

[5] Capra L, Emmerich W, Mascolo C: "Middleware for mobile computing: Awareness vs. transparency". Proceedings of the Eighth Workshop on Hot Topics in Operating Systems, 2001. HOTOS '01. Washington, DC, USA: IEEE Computer Society, 2001, p. 164, doi:10.1109/HOTOS.2001.990080.

[6] Mascolo C, Capra L, Emmerich W: Middleware for mobile computing (a survey). In Tutorial Proceedings of the International Conference of Networking 2002. Springer, 2002, pp. 20–58.

[7] Capra L, Blair G S, Mascolo V, Emmerich W, and Grace P: Exploiting reflection in mobile computing middleware. SIGMOBILE Mobile Computing Communication. Rev., vol. 6, no. 4, pp. 34–44, 2002.

[8] Campbell A T: Mobiware: Qos-aware middleware for mobile multimedia communications. In Proceedings of the IFIP TC6 seventh international conference on High performance networking, HPN '97, VII. London, UK, UK: Chapman & Hall, Ltd., 1997, pp. 166–183.

[9] Bellavista P, Corradi A, and Stefanelli C: Mobile agent middleware for mobile computing. Computer, vol. 34, no. 3, pp. 73–81, 2001.

[10] Chan A T, Chuang S N: Mobipads: A reflective middleware for context-aware mobile computing. IEEE Transactions on Software Engineering, vol. 29, no. 12, pp. 1072–1085, December 2003, doi:10.1109/TSE.2003.1265522.

[11] Rocha B P S, Rezende C G, Loureiro A A R: Middleware for multi-client and multi-server mobile applications. 2nd International Symposium on Wireless Pervasive Computing, 2007. ISWPC '07

[12] Gupta S, Joshi A, Santiago J, Patwardhan A: Query distribution estimation and predictive caching in mobile ad hoc networks. In Proc. of MobiDE, 2008, pp. 24–30

[13] Yin L and Cao G: Supporting cooperative caching in ad hoc networks. IEEE Transactions on Mobile Computing, vol. 5, no. 1, pp. 77–89, Jan. 2006, doi:10.1109/TMC.2006.15.

[14] Cheluvaraju B, Kousik A S R, and Rao S: Anticipatory Retrieval and Caching of Data for Mobile Devices in Variable-Bandwidth Environments. 5th Annual IEEE International Systems Conference (IEEE SysCon 2011), Montreal, Canada, April 2011.

[15] Huang J, Xiao Y, Liang Y: A Novel Secure Access Method for Remote Databases Based on Mobile Agents. Natural Computation, 2009. ICNC '09. Fifth International Conference on , vol.5, no., pp. 519-522, 14-16 Aug. 2009.

[16] Meier J D, Alex H, David H, Jason T, Prashant B, Lonnie W, Rob B J, Akshay B: App Arch Guide 2.0. [Online] http://apparchguide.codeplex.com/wikipage?title=Chapter 19 - Mobile Applications.

[17] SharpZipLib **ic#code:** The Zip, GZip, BZip2 and Tar Implementation For .NET [Online] http://www.icsharpcode.net/opensource/sharpziplib/.

[18] Greenberg M S, Byington J C, Harper D G**:** Mobile agents and security. IEEE Communications Magazine, Vol.36, July 1998, pp. 76~85.

Locally Complementary Multi-Path Routing in a MANET

Toshiaki Kagoshima
Graduate School of Engineering, Soka University
1-236 Tangi-cho, Hachioji-shi, 192-8577 Japan
e11m5210@soka.ac.jp

Kazumasa Takami
Graduate School of Engineering, Soka University
1-236 Tangi-cho, Hachioji-shi, 192-8577 Japan
k_takami@t.soka.ac.jp

***Abstract*— MANETs (Mobile Ad hoc NETworks) are drawing interest because it can provide a means of communication when an existing public network has been damaged by a disaster, or in areas where there is no fixed-line network available. For audio/video communication to be feasible in a MANET, the routing method used must provide high real-time performance. This paper proposes a locally complementary multi-path routing protocol, which can reduce the size of the network section for which route restoration is required in the event of a disaster. This makes it possible to achieve high real-time performance and maintain existing connections, thereby avoiding interruptions in audio/video communication. Specifically, this protocol secures a spare route for every two hops on the main route. Two alternative methods for securing spare routes are presented: an independent route securing method and a simultaneously route securing method. These as well as AODV (Ad hoc On-Demand Distance Vector) have been compared in terms of the number of control packets generated. The simultaneous route securing method has been built into a MANET emulator, and the percentage of successful establishment of the initial routes has been assessed.**

Keywords— MANET; multi-path routing protocol; SIP; securing spare route; AOMDV

I. INTRODUCTION

MANETs (Mobile Ad hoc NETworks) are drawing interest because it can provide a means of communication when an existing public network has been damaged by a disaster, or in areas where there is no fixed-line network available [1]. A MANET is made up of mobile terminals. Data are transferred between two terminals through a route involving multiple hops. The situations in which the network is economically viable are so limited that actual applications have not been developed. In the world of the Internet, connection-oriented applications (for audio/ video communication, etc.) have been developed based on SIP (Session Initiation Protocol) [2]. Provision of SIP-based services in a MANET should encourage many applications for a MANET to be developed. Although many studies have been made on handling SIP services in a MANET [3]-[9], a lack of experimental tools available has limited their scope to handling only the processes from a session establishment request to the start of the session. The authors have developed and extended a MANET emulator for use as a service experiment tool for a MANET [10]-[16].

To make audio/video communication feasible in a MANET, the routing method used must provide high real-time performance. In a MANET, a connection between SIP clients may be broken for a number of reasons: movements of the terminals involved in the connection, loss of wireless links, and running out of terminal batteries.

Several multi-path routing methods based on AODV (Ad hoc On-Demand Distance Vector) [17] have been proposed, such as AOMDV (Ad hoc On-Demand Multipath Distance Vector) [18] and the route disjoint protocol [19]. However, since the route disjoint protocol selects candidate paths for re-routing from the entire network, it takes considerable time in reconfiguring the network, a feature detrimental to real-time performance. A more reliable routing method is required if SIP services are to be provided in a MANET.

This paper proposes a locally complementary multi-path routing protocol, which can reduce the network section for which route restoration is required, in order to achieve high real-time performance, and make it possible to maintain existing connections, thereby avoiding interruptions in ongoing audio/ video communication. Section II reviews related studies on multi-path routing. Section III proposes a locally complementary routing protocol, which features high real-time performance, and two alternative methods of securing spare routes: an independent route securing method and a simultaneous route securing method. Section IV describes an evaluation system in which the simultaneous route securing method was implemented. Section V evaluates this method in terms of the percentage of successful establishment of the initial routes by conducting experiments on this evaluation system with different network models. In addition, the two spare route securing methods as well as AODV are compared in terms of the number of control packets generated.

II. RELATED WORK

A. AOMDV

AOMDV is a routing protocol that allows multiple non-overlapping routes to be secured between the originator and destination terminals, as between nodes S and D in Fig. 1. However, Node C in Fig. 1 is involved in two routes, and thus is an Achilles' heel. If Node C comes down, both routes are lost, and communication between the two end terminals is

disrupted. Therefore, it is necessary to avoid creating such critical nodes if real-time communication is to be provided.

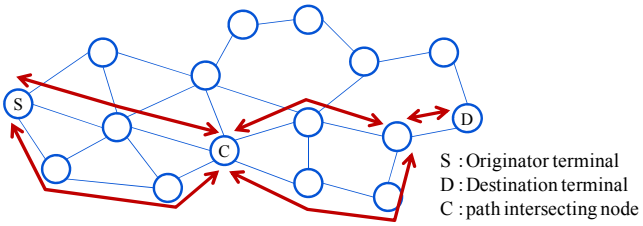


Fig. 1. Securing multiple routes with AOMDV.

B. Route disjoint protocol

Unlike AOMDV, the route disjoint protocol secures multiple routes without creating an Achilles' heel node. As shown in Fig. 2, the nodes of the two routes are disjoint. The downsides of this protocol are that it is necessary to look at the entire network to reconfigure routes, and that route restoration takes a long time because, when a link failure has been detected, Terminals S and D must be notified of it before the routes concerned are switched. It is necessary to reduce route restoration time if real-time communication is to be provided.

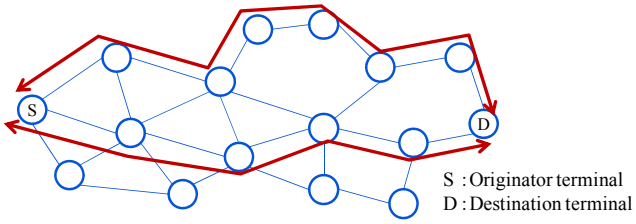


Fig. 2. Routes secured by the route disjoint protocol.

III. LOCALLY COMPLEMENTARY MULTI-PATH ROUTING

A. Overview

This section proposes a locally complementary multi-path routing protocol. It mitigates the disadvantages of AOMDV by reducing the size of the network sections for which route restoration is required. It secures and holds a spare route for every two hops on the main route in order to allow ongoing communication to be maintained even during route restoration. Since all nodes are mobile, this protocol is based on AODV, which is a reactive protocol. Examples of the main route and spare routes are shown in Fig. 3.

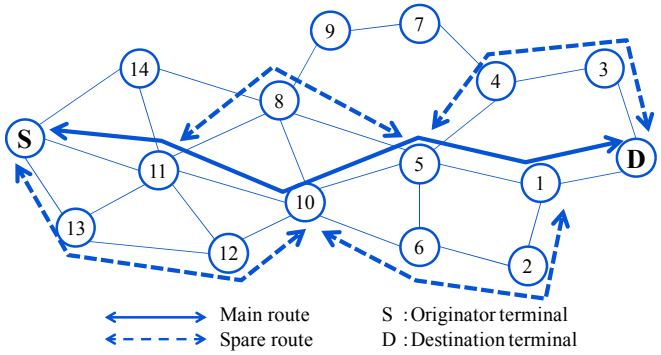


Fig. 3. Examples of the main route and spare routes.

Since a spare route is secured for each two hops of the main route, a route can be restored in a short time. Only up to 3 hops are allows for a spare route. For example, a spare route can be S-13-12-10 in Fig. 3. While AOMDV can involve Achilles' heel nodes in spare routes, the locally complementary multi-path routing protocol does not. Even when node 10 in Fig. 3 fails, for example, communication can be quickly restored because spare route 11-8-5 is available. Unlike the route disjoint protocol, which examines the entire network for route restoration, the proposed protocol needs to look at only two hops in the main route. Therefore, the route concerned can be reconfigured in a short time.

Two alternative methods of securing spare routes at the time of initial route configuration are described in Subsections III.*B* and III.*C*. They are called an independent route securing method and a simultaneous route securing method. To secure spare routes at the time of route configuration, we have revised messages RREQ (Route REQuest) and RREP (Route REPly), and added SpareRREQ and SpareRREP, as shown in Fig. 4. All the four messages are of extended packet format. The shaded parts show the added parts. Why these parts are added and how they are used are described in the next subsection. The number within each parenthesis is the number of bits. Where there is no such parenthesis, the number of bits is 1. The meaning and use of each existing field is the same as in RFC3561 [17], and thus are not explained here.

Type(8)	J	E	G	D	U	Reserved(11)	Hop Count(8)
RREQ ID(32)							
Destination IP Address(32)							
Destination Sequence Number(32)							
Originator IP Address(32)							
Originator Sequence Number(32)							
Two Hop IP Address(32)							

(a) RREQ

Type(8)	R	A	Reserved(9)	Prefix Sz(5)	Hop Count(8)
Destination IP Address(32)					
Destination Sequence Number(32)					
Originator IP Address(32)					
Lifetime(32)					
Two Hop IP Address(32)					

(b) RREP

Type(8)	J	E	G	D	U	Reserved(11)	Hop Count(8)
RREQ ID(32)							
Destination IP Address(32)							
Destination Sequence Number(32)							
Originator IP Address(32)							
Originator Sequence Number(32)							
Spare Destination IP Address(32)							
Spare Originator IP Address(32)							
Relay NG IP Address 1 (32)							
Relay NG IP Address 2 (32)*							

* Not needed in the independent route securing method

(c) SpareRREQ

Fig.4. Extended control packet formats(1/2).

Type(8)	R	A	Reserved(9)	Prefix Sz(5)	Hop Count(8)
Destination IP Address(32)					
Destination Sequence Number(32)					
Originator IP Address(32)					
Spare Destination IP Address(32)					
Spare Originator IP Address(32)					
Relay NG IP Addrress(32)					
Lifetime(32)					

(d) SpareRREP

Fig. 4. Extended control packet formats(2/2).

B. *Independent route securing method*

This method secures spare routes by flooding the network with a SpareRREQ after the main route has been secured. It secures the main route in the same manner as AODV. When a node on the main route receives an RREP, it writes a relevant value into the *Two Hop IP Address* field of the RREP because it wants to send a SpareRREQ to nodes two hops away. In the example of Fig.3, when Node 1 has received an RREP from Node D, it writes the address of Node D into its *Two Hop IP Address* field, and sends the message to Node 5. Node 5 writes Node D's address into the *Spare Destination IP Address* field, its own address into the *Spare Originator IP Address* field, and Node 1's address in the *Relay NG IP Address* field of a SpareRREQ, and floods the network with the message with TTL=3. Although Node 1 receives this SpareRREQ, it does not rebroadcast it because the value in the *Relay NG IP Address* field is its own address. Consequently, this SpareRREQ is relayed on a route (5-4-3-D), which is separate from the main route. Node D sends a SpareRREP as a reply, and a spare route is secured. A problem with this method is that a large number of control packets are generated and flow in the network because each of Nodes S, 11, 10 and 5 broadcasts a SpareRREQ to secure a spare route.

C. *Simultaneous route securing method*

This method does not use SpareRREQ. It secures spare routes by sending a SpareRREP at the same time when the main route is secured. When an RREQ is broadcast using AODV, nodes receive multiple RREQ messages, and determine whether to discard them by examining the sequence number of each message. Flooding of the network with RREQs is shown in Fig.5, and the resulting routing table is shown in Fig.6. Each solid arrow in Fig.5 corresponds to one of the encircled parts in Fig.6, and indicates a route secured when Terminal S has flooded the network with an RREQ that is intended for Terminal D. Dotted arrows show routes that have been secured by RREQs that are to be discarded. In the simultaneous route securing method, spare routes are secured using RREQs that are to be discarded because they share the same sequence numbers. When SpareRREPs is unicast, the main route shown by solid lines and spare routes shown by dotted lines in Fig. 3 are secured. At the same time, relevant values are written into the *SpareDestIPAddr* field and the *SpareRelayIPAddr* field in the routing table so that the destination and the IP addresses of the relaying nodes of a spare route can be retained. The first node that has sent an RREQ is designated as the relaying node on the main route, and the second node that has sent an RREQ is designated as the relaying node on the spare route. For Node D, for example, Node 1 is the relaying node on the main route, and Node 3 is the relaying node on the spare route. A relevant value is written into the *Two Hop IP Address* field of the RREQ so that a node two hops back can be identified.

The algorithm for simultaneous route securing is described in the following, using Fig. 5. First, Terminal S floods the network with an RREQ. A node that has received this RREQ writes the address of its relaying node into the *Two Hop IP Address* field, and broadcasts the message. If the sequence number of the message is new, this node address is written into the routing table, and the message is rebroadcast. If the sequence number is the same as that of a previously received message, the node is registered as the relaying node of the spare route in the table. When Terminal D has received an RREQ, and if the sequence number of the message is new, it sends back an RREP to the secured route (Node 1). If the sequence number of the RREQ is the same as that of a previously received message, Terminal D sends back a SpareRREP. The node on the main route that has received the RREP writes its own IP address into the *Spare Destination IP Address* field, the IP address of a node two hops back into the *Spare Originator IP Address* field, and the IP address of the relaying node into the *Relay NG IP Address* field, of a SpareRREP, and sends the message to the relaying node on the spare route. The node that has received this SpareRREP sends it to the adjacent node if the IP address written in the *Spare Originator IP Address* field of the message is that of this adjacent node. If it is not, the node transfers the message to either the relay node on the main route or that on the spare node whichever has an IP address different from that written in the *Relay NG IP Address* field. Note that a spare route has been secured in the forward direction (direction towards Terminal D) when a SpareRREP is received, and in the backward direction (direction towards Terminal S) when a SpareRREP is sent.

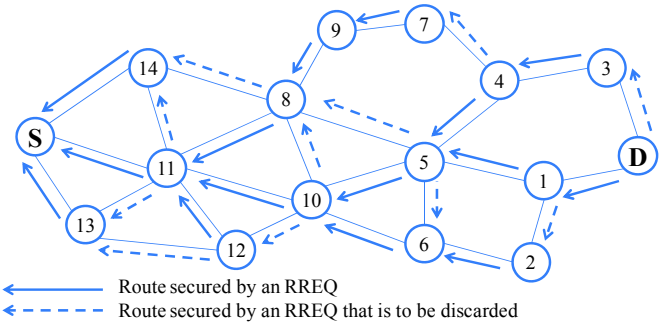


Fig. 5. Flooding with an RREQ.

Node	5			4			3		1			D		
Destination node	S	10	8	S	5	7	S	4	S	5		S	1	3
Relay node	10	10	8	5	5	7	4	4	5	5		1	1	3
Spare destination node (2hops back)	11			10			5		10			5		
Spare relay node	8			7					2			3		

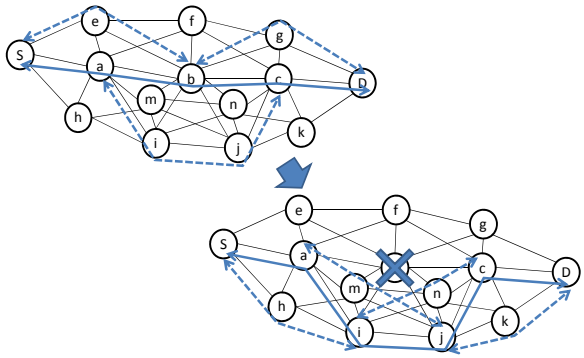
Fig. 6. Routing table created.

D. Route restoration method

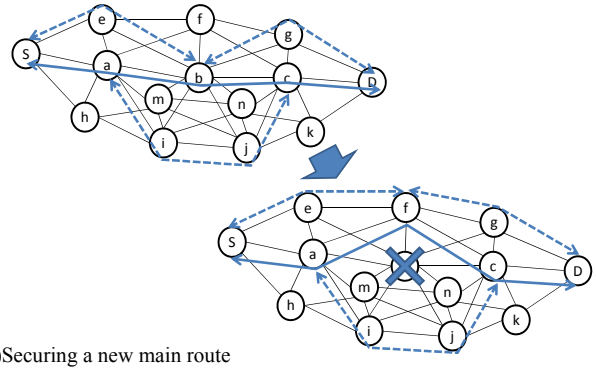
The main cause of a route being disrupted are a node dropping out of the network due to its battery running out, and a link being disrupted due to movements of a node on the route.

1) In the case of a node dropping out of the network: Examples of securing the main route and spare routes are shown in Fig. 7. When a node (e.g., Node b) on the main route drops out of the network, two alternative methods can be conceived to secure a new main route. Method 1 is to replace a part of the previous main route with its spare route, as shown in Fig. 7(a). Method 2 is to secure a new main route, as shown in Fig. 7(b). If the number of hops of the spare route that substitutes for the route provided by the dropout node is 2, the results of Methods 1 and 2 are more or less the same. However, if the spare route involves 3 hops, Method 1 increases the number of hops on the main route. This also means that the number of spare routes that stand by also increases. Not only do the spare routes become more redundant but also many control packets are generated and flow in the network to secure the spare routes. In contrast, with Method 2, if it is found that Node f can substitute for Node b, only Node f broadcasts a SpareRREQ. When Terminals S and D send back a SpareRREP, the spare route is restored. To sum up, when a node on the main route drops out, a new node that can substitute for the dropout node is searched for, and a different main route is secured anew using this node.

When a route on a spare route drops out, a new spare route is secured.



(a)A spare route is used to secure a new main route.



(b)Securing a new main route

Fig. 7. Examples of securing the main route and spare routes.

2) In the case of a link being disrupted: When a link on the main route is disrupted, and the two nodes connected to this link attempt to establish a new link, the number of hops between the two nodes will increase from 1 to 2. In addition, when the relevant nodes, including the node newly incorporated into the main route, try to secure a spare route for the new link, excessively many control packets will be generated in the network. However, if the above process is considered as an attempt to secure a 3-hop spare route to replace the original 2-hop spare route, all that is needed is to re-secure spare routes for the newly added node. If this process is continued, many redundant routes may be created. Therefore, it is necessary to refresh routes at certain intervals or at the trigger of a certain event.

When a link between two nodes on a spare route is disrupted, all that is required is to secure a new spare route between the spare originator node and the spare destination node.

E. Data transmission method

A node on the main route sends data to the relay node on the main route and that on the spare route while a node on the spare route sends data to the relay node on the spare route. The *Main or Spare* field is used to determine whether a particular route is the main route or a spare route. Since real-time communication is assumed, old packets can be discarded. There is no need to retransmit packets. Data packets with the same value in their *Data Sequence Number* fields are sent to both the main route and the spare route. If a node on the main route judges that the received packet is an old one, it regards it as a duplicate packet and discards it. An extended data packet format defined to provide the above function is shown in Fig.8. The meanings and purposes of other fields are the same as those of the corresponding fields in RFC3561 [17], and thus the description of these are omitted here.

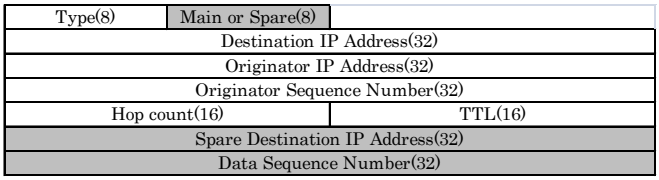


Fig. 8. Extended data packet format.

IV. Development of an Evaluation System

The proposed multi-path routing protocol was implemented on an already developed MANET emulator [11] using VC++6.0. The system configuration of the revised MANET emulator is shown in Fig. 9. The proposed protocol was implemented by extending AODV of the routing module [Routing]. In addition, the emulator can simulate the remaining battery level [Battery], the interface with the monitor [Monitor IF], conversion between the virtual and real IP addresses [CNV], a mobility model [Movement], the MAC (Media Access Control) layer [MAC], radio coverage [Zone], etc. Module [EN] simulates inter-SIP client communication within the emulator.

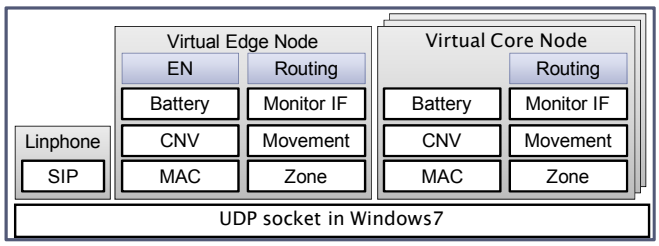


Fig. 9. System configuration of the revised MANET emulator.

V. EVALUATION

A. *Initial spare route configuration success ratio for the simultaneous route securing method*

The ratio of successful spare route configuration attempts was measured for the different network models with different numbers of hops on the main route, shown in Fig.10. The result is shown in Fig.11. If there are *N* hops on the main route, *N-1* spare routes can be secured. The *Initial spare route configuration Success Ratio* (ISR) was calculated using Eq. (1). It was considered successful if *N-1* spare routes were secured without sending SpareRREQs. In the evaluation experiment, five route configuration attempts were made for each model.

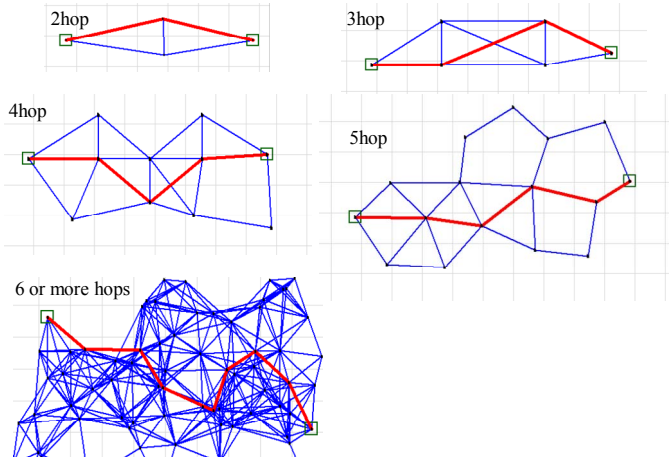


Fig. 10. Network models with different numbers of hops on the main route.

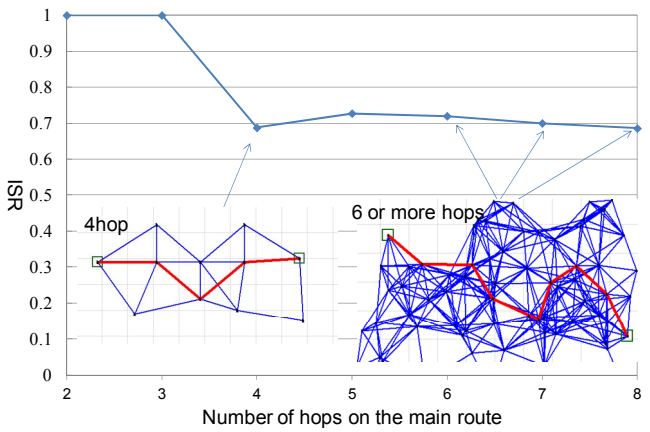


Fig. 11. ISR for the simultaneous route securing method.

$$ISR = \frac{N_A}{N_C} \quad (1)$$

where

N_A= Total number of spare route configuration attempts

N_C= Total number of spare route configuration candidates

Network models were fixed for networks with up to 5 hops. In the network model in which the number of hops is larger than 5, all nodes other than the originator and destination nodes were placed at random, and then the initial route configuration was executed. Network configuration examples for 4 hops and 6 or more hops on the main route are shown in Fig. 11.

In the network model of 4 hops, 3 spare routes can be secured. Among these, the number of spare routes that were secured only with RREQ, RREP and SpareRREP was counted. The initial spare route configuration situation for every attempt on 4 hop model is shown in Fig. 12. There are 3 candidate spare routes. Since 2 spare routes could be secured at first, second and 4th attempt, the ISR was calculated to 0.66. Also, at 5th attempt, 3 spare routes could be secured, and then the ISR was 1. At third attempt, the numbers of hops on the main route are 4 hops. However, 2 spare routes could be secured, and then the ISR was 0.5. Therefore, when five attempts were added together, the ISR was as follows.

$$ISR_{\_4hop} = (2+2+2+2+3) \div (3+3+4+3+3) = 0.68$$

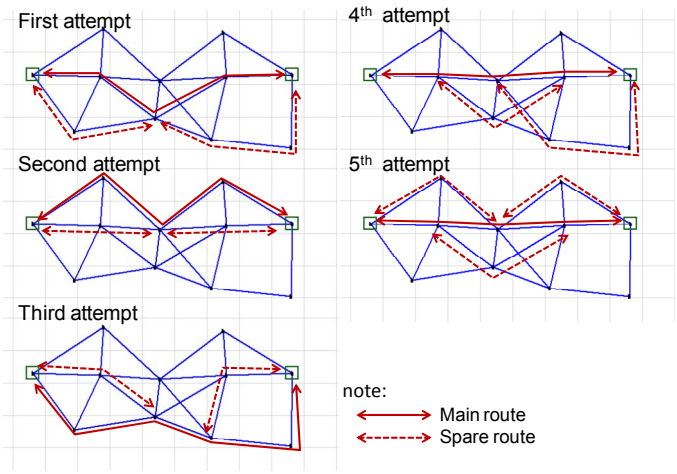


Fig. 12. Initial spare route configuration situation for every attempt on 4 hop model.

For the network models with 2 hops and 3 hops, the success ratio was 1 because the proposed algorithm can secure spare routes without fail. For network models with 4 or more hops, the average ratio of successfully securing spare routes was 0.68. This is because the proposed algorithm selects any second node that sent an RREQ as a relay node on a spare route, which may cause a relay node to be used for multiple spare routes. When this happens, not all spare routes can be secured successfully. A way to avoid a relay node being used for multiple spare routes and thereby to raise the ratio of successfully securing spare routes is to define a new response message to a SpareRREP, consider the third and fourth nodes that sent an RREQ as a candidate relay node on spare routes, and send a SpareRREP to the next node if there was no reply.

B. Comparison in terms of the number of control packets generated

The number of control packets sent during route configuration was counted. Different route securing methods are compared in terms of the number of control packets generated in Fig. 13. The network models used were those shown in Fig. 10.

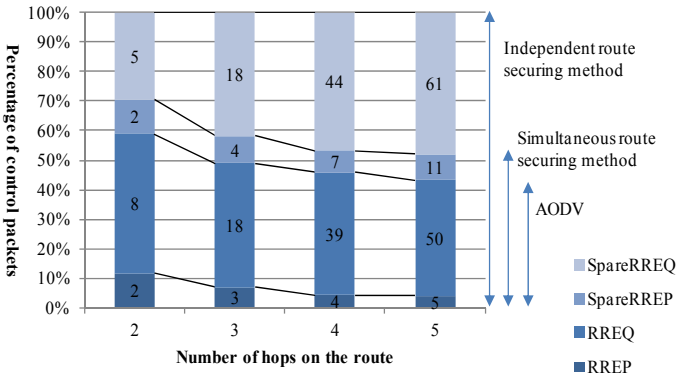


Fig. 13. Comparison in terms of the number of control packets generated.

With the independent route securing method, nodes on the main route flood the network with a SpareRREQ to secure spare routes. Therefore, as the number of hops on the main route increases, so does the number of control packets generated. With the simultaneous route securing method, spare routes are secured only with RREQs and SpareRREPs. Therefore, this method generates fewer control packets than the independent route securing method. It can be seen in Fig. 12 that spare routes can be secured by sending only as many SpareRREPs as are required.

VI. CONCLUSIONS

This paper has proposed a locally complementary multi-path routing protocol, which can reduce the size of the network section for which route restoration is required in the event of a disaster, in order to achieve high real-time performance and make it possible to maintain existing connections, thereby avoiding interruptions in ongoing audio/video communication. This protocol secures multiple routes by unicasting a SpareRREP for route configuration using AODV. Two alternative spare route securing methods have been presented: the independent route securing method and the simultaneous route securing method. These two methods as well as AODV were compared in terms of the number of control packets generated. The simultaneous route securing method can reduce the generation of control packets. This method was implemented in a MANET emulator to measure the initial route configuration success ratio. It was shown that the success ratio was about 70% even when the number of hops on the main route is 6 or more.

The issues that remain to be studied include implementation of SIP-based real-time communication using the proposed protocol to verify the protocol's feasibility in real communication.

REFERENCES

[1] S. Kumar Sarkar, T. G. Basavaraju, and C. Puttamadappa, "Ad Hoc Mobile Wireless Networks, Principles, Protocols and Applications," Auerbach Publications, Boston, MA, 2007.

[2] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," RFC Editor, 2002.

[3] C. Fu, R.H. Glitho, and R. Dssouli, "A novel signaling system for multiparty sessions in peer-to-peer ad hoc networks," IEEE Wireless Communications and Networking Conference, Vol. 4, pp. 2287–2292, 2005.

[4] S. Leggio, J. Manner, A. Hulkkonen, and K. Raatikainen, "Session Initiation Protocol Deployment in Ad-Hoc Networks: a Decentralized Approach," 2nd International Workshop on Wireless Ad-hoc Networks (IWWAN), London, May 2005.

[5] N. Banerjee, A. Acharya, and S. K. Das, "Enabling SIP-based sessions in ad hoc networks," Wireless Networks, vol.13 no.4, p.461-479, August 2007.

[6] S. Leggio, J. Manner, A. Hulkkonen, and K. Raatikainen, "Session initiation protocol deployment in ad-hoc networks: a decentralized approach," Proceedings of the International Workshop on Wireless Ad-Hoc Networks (IWWAN2005), London, UK, May 2005.

[7] N. Banerjee, A. Acharya, and S. Das, "Peer-to-peer SIP-Based Services over Wireless Ad Hoc Networks," Proceedings of the 2004 First Annual International Conference on Broadband Networks, San Jose, USA, October 2004.

[8] P. Stuedi, M. Bihr, A. Remund, and G. Alonso, "SIPHoc: efficient SIP middleware for ad hoc networks, Proceedings of the ACM/IFIP/USENIX 2007 International Conference on Middleware, November 26-30, 2007, Newport Beach, California.

[9] K. Balov, K. Kawagoe, and T. Nishimura, "SIP Deployment in Integrated Mobile Ad Hoc Networks: Centralized and Quasi-Decentralized Approaches," 11th IEEE International Conference on Advanced Communication Technology, pp.203-207, Feb. 2009.

[10] D. Kasamatsu, N. Shinomiya, and T. Ohta, "Routing Algorithm in Ad Hoc Networks," Proc. of Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT2005), 310-315, 2005.

[11] D. Okada, D. Kasamatsu, N. Shinomiya, and T. Ohta, "Programmable Ad hoc Networks," Proc. of International Conference on Software Engineering Advances (ICSEA2006), 6-4, 2006.

[12] D. Kasamatsu, N. Shinomiya, and T. Ohta, "A Broadcasting Method considering Battery Lifetime and Distance between Nodes in MANET," Proc. of International Workshop on Specialized Ad Hoc Networks and Systems (SAHNS2009), 1-1, 2009.

[13] D. Kasamatsu and N. Shinomiya, "Implementation and Evaluation of Emulator for Testing Service Programs in MANET," Proc. of International Workshop on Performance Analysis and Enhancement of Wireless Networks (PAEWN2010), 496-501, 2010.

[14] D. Kasamatsu, Y. Kawamura, M. Oki, and N. Shinomiya, "Broadcasting Method based on Topology Control for Fault-tolerant MANET," Proc. of International Workshop on Specialized Ad Hoc Networks and Systems (SAHNS2011), (in press), 2011.

[15] T. Kagoshima, D. Kasamatsu, and K. Takami, "Architecture and emulator in ad hoc network for providing P2P type sip_voip services," TENCON2011. Bali (Indonesia), PROGRAMME&ABSTRACTS. T02.01.5, pp.164-168. 2011.

[16] H. Todoroki, T. Kagoshima, D. Kasamatsu, and K. Takami, "Implementation of a Peer-to-Peer-type SIP client application on a MANET emulator," TENCON2012, Cebu Island (Philippines), PROGRAMME&ABSTRACTS. NC7-3, pp.1630-1650. 2012.

[17] C. E. Perkins, E. M. Royer, and S. R. Das, "Adhoc On-Demand Distance Vector Routing," RFC3561, http://www.ietf.org/rfc/rfc3561.txt.

[18] M. Marina and S. Das, "On-demand Multipath Distance Vector Routing in Ad Hoc Networks," IEEE ICNP,2001.

[19] K. Noguchi, T. Yamada, S. Koizumi, Y. Kotani, and J. Katto, "Experimental Evaluation of an On-demand Multipath Routing Protocol using Received Signal Strength for Ad hoc Networks," IEICE Technical Report, IN2006-4, May 2006.

A Sequence Frequency Reuse Scheme for Coordinated Multi-Point Transmission in LTE-A

Ying-Hong Wang
Dept. of Computer Science and Information Engineering,
Tamkang University,
Taiwan, R.O.C.
inhon@mail.tku.edu.tw

Chih-Hsiao Tsai
Dept. of Information Technology,
Takming University of Science and Technology,
Taiwan, R.O.C.
chtsai2104@gmail.com

Liang-Chain Chen
Dept. of Computer Science and Information Engineering,
Tamkang University,
Taiwan, R.O.C.
polo90169@hotmail.com

***Abstract*—**OFDMA communication technology becomes for the next generation mobile communication systems (4G) standards. Decreasing Inter-Cell Interference (ICI) is one of important issues in OFDMA. A new technology of LTE-A (Lone Term Evolution-Advanced) is CoMP(Coordinate Multi-Point, CoMP) transmission, which purpose is to solve low performance and poor quality of transmission due to interference cause by. Even through there are many LTE research papers are presented to explore the use of frequency reuse mechanism to solve the ICI, but they are inappropriate for CoMP technology. In this paper, we propose an efficiently Sequence Frequency Reuse (SeFR) scheme to improve low performance that cause by ICI, and making restrict in ICI seriously region to work smoothly on CoMP. According to the results of simulation, the proposed mechanism makes the CoMP transmission to achieve higher efficiency in the use of, and improves the SINR value of CEU (Cell Edge Users).

Keywords: CoMP, LTE-A, Frequency reuse, ICI

I、Introduction

In recent years, wireless networks and broadband networks have the majority of the market as well as the public high acceptance. In response to the needs of the general public, the constant evolution and progress of the wireless communication network are approaching to the next Internet generation. The aim is to provide user convenience and high-speed transmission. Orthogonal Frequency Division Multiplexing Access (OFDMA) communication technology is the standard of next generation mobile communication system [1-2]. Both of the standards of Institute of Electrical and Electronics Engineers (IEEE) 802.16e and Long Term Evolution (LTE) of the 3rd Generation Partnership Project(3GPP) select OFDMA as Downlink transmission scheme [1-2]. Due to the Cellular system network exists the problem of signal interference. OFDMA could be an effective solution to Intra-Cell Interference (ISI) by the characteristics of the orthogonal but Inter-Cell Interference (ICI) problem still exists, the cell-edge (CE) is particularly serious especially. ICI problem in general is solved by frequency reuse mechanism [8-10, 12, 16].

Including the Coordinated Multi-Point (CoMP) transmission is one of major technology characteristics in the evolution of LTE to the LTE-A process. The technological purpose of CoMP is to solve the low performance and poor quality of transmission caused by ICI. However, a number of frequency reuse mechanisms of LTE are not entirely support the CoMP, this new technology of the LTE-A. So that Jingya Li etc. proposed a Cooperative Frequency Reuse (CFR) [11] to apply CoMPinto frequency reuse mechanism.

Although the CFR can support CoMP technology, but it can not effectively use the CoMP technological characteristic. To use technically the CoMP to coordinate and allocate the resources among neighboring cells, it is necessary to design a more efficient frequency reuse mechanism. The purpose of our study is based on the CoMP technology to improve the existed frequency reuse mechanism and propose a suitable frequency reuse mechanism to enhance the overall effectiveness of the overall cell.

To accomplish this objective, this paper proposes an efficient Sequence Frequency Reuse (SeFR) mechanism to improve the CFR mechanism, ensure that the CoMP to be applied fully, and solve the low performance problem caused by interference between cells. The proposed SeFR offers different frequency allocation of resources in different regions of a cell and give the order of use restrictions [14], use different frequency reuse assignment to reduce the ICIs among the neighboring cells. So that the operations of CoMP not only can reach better results, but also will to avoid interference.

The remains of this paper is divided into four sessions. Session II presents the related technologies and researches, they include several different frequency reuse mechanisms. The proposed mechanism contains the definition and related processes are proposed detail in session III. Session IV simulates and analyzes SeFR with different frequency reuse mechanisms and make comparison. Finally, session V is conclusion summarizes the objective and contribution of the proposed SeFR mechanism and illustrates the future research directions.

II、 Related Work

A. Partial Frequency Reuse

The idea of Partial Frequency Reuse [9, 12] of all available spectrum resources F is split into two parts, denoted as F_1 and F_3, and F_3 is divided into three subset: $F_{3\text{-}A}$, $F_{3\text{-}B}$ and $F_{3\text{-}C}$, Cell Edge (CE) will be assigned to the resources of the F_3, Cell Edge Users (CEUs) can only use their own part of the F_3, Cell-Center users (CCUs) use only F_1, too. Shown as figure 2.1 [9].

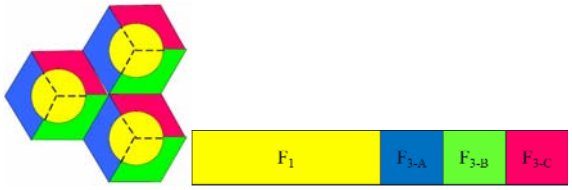


Fig. 2.1 PFR

B. Soft Frequency Reuse

Soft Frequency Reuse (SFR) [3] collects three neighboring cells into a set, all of the resources are available to the band will be divided into three sections: F_1, F_2, F_3, shown as figure 2.2. The CE of the three cells is allocated one of the three different sections, respectively.

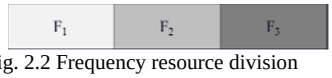


Fig. 2.2 Frequency resource division

Assume the CE in cell A is using the resource section, F_1, and the assignment to the Cell-Center (CC) of cell A are sections, F_2 and F_3 integrality. Figure 2.3 describes this assignment.

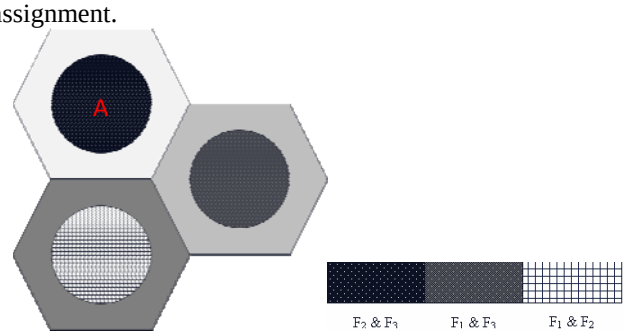


Fig. 2.3 SFR

C. Incremental Frequency Reuse

Incremental Frequency Reuse (IFR) [10] assigns also three neighboring cells as a set. It defines a different starting point to each cell in same set. When it will allocate resource, illustrated in figure 2.4 [10], black frame represents the starting position of each cell to allocate resources, so that can avoid ICI when traffic load is not heavy. However, the interference problem will be caused more serious than SFR or PFR when the load is increasing.

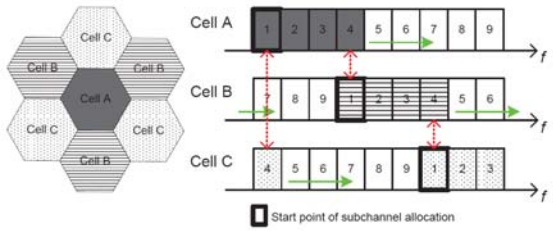


Fig. 2.4 IFR

D. Cooperative Frequency Reuse

Cooperative Frequency Reuse (CFR) is published in 2009, and proposed by VTC conference in 2010[3]. The mechanism divided into three steps.

The **first step** assumes that every three neighboring cells as a cluster. The CE regions will be separated according to six different positions of neighboring cells. The CE is cut into six zones, shown in figure 2.5. Each cell periphery is expressed as A_i^j. For examples:

Zone A_1^2 : The zone of cell edge Belongs to cell 1 and the possible interference of user is come from recent neighboring cell 2.

Zone A_2^3 : The zone of cell edge Belongs to cell 2 and the possible interference of user is come from recent neighboring cell 3.

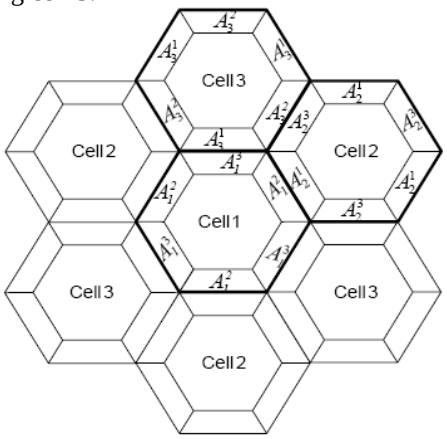


Fig. 2.5 Cell-Edge areas partition for each cell

The **second step** assigns the resource to users of each cell cluster according to the following rules:

Step 1 : All the resources in each cell is divided into two sets: G and F, G is assigned to the CCUs of each cell, while the resources of F is assigned to the CEU of each cell.

Step 2 : F will cut further into three equal portions, labeled F_1, F_2 and F_3, respectively. Any two F_i and F_j, $i \neq j$, are disjoint.

Step 3 : For each cell cluster, F_i of cell i is a cooperation frequency, which is make use of CoMP with CEUs of adjacent cells for coordinated multi-point data transmission.

Step 4 : F_j is assigned to the CEU, expressed as A_i^j. It is illustrated as figure 2.6.

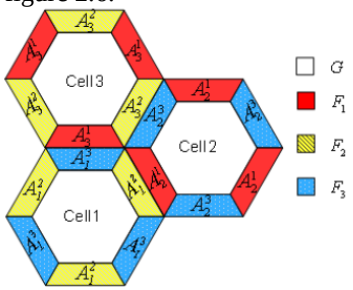


Fig. 2.6 Frequency allocation for each cell in the cluster

A CEU who locates in the heart of A_i^j, the cell i and cell j are classified in the CCS (CoMP Cooperating Set). It means both of cell can execute CoMP cooperation transmission for the CEU. It is the cell i and cell j could joint transmission in the same frequency resources. As show in figure 2.7, the CEU1 is position in the A_1^3, it can be regarded as CoMP CEU, the original service cell 1 and the original interfered cell 3 will be classified together in the CCS

collection.

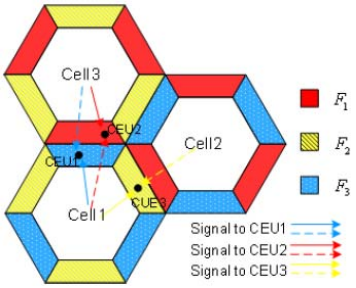


Fig. 2.7 CoMP joint transmission in the CFR system

The **third step** uses the algorithm as follows, proposed by G. Piñero [15] to the implementation of the CoMP Cooperating Set (CCS) selection.

Algorithm: CCS Selection

① Initialization

$\Psi_k \leftarrow \varnothing$, $count \leftarrow 0$.

② Calculate G_k^i between the k^{th} CEU and the i^{th} cell, for $i = 0, \ldots, N-1$.

$G \leftarrow \{G_k^0, G_k^1, \ldots, G_k^{N-1}\}$

③ Find serving cell for k^{th} CEU

$$i \leftarrow \arg\max\left(G_k^i\right), G_k^i \in G$$

$$s \leftarrow i$$

④ Update

$$\Psi_k \leftarrow \Psi_k \cup \{i^{th}\ cell\}$$

$$count \leftarrow count + 1$$

⑤ If $count < M$,

$$G \leftarrow G - \{G_k^i\}$$

$$i \leftarrow \arg\max\left(G_k^i\right), G_k^i \in G$$

Else stop.

⑥ If $G_k^s - G_k^i \le thr$, go to ④

Else stop.

The 3GPP document [5] mentioned that if the number of cell in UE CCS is 2, it is enough to reach the CoMP gain effect. Therefore, the CFR set the value of the M is 2.

However, one drawback of CFR is it lacks a rule to allocate the frequency resources of CE. Thus, it is possible that the frequency in the CE CCS maybe allocated by other CEU and CoMP transmission can't work effectively and the totally throughput is decreasing.

SeFR will proposed to offers different frequency allocation of resources in different regions of a cell and give the order of use restrictions [14], use different frequency reuse assignment to reduce the ICIs among the neighboring cells and increases the total throughput.

III、Sequence Frequency Reuse

This section, the proposed Sequence Frequency Reuse (SeFR) mechanism is divided into four steps, shown as figure 3.1.

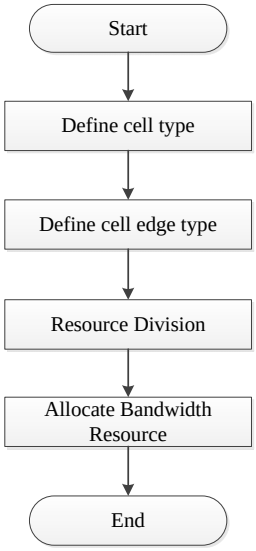


Fig. 3.1 SeFR flow chart

1. Define Cell Tye

In phase one, all of cells are classified into three categories: type 1, type 2, and type 3. We can start at any one cell and define it as type 1. Next, the six neighboring cells can be defined as type 2 and 3, separated by clockwise. After that, the neighboring cells will apply the method to define their other neighboring cells which are undefined. Figure 3.2 presents the basic definition.

2. Define Cell Edge type

The second phase, SeFR will divide CE into six zones according adjacent cell and give zones sequence number (number 1~6) from top (clockwise). It shows in figure 3.3.

Fig. 3.2 Cell type definition

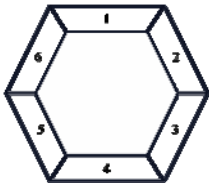


Fig. 3.3 Division and Number

Fig. 3.4 Define cell-edge type flow chart

Some users suffer from seriously interference in region that covered from three cells because the user can receive three cell signals at the same time, as show in figure 3.4. The region we call Cell-Corner (CCr) [6], and the users located at CCr are called Cell-Corner User (CCrU).

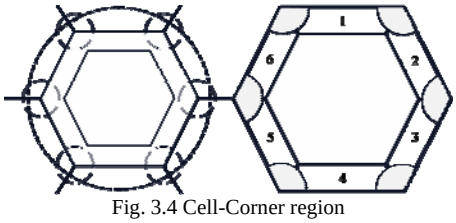


Fig. 3.4 Cell-Corner region

3. Resource Division

The Physical Resource Block (PRB) is minimum resource unit in LTE. We divide the resource into: Cell-Corner Resource (CCrR), Cell-Edge Resource (CER), Cell-Center Resource (CCR). At first, we divide the resource into two parts, called G and F. G is supported as CCR. F further divides into 4 parts, called R_1, R_2, R_3 and CCrR. The $R_{1\sim3}$ is allocated for CE, and CCrR is allocated for CCr. $R_{1\sim3}$ further divide into 6 partitions, and give a sequence number 1~6, shown in figure 3.5.

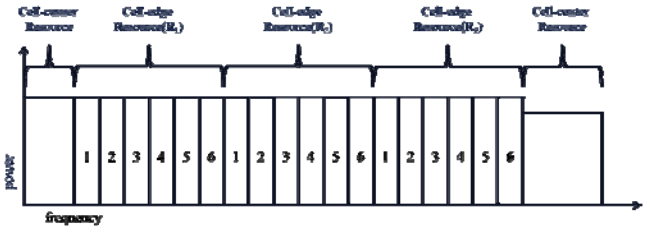


Fig. 3.5 Division of PRB

4. Allocate Resource

Because zone 1 of any cell type is serial number is 1, so that the frequency resource starts allocated from partition 1 of resource R_i, j =1. 2, or 3.

A. Resource allocated on CE

Formula (1) can use to decide the resource allocation:

$$\begin{cases} r = (x\%3)+1, & \text{if i is odd} \\ r = [(x+1)\%3]+1, & \text{if i is even} \end{cases}$$

x is express cell type, i and r express CE zone i use R_r resource. Figure 3.6 can express resource allocation of the cell type 1, for example. The black partition is restriction, means that the cell didn't use this part but use this part to do CoMP transmission.

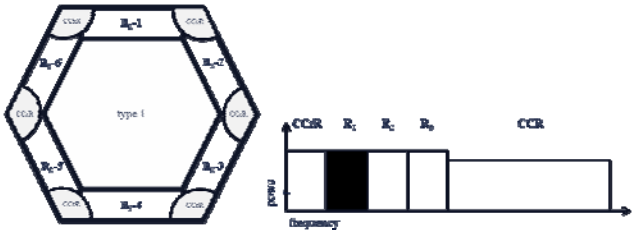


Fig. 3.6 resource allocated of the cell

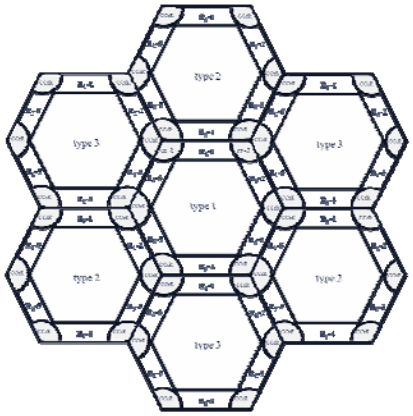


Fig. 3.7 cell cluster

For example, the center cell of figure 3.9, it is one of type 1 cell, we can use Formula (1) to decide what resource could allocate if the zone 1 need resource. Such as zone 1 of this cell, it will allocate resources starting from partition 1 of R_2. The adjacent CE is zone 4 of a type 2 cell, this CE will allocate resource starting from partition 4 of R_3. We can observe from figure 3.7 that does not exist two adjacent cells of the CE are using the same frequency band resources. SeFR gives each cell edge a serial number and using resource allocation rule that can efficiently decrease interference. It will be better than CFR.

CoMP transmission from center cell to surround cells is shown in figure 3.8(a). CoMP transmission from surround cells to center cell presents in figure 3.8(b). For the center cell CE, all of CE surrounding center cell are use different resource. Type 1 didn't use R_1 resource, due to R_1 of type 1 cell will execute CoMP transmission to all surrounding CE of neighboring cells completely because that all of surrounding CE allocated resources are start from different partition.

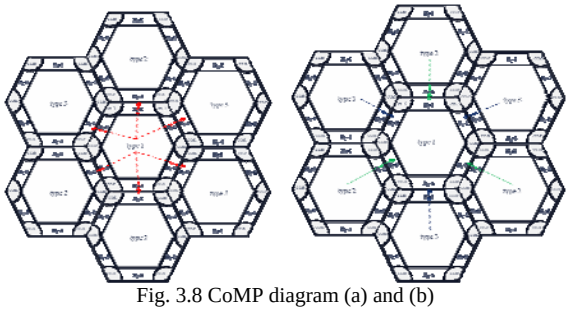


Fig. 3.8 CoMP diagram (a) and (b)

SeFR applies a resource rule to avoid resource allocated not uniform. For example, zone 1 and zone 3 and zone 5 allocated resources with starting from R_2-1, R_2-3, and R_2-5 of R_2, respectively in the cell of type 1. Therefore, they have highest priority for these partitions that other CE couldn't use. In zone 1 of type 1 cell, R_2-1 with highest priority; R_2-2 with high priority because R_2-2 is second sequence for zone 1, and R_2-3 and R_2-5 without priority because R_2-3 and R_2-5 belong to others zones highest priority. R_2-4 and R_2-6 with low priority because R_2-4 and R_2-6 belong to others zones high priority. Table 3-1 illustrates the resource rule.

Table 3.1 type 1 CE allocated resource priority

	R_2-1	R_2-2	R_2-3	R_2-4	R_2-5	R_2-6
zone 1	**highest**	high	without	low	without	low
zone 3	without	low	**highest**	high	without	low
zone 5	without	low	without	low	**highest**	high

Resource allocated on CCr

SeFR allocates resource to independent of cell corner (CCr) because the user in this region have lower SINR that easy get higher interference by more than two cells.

Assume the resource *r* has already to be allocated to user *u* that *u* have nice signal with three source signals (CoMP transmission by *c1*, *c2* and *c3*), while user *v* also need resource *r* but *v* could not get three source signals because only *c4* and *c5* can joint transmission. Whether *u* or *v*, both users can achieve a good SINR. It shows in figure 3.9.

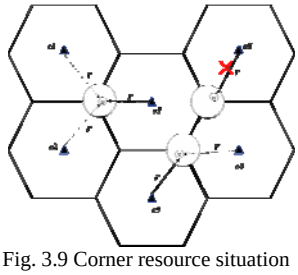


Fig. 3.9 Corner resource situation

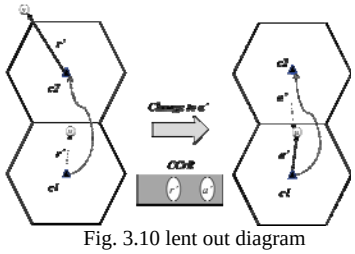


Fig. 3.10 lent out diagram

However, if user *w* also needs same resource *r*, *w* only has one signal resource, but have more interference existed. The SINR of user w must be terrible, moreover it will waste resource, so that it will not allocate resource *r* to *w* instead change another resource or wait for release. In this case, in order to achieve better resource utilization, *r* resource of *c6* cell will be allocated to another user, not for the user w.

To solve this problem of Cell Corner, SeFR uses a lend mechanism [13], when CE needs more resources and CCrR have free resource. The priority of borrowed resource is low. The user *u* needs more resources and the CCrR of serving cell have free resource, so that *u* can borrow resource. Before CCrR lend out resource *r'*, serving cell will ask the adjacent cells that the resource *r'* is free or not? If yes, the serving cell will lent out *r'* and sent requirements to adjacent cells to do CoMP transmission on *r'*. Otherwise, the serving cell will change another free resource to *u*. Figure 3.10 shows the lend mechanism. Thus, SeFR gets best SINR using by CoMP transmission with borrowed resource.

Of course, SeFR provides the resource return mechanism for previous lend process. Figure 3.11 presents the return mechanism.

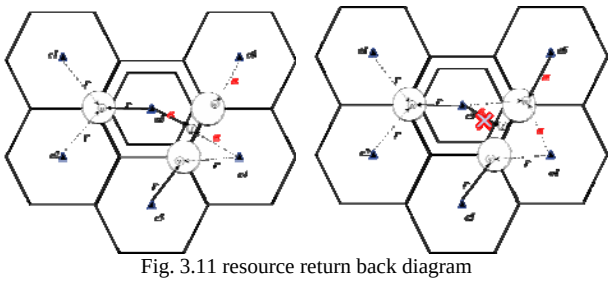


Fig. 3.11 resource return back diagram

IV、Simulation

We write a simulation program tool by C++, through the actual simulation to analyze the performance of SeFR. The comparison proves SeFR can achieve better results than CFR in the cell edge users.

A .Simulation Environment

Table 4.1 SIMULATION PARAMETERS

Parameters	Values
Carrier Frequency	2GHz
Bandwidth	10MHz
Subcarrier Spacing	15kHz
Number of PRBs	50
Number of Cells	19
Cell Radius	500m
Maximum power in BS	46dBm
Distance-dependent path loss	L=128.1+37.6log10*d*(dB). *d* in km
Shadowing Factor Variance	8dB
Shadowing Correlation Distance	50m
Inter cell Shadowing Correlation	0.5

We use the formula (2) [7] to calculate the SINR of each user. According to the Shannon theory [5], it calculates the SINR and the allocation of bandwidth to the user's capacity and the throughput of the cell as a whole.

For formula (2), *r* is the SINR, k^{th} user, i^{th} is serving cell which uses l^{th} PRB. The numerator represents the source of the signal strength, s^{th} represents as cells of CCS belong to k^{th} user. For CEU and CCrU, we take two and three cells, respectively [4]. $P_{s,l}$ is transmission power of s^{th} on l^{th} PRB, G_s^k is long term gain between s^{th} cell and k^{th} CoMP user. Denominator is that all sources of interference the total, N_0 is noise, *n* is the cell not exist in the CCS of k^{th} user, $x_{n,l}$ is n^{th} cell that using l^{th} PRB, $P_{n,l}$ is transmission power on l^{th} of n^{th} cell, G_n^k is long term gain between n^{th} cell and k^{th} user.

Formula (2) SINR to each user:

$$r_{i,l}^{k} = \frac{\sum_{s \in CCS} P_{s,l} G_{s}^{k}}{N_0 + \sum_{n \in CCS} x_{n,l} P_{n,l} G_{n}^{k}}$$

Formula (3) [7] is that calculate the user capacity. *C* is capacity, *k* is user number, *i* is serving cell number, l^{th} is PRB number, and *B* is the bandwidth.

Formula (3) user capacity of cell:

$$C_{i,l}^{k} = B\log_2(1+r_{i,l}^{k})$$

B. Analysis of simulation results

When users are uniform distribute in cell edge graph, it is shown in figure 4.1. SeFR avoids surrounding cell users use the l^{th} PRB of i^{th} the cells, so that the limit resources can average and complete allocated to neighboring CEU for i^{th} cell. Moreover, it can avoid the problem with i^{th} cell doing CoMP transmission.

SeFR mechanism can improve CoMP on the user's operational efficiency. The SINR is raise a lot in less number od users. Even the user number is increasing, SeFR still maintain the user in a higher SINR than CFR.

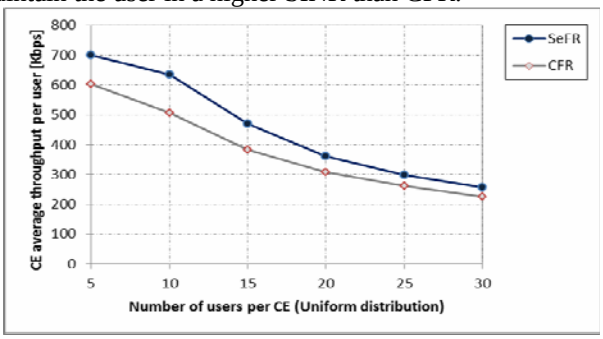


Fig. 4.1 Cell-Edge average throughput per user on uniform distribution

Observed in figure 4.2, the users are uniform distribution of the overall throughput of the cell edge in the case that Strengthen the the CoMP ability to CEU to SeFR mechanisms, while also focusing on users in less time as the user can accurately enhance CoMP the ability of the cell corner regions, the average throughput of each user is higher than the CFR.

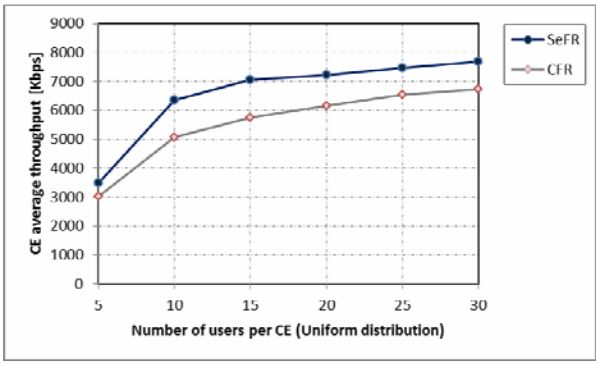


Fig. 4.2 Cell-average throughput as the number of users per cell

V、Conclusion

This paper presents a Sequence Frequency Reuse (SeFR) mechanism to improve the original technology that works on CoMP of LTE-A. SeFR uses the re-division of the frequency resources to the given six cell edge zones and allocates frequency resources to each corresponds to the zone number. It can make the CoMP transmission to achieve higher efficiency in the use of, and does improve the SINR value of CEU. SeFR also applies the independent of the cell corners, and give part of a separate allocation of resources, at the corner of the cell. The user can effectively use the the CoMP bring the advantages of, and the reduce interference. The SINR value can be upgraded. Finally the proposed mechanism compares to the existing mechanisms to prove that it is more suitable for LTE-A CoMP technology.

In the future, this mechanism will consider to including transmission power and others technologies of LTE-A that can make some improves to this mechanism.

References

[1] 3GPP TR 25.814 v7.1.0, “Physical layer aspects for evolved UTRA (Release 7),” Sept. 2006.

[2] 3GPP TR 36.814 v1.0.1, “Further Advancements for E-UTRA Physical Layer Aspects (Release 9),” Mar. 2009.

[3] R1-050507, “Soft frequency reuse scheme for UTRAN LTE,” Huawei, 3GPP TSG RAN WG1 Meeting #41, May 2005.

[4] R1-091688, “Potential gain of DL CoMP with joint transmission,” NEC Group, 3GPP TSGRAN WG1 #57, May. 2009.

[5] 3GPP RAN1 R1-050764, "Inter-cell interference Handling for E-UTRAN," Ericssion.

[6] W. Fu, Z. Tao, J. Zhang, D. P. Agrawal, "Clustering Based Fractional Frequency Reuse and Fair Resource Allocation in Multi-cell Networks," in Proc. IEEE International Conference on Communications, pp.1-5, May 2010.

[7] D. Jia, G. Wu, S. Li, G.Y. Li, X. Zhu, ” Dynamic Soft-Frequency Reuse with Inter-Cell Coordination in OFDMA Networks,” in Proc. International Conference on Computer Communications and Networks, pp.1-6, Aug. 2011.

[8] H. Jia, Z. Zhang, G. Yu, P. Cheng, and S. Li, “On the Performance of IEEE 802.16 OFDMA System under Different Frequency Reuse and Subcarrier Permutation Patterns,” in Proc. IEEE International Conference on Communications, pp. 5720- 5725, Jun. 2007.

[9] K. T. Kim, S. K. Oh, “A Universal Frequency Reuse System in a Mobile Cellular Environment,” in Proc. IEEE Vehicular Technology Conference, pp. 2855-2859, April 2007.

[10] K. T. Kim, S. K. Oh, “An Incremental Frequency Reuse Scheme for an OFDMA Cellular System and Its Performance,” in Proc. IEEE Vehicular Technology Conference, pp. 1504-1508, May 2008.

[11] J. Li, H. Zhang, X. Xu, X. Tao, T. Svensson, C. Botella, B. Liu, “A Novel Frequency Reuse Scheme for Coordinated Multi-Point Transmission,” in Proc. IEEE Vehicular Technology Conference, pp. 1-5, May 2010.

[12] H. Le, L. Zhang, X. Zhang, and D. Yang, “A Novel Multi-Cell OFDMA System Structure using Fractional Frequency Reuse,” in Proc. IEEE Personal, Indoor and Mobile Radio Communications, pp. 1-5, 2007.

[13] Z. Lu, H. Tian, Q. Sun, B. Huang, S. Zheng, "An Admission Control Strategy for Soft Frequency Reuse Deployment of LTE Systems," in *Proc. IEEE Consumer Communications and Networking Conference*, pp. 1-5, Jan. 2010.

[14] C. Nie, P. Liu, S. Panwar, “Interference Management using Frequency Planning in an OFDMA based Wireless Network,” in Proc. IEEE Wireless Communications and Networking Conference , pp.998-1003, March 2011.

[15] G. Piñero, C. Botella, A. González, M. de Diego and N. Cardona, “Downlink power control and beamforming for a cooperative wireless system,” in Proc. IEEE Personal, Indoor and Mobile Radio Communications, pp. 974-978, Setp. 2004.

[16] P. Wang, C. Liu, R. Mathar, “Dynamic Fractional Frequency Reused Proportional Fair in Time and Frequency Scheduling in OFDMA Networks,” in Proc. IEEE International Symposium on Wireless Communication Systems, pp.745-749, Nov. 2011.

A Comparative Study of Proactive and Reactive Geographical Routing Protocols for MANET

Rob Hussey, Earl Huff, Zabih Shinwari, and Vasil Hnatyshin†
{hussey20, huffe72, shinwa59}@students.rowan.edu, †hnatyshin@rowan.edu
Department of Computer Science
Rowan University
Glassboro, NJ 08028

***Abstract*—*In the world of mobile wireless communication, it has become more and more important to establish networks that are not only capable of delivering information across vast distances but can also perform this task efficiently. Many routing protocols for mobile ad hoc networks (MANETs) rely on additional information such as geographical locations obtained via GPS to improve the overall performance of the route discovery process. This paper is an extension of our previous study of location-aided MANET routing protocols. In this paper we continue our research endeavors by comparing the performance of several AODV-based reactive, location-aided MANET routing protocols and Geographical Routing Protocol (GRP), an OPNET implementation of a proactive, geographical location-based routing protocol for MANET.*

Keywords: location-aided routing; geographical routing; LAR; GeoAODV; AODV; GRP

1 Introduction

As the world becomes more and more reliant on wireless communication, efficient delivery of information from one network device to another becomes critical. Since these devices are often mobile, it becomes even more important to develop the means for data delivery in the environments that experience frequent topological changes [3]. Mobile ad hoc networks (MANETs) are collections of autonomous mobile nodes which work together to transport information through wireless environments [11]. The dynamic nature of MANETs makes finding a route from source to destination a challenging task.

Generally, MANET routing protocols are divided into two broad categories: *reactive* – the source only tries to find a route to the destination as needed and *proactive* – the nodes continually maintain the routes in the network regardless of whether there is traffic traveling to the destination or not. The main advantage of reactive routing protocols is that they do not waste resources, which are typically very scarce in MANETs, on the routes which may not be needed. However, when a source node has data to be transmitted, a route to the destination may not be readily available. This may result in the transmission being delayed until a route to the destination is found. On the other hand, when proactive routing protocols are used, the data can be transmitted right away since each node maintains and continually updates the routes to all reachable nodes in the network. The main disadvantage of proactive routing protocols is that the nodes maintain the routes even if they are not used, which results in unnecessary waste of available resources such as bandwidth, battery power, etc.

The route discovery process in a MANET environment often relies on flooding to find a path to the destination. Typically, flooding also unnecessarily consumes available resources because it searches the whole network, including the portions of the network which are unlikely to contain a route to the destination. In recent years there have been a large number of proposals which attempt to improve the performance of the route discovery process by utilizing geographical information. In this paper we examine and compare the performance of several location-aided, *reactive* routing protocols based on Ad hoc On-demand Distance Vector (AODV) and Geographical Routing Protocol (GRP), a *proactive*, geographical location-based routing protocol for MANET. Improving MANET routing through the use of location information have been an active area of research [1-2, 4-6, 8-11]. However, in this paper we examine and study through simulation two variations of the Location-Aided Routing (LAR) protocol [9,10], two variations of Geographical AODV (GeoAODV) routing protocols [1,5], and an OPNET implementation of GRP [13]. The results presented in this paper were collected using the OPNET Modeler version 16.1 network simulation software [12].

The rest of the paper is organized as follows. We provide a brief overview of studied routing protocols in Section II. Set-up of the simulation study and analysis of results are presented in Sections III and IV. The paper discusses the plans for future work and concludes in Section V.

2 Overview of Location-Aided Routing Protocols for MANET

2.1 LAR

Ad hoc on-demand distance vector (AODV) is a reactive routing protocol for MANETs [3, 14 - 15]. AODV performs route discovery using flooding. When a source node, let us call it the originator, needs to send data but does not have a route to destination, it initiates the route discovery process, which works as follows. The originator node broadcasts a route request (RREQ) message to its immediate neighbors, which in turn, rebroadcast the message farther until the node that has a path to the destination or the destination itself is reached. At this point, a route reply (RREP) message is unicast back to the originator

node, establishing a path between the source and destination nodes. The route discovery process completes when the originator node receives the RREP message, at which point it can start transmitting the data.

The Location-Aided Routing (LAR) protocol [6, 9 - 10] is an extension of the AODV protocol, which relies on the geographical position of the nodes and their traveling velocities to limit the search area during the route discovery process. LAR assumes that all the nodes know the Global Positioning System (GPS) locations and average traveling speed of all the other nodes in the network. LAR performs route discovery in a fashion similar to that of AODV. However, in AODV, RREQ messages are forwarded to all the nodes in the network, while LAR uses geographical information to limit the RREQ flooding to only those nodes that are likely to be part of the path to the destination. This technique significantly reduces the control message overhead of the route discovery process by forwarding the RREQ messages only in a portion of the whole network.

There are two main variations of the LAR protocol which we call *LAR zone* and *LAR distance*. LAR zone uses the destination's last known coordinates and traveling speed to determine an *expected zone*, an area which is likely to contain the destination node. The expected zone is defined as a circle with radius **R**, centered in the last-known GPS location of the destination node recorded at time $\mathbf{t_0}$. The value of **R** is computed as shown in equation (1):

$$R = v \times (t_1 - t_0) \qquad (1)$$

In equation (1) **v** is the average traveling speed of the destination node and $\mathbf{t_1}$ is the current time. Based on the expected zone area, LAR computes the *request zone*, a rectangular area which is likely to contain the path to the destination. A request zone is the smallest rectangle that encompasses the expected zone such that the sides of the request zone are parallel to the X and Y axes. Only nodes located inside of the request zone participate in RREQ flooding, while all the other nodes simply discard arriving RREQ messages. Figure 1 illustrates two possible scenarios of the expected and request zone locations: (a) the source node S is *outside* of the expected zone for destination node D and (b) the source node S is *inside* of the expected zone for destination D.

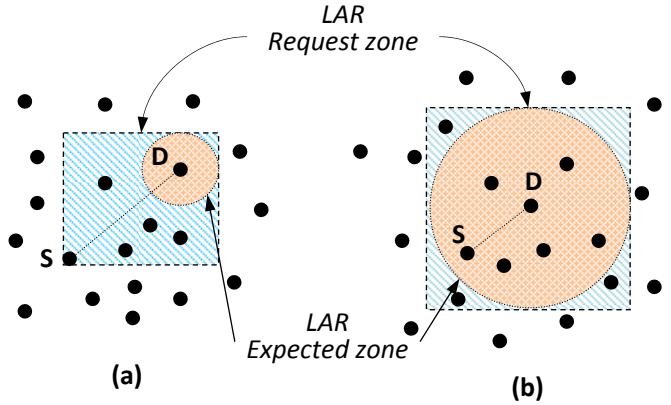


Figure 1: LAR zone: Expected and Request Zones

In the LAR distance approach a node participates in the flooding, i.e., rebroadcasts the RREQ message, only if it is located not father away from the destination than the node that forwarded an RREQ. Generally, LAR distance relies on inequality (2) to determine if node $\mathbf{N_1}$ that receives an RREQ from node $\mathbf{N_0}$ will rebroadcast the message:

$$\alpha \times |N_0 D| + \beta \geq |N_1 D| \qquad (2)$$

In inequality (2) we denote distance between nodes A and B as |**A B**|, while α and β are configuration parameters. We provide an example of LAR distance operation in Figure 2. Source node **S** initiates route discovery by broadcasting an RREQ. At some point node $\mathbf{N_0}$ receives this RREQ and rebroadcasts it farther. When node $\mathbf{N_1}$ receives an RREQ from node $\mathbf{N_0}$ it rebroadcasts the message because $|\mathbf{N_1\ D}| \leq |\mathbf{N_0\ D}|$. However, nodes $\mathbf{N_2}$ and $\mathbf{N_3}$ will discard an RREQ forwarded by $\mathbf{N_0}$ because $|\mathbf{N_2\ D}| > |\mathbf{N_0\ D}|$ and $|\mathbf{N_3\ D}| > |\mathbf{N_0\ D}|$, respectively.

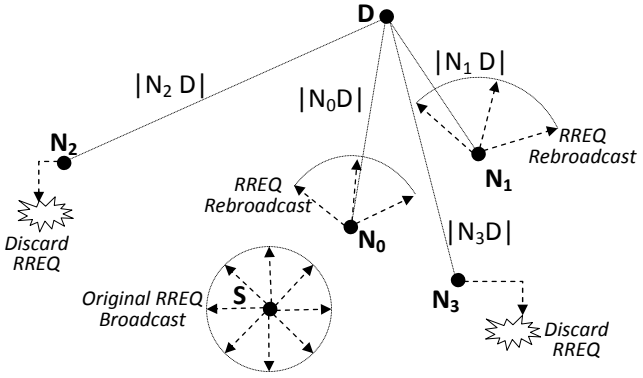


Figure 2. Example of LAR distance scheme

The major difference between these approaches is that LAR zone assumes universal availability of GPS coordinates and traveling velocities needed for computation of a search area where the path to the destination may reside. LAR distance only relies on the availability of GPS coordinates for computing the distances between the nodes. LAR zone has no restrictions as to how the path to destination is constructed; the route can move farther away from the destination before actually reaching it. LAR distance on the other hand, constructs the path by attempting to come closer and closer to the destination during each RREQ rebroadcast. Such an approach may result in a failure to find a route to the destination even though it exists. LAR distance attempts to mitigate this issue by parameterizing the inequality (2) through configuration parameters α and β. However, in practice, determining the optimal values for α and β is a challenging task. Furthermore, LAR zone also suffers from a similar problem: it may fail to find the path to the destination if a portion of the path resides outside the request zone area. Both LAR schemes have no mechanism for expanding the search after a failed attempt to find a route; the route discovery process is stopped if a limited RREQ flood did not find a route to destination. Geographical AODV (GeoAODV) attempts to address this issue by increasing the search area after each failed attempt until GeoAODV morphs into regular AODV.

2.2 Geographical AODV

GeoAODV is based on the same idea as the LAR zone protocol: only nodes within the search area, i.e., the request zone, participate in route discovery. However, unlike LAR zone, GeoAODV does not assume that GPS locations and traveling velocities of the nodes are readily available to all the other nodes in the network. Instead, GeoAODV assumes that the nodes only

know their own location information. In GeoAODV, the location information is dynamically distributed during the route discovery process; i.e., the RREQ and RREP messages are modified to also carry the location information which is recorded by all intermediate nodes that receive these messages.

GeoAODV defines the request zone in the shape of a cone as shown in Figure 3. The originator node **S** serves as an apex of the cone-shaped request zone. The "width" of the area is controlled through the configuration parameter **α** called the flooding angle, which is evenly divided by the straight line between originator **S** and destination **D**. After each failed attempt to find a route to destination, i.e., a single round of route discovery, the value of the flooding angle increases, expanding the search area and the process is repeated again, i.e., the next round of route discovery is started. This continues until either a path to the destination is found or the route discovery fails to find the path with the flooding angle value of 360 degrees, in which case the whole network has been searched). GeoAODV eventually may search the whole network and thus, it guarantees that a route to the destination will be found if one exists.

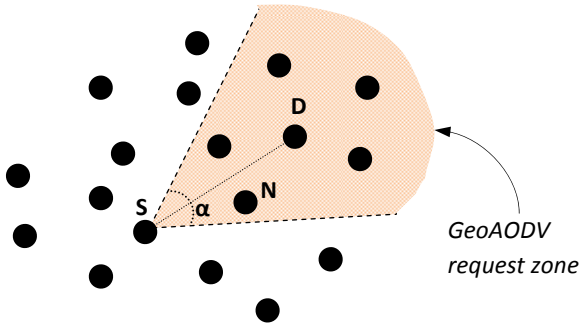


Figure 3: GeoAODV request zone

There are two variations of the GeoAODV protocol: *GeoAODV static* and *GeoAODV rotate*. In GeoAODV static, the originator node always serves as an apex of the request zone cone. This means than the request zone remains the same through each round of route discovery. GeoAODV rotate dynamically adjusts the search area during the route discovery process. Specifically, in GeoAODV rotate each intermediate node re-computes the request zone area based on the location of the previous hop, instead of the originator node, which effectively realigns the search area towards the destination node. Figure 4 illustrates the idea of GeoAODV rotate: node $\mathbf{N_1}$ belongs to the request zone computed based on location of node **S** while node $\mathbf{N_2}$ belongs to the new, re-adjusted request zone computed based on the location of node $\mathbf{N_1}$. Both $\mathbf{N_1}$ and $\mathbf{N_2}$ participate in route discovery even though they belong to different request zones. On the other hand, $\mathbf{N_3}$, which receives an RREQ from $\mathbf{N_1}$, will not participate in the route discovery because it does not belong to the request zone computed based on the location of its previous hop, which is node $\mathbf{N_1}$. However, when GeoAODV static is used, $\mathbf{N_3}$ is part of the request zone computed based on the location of originator node **S** and thus will be a part of the route discovery process.

Unlike LAR, which assumes that location information and traveling velocities are readily available everywhere in the network, GeoAODV makes more realistic assumptions about the availability of GPS location information, in that the nodes only know their own location information, which is distributed during the route discovery process. Furthermore, by increasing the search area after each failed attempt, GeoAODV guarantees that a route to the destination will be found if it exists.

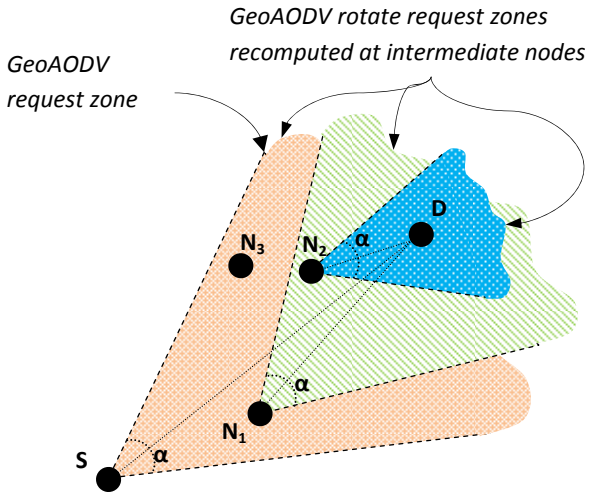


Figure 4: GeoAODV rotate request zone

2.3 GRP

The Geographic Routing Protocol (GRP) is a custom location-based MANET routing protocol developed by OPNET Technologies**,** Inc [12]. GRP is a proactive, distance-based, greedy algorithm which assumes that each node in the network knows its own GPS location. GRP relies on physical distances for routing: the next hop on the path to the destination is selected as the node geographically closest to destination.

GRP relies on the concept of quadrants or neighborhoods for routing. The network area is divided into square quadrants as shown in Figure 4. Given the GPS coordinates of the node, GRP can easily determine the quadrant it belongs to. Every four quadrants of the lower level form a square or quadrant of a higher level. As Figure 4 illustrates, quadrants **Aa1**, **Aa2**, **Aa3**, and **Aa4** from level 1 form the single level 2 quadrant **Aa**. The size of the lowest-level quadrant is a configurable parameter.

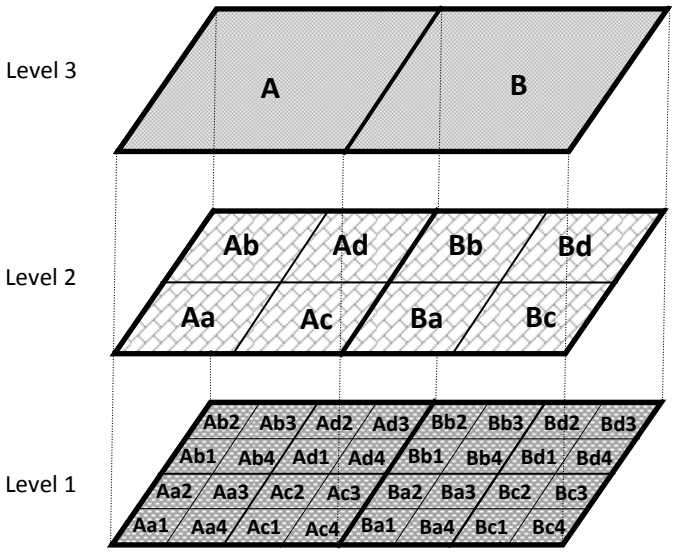


Figure 5: Quadrant division in GRP

GRP maintains forwarding tables as the geographical positions of the nodes in the network. Specifically, the

forwarding table of a node stores precise GPS locations of all the other nodes in the same quadrant and the highest level neighboring quadrant label for the nodes located in different quadrants. For example, assume that nodes $\mathbf{N_1}$ and $\mathbf{N_2}$ are located in quadrant **Ac2**, while nodes $\mathbf{N_3}$ and $\mathbf{N_4}$ are located in quadrants **Ad1** and **Bb4**, respectively. In this case, node $\mathbf{N_1}$ will store the following location information: $\mathbf{N_2}$ – precise coordinates since $\mathbf{N_1}$ and $\mathbf{N_2}$ are both in the same quadrant; $\mathbf{N_3}$ – quadrant **Ad** because **Ad** is the highest level quadrant adjacent to **Ac2**; $\mathbf{N_4}$ – quadrant **B** because **B** is the highest level quadrant neighbor of **Ac2**.

The GRP forwarding process works as follows. If the source and destination nodes are located in the same quadrant then the source sends the data to its immediate neighbor geographically closest to the destination. The intermediate node does the same by forwarding the data to its immediate neighbor closest to the destination, and this process will continue until the data arrives at the destination. If source and destination are located in different quadrants then the source node sends the data to its immediate neighbor closest to the entry point into the highest-level quadrant in which the destination node resides. As the data traverses the quadrant boundaries the location information about the destination becomes more specific until eventually the data arrives at the destination's quadrant and is routed using precise location information.

For example, consider the situation when $\mathbf{N_1}$ from quadrant **Ac2** sends data to node $\mathbf{N_4}$ in quadrant **Bb4**. In this case $\mathbf{N_1}$ will send the data to the node closest to quadrant **B**. Eventually, the data will reach an intermediate node in, let us say, quadrant **Ba2**, which will have more precise location information about $\mathbf{N_4}$. Specifically, the intermediate node in quadrant **Ba2** will have the location of $\mathbf{N_4}$ recorded as quadrant **Bb**. Similarly, when data arrives at an intermediate node in quadrant **Bb1**, the forwarding information will state that $\mathbf{N_4}$ is located in quadrant **Bb4**. Eventually the data will arrive at some intermediate node in quadrant **Bb4**, at which time it will be forwarded according to precise location information.

It is possible that the data may reach an impasse, i.e., a blocked route, in which the current intermediate node has no neighbors besides the node from which the data arrived. In this case, the forwarding algorithm backtracks to the previous node which forwards the data to the next closest neighbor on the path to the destination. GRP allows recursive backtracking all the way back to the source node such that if an intermediate node receives a backtrack request and there are no more neighbor nodes to try, then in an attempt to find an alternative route, it forwards the packet back to the node from which it originally arrived. If the source node receives a backtrack packet and it has no more neighboring nodes to try, then it is determined that there is no path to the destination and the data is discarded.

To create forwarding tables, GRP also relies on flooding. Initially, GRP performs a network wide flooding to discover location information of all the reachable nodes in the network. After initial route discovery, GRP periodically conducts limited flooding in order to update the forwarding tables. GRP initiates limited flooding based on node movement, i.e., whenever a node moves a set distance or crosses a quadrant boundary. The area of the limited flooding is determined based on the quadrant boundary that was crossed. For example, if the node did not cross the quadrant boundary, that is, the limited flooding was initiated based on the distance traveled, then the flooding is restricted to the node's quadrant only. If the node crossed a quadrant boundary, then the flooding is performed in the highest level quadrant which is common to the quadrants on each side of the boundary. For example, if a node crosses the quadrant boundary between **Aa2** and **Ab1**, then the flooding will be limited to quadrant **A**. The route discovery messages received outside the flooding area are discarded. Finally, in order to keep location information about its immediate neighbors up-to-date, GRP requires every node to broadcast periodic hello beacon messages [13].

TABLE 1: SUMMARY OF NODE CONFIGURATION

Configuration Parameter	Value
Channel Data Rate	11 Mbps
Transmit Power	0.001 Watts
Packet Reception Power Threshold	-95 dBm
Start of data transmission	normal(100, 5) seconds
End of data transmission	End of simulation
Duration of simulation	300 seconds
Packet inter-arrival time	exponential(1) second
Packet size	exponential(1024) bytes
Mobility model	Random Waypoint
Pause Time	exponential(10)
Destination	Random

3 Simulation Set-up

We compared the performance of LAR, GeoAODV, and GRP protocols using OPNET Modeler version 16.1 [12]. The network topology in our study contained 50 WLAN nodes randomly placed within a 1500 meters x 1500 meters area. We examined scenarios with 2, 5, 15, and 30 randomly selected communicating nodes. The communicating nodes began data transmission 100 seconds after the start of the simulation, which itself ran for 300 seconds. The nodes in the network moved according to the Random Waypoint model with pause time computed using exponential distribution with the mean outcome of 10 seconds. We examined the performance under two sets of scenarios: (1) all the nodes in the network are stationary and (2) all the nodes in the network travel with the speed 20 meters per second. Summary of individual node configuration presented in Table 1.

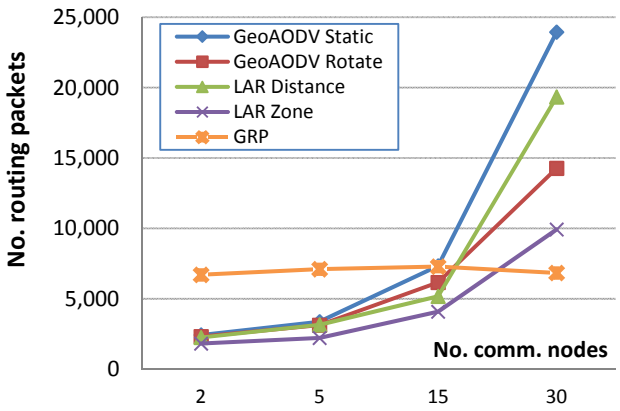


Figure 6: Routing traffic sent when nodes travel at 20 m/s

The geographical location-based routing protocols examined in our study were configured as follows. In LAR scenarios, individual nodes distributed their precise location information once every second. We set α and β parameters of LAR distance

protocol to 1 and 0, respectively. GeoAODV protocols were configured to have the initial value of the flooding angle set to 90 degrees. After each failed round of route discovery the value of flooding angle was increased by 90 degrees, until it reached 360 degrees, at which point GeoAODV morphed into regular AODV protocol. GRP was configured to perform a single initial flood. Limited flooding was triggered whenever a node traveled 250 meters or crossed the boundary of a 375 meters x 375 meters quadrant, i.e., the network area was divided into four GRP quadrants. The remaining configuration attributes were set to their default values.

4 Analysis of Results

The results collected in our study suggest that the reactive protocols generate less control traffic than GRP in scenarios where the nodes are moving around and there are less than 30 traffic generating sources. However, GRP performed better in all scenarios with stationary nodes and in scenarios with 30 communicating nodes. A summary of collected simulation results is presented in Figures 6 and 7.

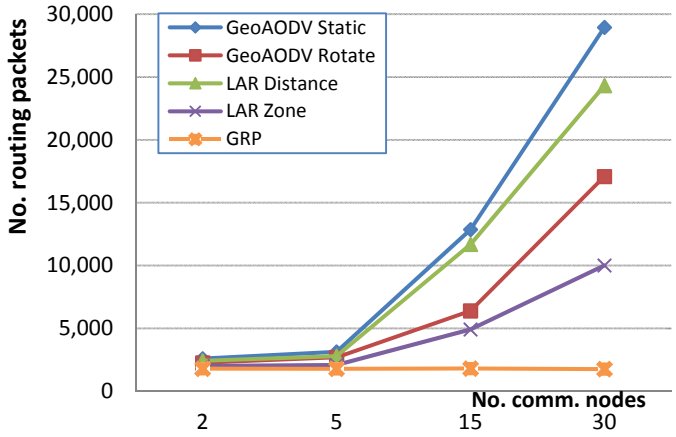


Figure 7: Routing traffic sent when nodes are stationary

GRP performs route discovery during initialization and based on the node movement. However, the number of communicating nodes is not tied in any way to the amount of control traffic generated by GRP. This is clearly reflected in collected results: the number of routing packets generated by GRP remains more or less the same in respect to the number of communicating nodes. However, GRP generates significantly more control traffic in scenarios with mobile nodes than when the nodes are stationary. This happens because in scenarios with stationary nodes GRP generates control traffic only upon initial network-wide route discovery and during periodic "pinging" of neighboring nodes; while when the nodes move around GRP performs additional route discoveries whenever the nodes travel a certain distance or cross quadrant boundaries.

The reactive protocols initiate route discovery whenever there is data to send but route to destination is unknown. Thus, their performance directly relates to the number of traffic generation sources. As the number of communicating nodes increases so do the frequency of route discoveries and the total number of routing packets sent into the network. Collected results suggested that LAR zone protocol generates the least amount of control traffic while GeoAODV rotate is a close second. Even though LAR zone performs the best it relies on the assumption that the GPS location and traveling velocities of all the nodes in the network are readily available, while GeoAODV makes no such assumption and distribute location information during route discovery. Thus, even though GeoAODV rotate generates slightly more control traffic than LAR zone, it may be a better choice in certain environments.

We compared the length of the path taken by the data packets when routed using the proactive GRP protocol and reactive location-aided protocols. The length of the path taken when using GRP was about twice as long as that of any reactive protocol. This phenomenon is most likely due to the greedy nature of GRP: intermediate nodes route the data packet to the next hop node that is closest to destination and backtrack if an impasse is encountered. GRP does not maintain the next hop id in its routing table; instead, the routing tables store the location information. GRP performs some limited route discovery while forwarding the data: it tries to find the next hop which is closest to destination and is a part of the path. This results in occasional detours and backtracks which extend the length of the path. Reactive protocols examined in this study actually find the shortest path to destination before forwarding the data. That is why the path length for all examined reactive protocols is about the same and significantly shorter than that of GRP.

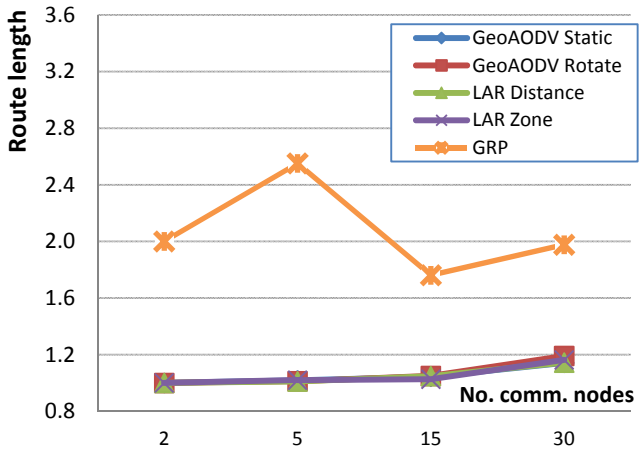


Figure 8: Route length when node travel at 20 m/s

5 Conclusions

This paper compares the performance of reactive and proactive geographical location–aided routing protocols for MANETs through simulation using the OPNET Modeler version 16.1 software package [12]. Collected results suggest that proactive protocols will generate less control traffic in the network environment where the nodes are stationary. This occurs because once proactive protocol collects routing information it does not need to be updated very frequently, since the routes remain the same. On the other hand, reactive protocols perform route discovery every time there is data to send. However, in environment where the nodes constantly move, reactive protocols perform better due to the fact that proactive protocols will have to update routing information proportionally to the node movement, while reactive protocols update routing information only when there is data to send. However, a more detailed study of this phenomenon is needed.

Currently, we are investigating other aspects of the reactive protocols which might affect their performance. In particular,

we are looking into various optimizations of the GeoAODV protocol in respect to the initial value of the flooding angle and how the flooding angle is expanded after initial failure to find a route. Similarly, we are examining the possibility of improving the LAR distance protocol by dynamically adjusting the values for α and β parameters and modifying LAR zone to extend the request zone after route discovery failures. Finally, we plan to expand our study by comparing the performance of other proactive routing protocols, such as the Greedy Perimeter Stateless Routing (GPSR) protocol [6].

6 References

[1] H. Asenov and V. Hnatyshin, "GPS-Enhanced AODV routing," in Proceedings of the 2009 International Conference on Wireless Networks (ICWN'09), Las Vegas, Nevada, USA, July 13-16, 2009.

[2] L. Barriere, P Fraigniaud, L. Narajanan, and J. Opatrny, "Robust position based routing in wireless ad hoc networks with unstable transmission ranges", in Proceedings of the 5th international workshop on Discrete algorithms and methods for mobile computing and communications (DIALM '01). New York, NY, USA: ACM Press, 2001, pp. 19-27.

[3] D. Chakeres and E. M. Belding-Royer, "AODV routing protocol implementation design," In Proceedings of the 24th International Conference on Distributed Computing Systems Workshops (ICDCSW'04), IEEE Computer Society, USA, 698–703.

[4] F. De Rango, A. Iera, A. Molinaro, S. Marano, "A modified location-aided routing protocol for the reduction of control overhead in ad-hoc wireless networks," Proc. of the 10th International Conference on Telecommunications, 2003, pp. 1033 – 1037.

[5] V. Hnatyshin, R. Cocco, M. Ahmed, and D. Urbano, "Improving Geographical AODV Protocol by Dynamically Adjusting the Request Zone," In Proc. of OPNETWORK 2012, Washington, DC, August 2012.

[6] A. Husain, B. Kumar, A. Doegar, "A Study of Location-Aided Routing (LAR) Protocol for Vehicular Ad Hoc Networks in Highway Scenario," International Journal of Engineering and Information Technology, 2(2), 2010, pp 118-124.

[7] B. Karp and H. T. Kung, "GPSR: greedy perimeter stateless routing for wireless networks," in MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking. New York, NY, USA: ACM Press, August 2000, pp. 243–254.

[8] Y.-J. Kim, R. Govindan, B. Karp, and S. Shenker, "Geographic routing made practical," in Proceedings of 2nd Symposium on Networked Systems Design and Implementation. USENIX, 2005, pp. 217–230.

[9] Y. Ko and N. H. Vaidya, "Location-aid routing (LAR) in mobile ad hoc networks," Wireless Networks, 6, 2000, 307-321.

[10] Y. Ko and N. H. Vaidya, "Flooding-based geocasting protocols for mobile ad hoc networks," Mobile Networks and Applications, 7(6), Dec. 2002, pp. 471-480.

[11] M. Mauve, J. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks". IEEE Network, 2001.

[12] OPNET Modeler ver. 16.1. OPNET Technologies, Inc, www.opnet.com last visited 2/19/13

[13] OPNET Modeler 16.1 Documentation, OPNET Technologies, Inc., 2011.

[14] C. Perkins, E. Belding-Royer, S. Das. (July 2003). Ad hoc On Demand Distance Vector (AODV) Routing. IETF RFC 3561 (http://www.ietf.org/rfc/rfc3561.txt). Last accessed 2/19/13

[15] C. E. Perkins and E. M. Royer. "Ad hoc On-Demand Distance Vector Routing," Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, Feb. 1999, pp. 90-100.

An Email-Based Performance Analysis of Routing Protocols in Mobile Ad Hoc Networks using OPNET Modeler.

Michel Mbougni
Department of Computer Science
North West University
Mafikeng Campus, South Africa
21248435@nwu.ac.za

Obeten Ekabua
Department of Computer Science
North West University
Mafikeng Campus, South Africa
24069469@nwu.ac.za

***Abstract*— Mobile Ad Hoc Networks (MANETs) are becoming very popular in the world of wireless networks and telecommunication. MANETs consist of mobile nodes which are able communicate with each other without the need of any infrastructure or centralized administration. In MANETs, the movement of nodes is unpredictable and complex; thus making the routing of the packets challenging. As a result, routing protocols play a crucial role in managing the formation, configuration, and maintenance of the topology of the network. A lot of routing protocols have been proposed as well as compared in the literature. However, most of the work done on the performance evaluation of routing protocols is done using the Constant Bit Rate (CBR) traffic. This paper presents the performance analysis of MANETs routing protocols such as Ad hoc on Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Temporary Ordered Routing Algorithm (TORA), and Optimized Link State Routing (OLSR) using Electronic mail (Email) traffic. The performance metrics used for the analysis of these routing protocols are delay and throughput. The overall results show that the proactive routing protocol (OLSR) performs better in terms of delay and throughput than the reactive routing protocols AODV, DSR and TORA.**

***Index Terms*—mobile ad hoc network, routing protocols, Email traffic.**

I. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) are becoming very popular in the world of wireless networks. MANETs are ad hoc networks consisting of mobile nodes which can communicate with each other without any infrastructure.

In MANETs, there is no need for infrastructure or central administration since the temporary network formed by the mobile nodes are self-configuring, self-routing and self-organizing.
Every node in a MANET acts as a router or as a relay station [1]; each node participates in routing packets [2]. That is, the sender node can either forward the packet directly to the destination when it is close enough or through intermediate nodes when the destination node is out of reach [3]. MANET nodes can form the network at anytime and anywhere thus making the network topology highly dynamic and the routing of packets complex. Hence there is a need for MANETs to have routing protocols which can adapt to the mobility and dynamically changing topology of the network.
A number of routing protocols have been proposed, evaluated and implemented. Some researchers have classified routing protocols into two categories: link-state protocols and distance-vector protocols [4], whereas others [5] classified them into four categories: proactive protocols, reactive protocols, hybrid protocols and cluster-based protocols.
In MANETs, the movement of the nodes is unpredictable; so reliable routing protocols should be able to adapt to the unpredictable and dynamic topology of the network caused by the random displacement of mobile nodes within a specific area [3]. As stated earlier, many routing protocols have been proposed and implemented by researchers; however most of them use Constant Bit Rate (CBR) traffic [2], [3], [5], [6], [7] , [8] and [9] because CBR traffic preserves constant bandwidth and minimizes the packets loss during transmission. However, with the increased use of Email applications over the past decade, there is a need to study routing protocols using Email traffic.

This paper propose the performance analysis of MANET's routing protocols e.g., Ad Hoc on Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Temporally Ordered Routing Algorithm (TORA) and Optimized Link State Routing (OLSR) protocols in terms of delay and throughput for a common and simple application such as Email.

II. ROUTING PROTOCOLS OVERVIEW

The challenges and flexibility of MANETs have generated a lot of research in routing protocols for such networks. The network research community has been working intensively on modeling, designing and implementing new routing protocols for MANETs. De Rango et al. [5] classify MANET routing protocols into four categories: proactive protocols, reactive protocols, hybrid protocols and cluster-based protocols. Three popular reactive routing protocols, DSR, AODV and TORA

and a popular proactive routing protocol, OLSR, will be briefly discussed in the next section.

A. Ad hoc On Demand Distance Vector (AODV)

AODV routing protocol is a reactive routing protocol which was first proposed by an IETF Internet draft in 1997. According to Belding-Royer and Perkins [4], AODV was proposed to meet the following goals:

- Minimal control overhead.
- Minimal processing overhead.
- Multi-hop path routing capability.
- Dynamic topology maintenance.
- Loop prevention.

The operation of AODV is done using the following two mechanisms: route discovery and route maintenance [4], [8].
Route discovery: This is a mechanism by which a source node wishing to send a packet to a destination node obtains dynamically a source route when it does not have a route in its routing table.

Route maintenance: Once a route has been established, the source node will maintain the route for as long as it needs it. The movement of nodes not lying along the active route does not affect the routing to that path's destination.

B. Dynamic Source Routing (DSR).

DSR is a reactive routing protocol developed at Carnegie Mellon University, Pittsburgh USA, for the use of multi-hop wireless MANETs. DSR allows the network to be completely self-organizing and self-configuring [6]. The operation of DSR is done using the following two mechanisms: route discovery and route maintenance [5].
Route discovery: This is a mechanism by which a source node wishing to send a packet to a destination node dynamically obtains a path to the destination. Route discovery is used only when the source node does not know a route to the destination.
Route maintenance: This is performed when there is an error with an active route. When a node of the network that is part of some route notices that it cannot send packets to the next hop, it will create a message containing the addresses of the node that sent the packet and of the next hop that is unreachable; and send that to the source node.

C. Temporally-Ordered Routing Algorithm (TORA).

TORA is an efficient, highly adaptive, and scalable routing protocol based on the link reversal algorithm [10]. TORA provides multiple routes to transmit data packets between source and destination nodes of the MANET.

According to [6], the TORA protocol consists of three basic functions: creating routes, maintaining routes, and erasing routes. Creating routes corresponds to the selection of heights to form a directed sequence of links leading to the destination in a previously undirected network or portion of the network. Maintaining routes refers to adapting the routing structure in response to network topological changes. During this erasing routes process, routers set their heights to null and their adjacent links become undirected.

D. Optimized Link State Routing (OLSR).

OLSR is an MANET proactive routing protocol that uses the concept of Multi Point Relays (MPRs). MPR is an optimized flooding control protocol used by OLSR to construct and maintain routing tables by diffusing partial link state information to all nodes in the network [5].
The functioning of OLSR can be divided into the following three mechanisms:

- Neighbor/Link sensing.
- Efficient control flooding using MPR.
- Optimal route calculation using the shortest route algorithm.
-

III. RELATED WORK

Many researchers have studied MANETs routing protocols especially in terms of performance analysis. The next section presents some of the related work done on MANETs routing protocols.

A study by Gupta et al. [6] analyzed the performance of AODV, TORA and DSR using simulation. The simulator used for evaluation was Network Simulator version 2 (NS-2). The simulation was done in a rectangular field of 500m x 500m with 50 nodes. The traffic source used was CBR traffic and the simulation time was 2000s. The performance metrics used were Packet Delivery Fraction (PDF) and average end-to-end delay. From the results generated, it was concluded that the AODV protocol has the best overall performance. The result also demonstrated that the DSR protocol is suitable for networks with moderate mobility rate and since it has a low overhead that makes it suitable for low bandwidth and low power networks. The results also proved that TORA protocol is suitable for operation in large mobile networks having a dense population of nodes.

Ahmed and Alam (2005) [11] evaluated the effect of the load on the performance of TORA, DSR and AODV through simulation and the tool used for the simulation was OPNET modeler 10.5. For all the scenarios, the same movement models were used, and the MANET load was successively increased from 40, 60, 80 to 100 nodes. A square of 10 meters was used to define the area of the node's mobility. The simulation characteristics used in this research, were the control traffic received and sent, data traffic received, throughput, retransmission attempts, utilization, average power, route discovery time, and ULP traffic received. The results show that TORA shows a good performance for the control traffic received, control traffic sent, and data traffic sent. However, AODV shows better performance for data traffic received and throughput. DSR and AODV show poor performance as compared to TORA for the control traffic sent and throughput. However, TORA and AODV show an average level of performance for the data traffic received and

data traffic sent, respectively. The result also showed that for DSR, the number of packets in routing traffic received and sent, as well as the number of packets in total traffic received and sent, increase with increasing load.

De Rango et al. [5] presented a comparative analysis of DSR and OLSR from an energy point of view in MANETs. The objective of their study was to evaluate how DSR and OLSR affect the energy use of mobile nodes. The performance evaluation was through simulation and the simulator used was NS-2. The packet size was set to 512 bytes and the metrics used were: control overhead, data packets received, average end-to-end delay, throughput, connection expiration time, number of live nodes and energy consumption. The traffic used was CBR, fixed connection pattern and variable connection pattern. The results illustrated that the DSR protocol takes advantage of its routing policy, but the OLSR protocol can perform well with high traffic load and a variable traffic pattern. In the same work, De Rango et al. also stated that the route cache reply mechanisms activated on DSR can increase the data packet delivery and the protocol control overhead. However, the drawback of this approach is the increasing end-to-end data packet delay. The presented results also show that for the OLSR protocol, the link failure notification at the data link layer permits the delivered data packets to be considerably increased and the data throughout to be increased without expending more energy.

Kulla et al. [12] compared the performance of AODV and OLSR for different source and destination moving scenarios. They implemented a MANET testbed which provides the environment to make different measurements for indoor and outdoor communications. AODV and OLSR were implemented using four scenarios: Static Scenario, Source Moving Scenario, Destination Moving Scenario and Source-Destination Moving Scenario. The researchers performed the experiments in an indoor environment with the size nearly 70 m × 25 m. The packet size was fixed to 512 kilobytes and they used CBR over UDP to create the traffic. The performance metrics used were bit rate, delay, and packet loss. The results indicated that OLSR performs better than AODV in all the scenarios when both source nodes and destination nodes are moving during the communication.

A study by Naumov and Gross [2] analyzed the impact of the network size (up to 550 nodes), nodes mobility, nodes density and suggested data traffic on AODV and DSR performance. NS-2 was used since it supports the popular WaveLAN cards to study the performance of AODV and DSR in the areas of 2121 m × 425 m, 3000 m × 600 m, 3675 m × 735 m, 4250 m × 850 m, and 5000 m × 1000 m populated by 100, 200, 300, 400, and 550 mobile nodes, respectively. CBR was used for traffic sources. The performance metrics used were PDF, routing overhead and average end-to-end delay. The results indicated that in stationary scenarios with a low number of traffic sources, both protocols demonstrate good scalability with respect to the number and density of nodes. But as the mobility rate increases, the routing overhead of DSR prevent this protocol from delivering data packets effectively.

IV. METHODOLOGY

Routing algorithms are usually difficult to be formalized into mathematics [6]; they are instead tested using extensive simulation. Besides the difficulty to formalize these routing protocols into mathematics, there are two other great challenges: the cost and the difficulty of managing these routing protocols on large scale networks. From the related work done earlier, it appears that most of the research done in wireless networks today is done using simulators. This section presents the performance metrics used in this paper and the simulation setup of the MANET designed.

A. Performance Metrics.

The performance metrics evaluated in this paper are:

- Throughput: This is the sum of data packets generated by every source in the network. It is expressed in bits per second. So high throughput is desirable in wireless networks. The throughput reflects the completeness and accuracy of the routing protocol [6].
- Delay: This is the time it takes for a packet to be transmitted from the source node to the destination nodes. It is expressed in seconds. Short delay is desirable.

The throughput and the delay metrics are the most important performance metrics for traffic modeling [13].

B. Simulation Setup.

The MANETs to be modeled consists of nodes (in this paper, laptops were used) and a Wireless Local Area Network (WLAN) server. The nodes have applications running over TCP/IP and UDP/IP. The WLAN server has applications running over TCP. Depending on the scenarios, the WLAN server should be able to support Email applications. The performance evaluation of the routing protocols mentioned earlier was done using the discrete even simulator OPNET (Optimized Network Engineering Tools) version 14.0 [14]. The simulation models in this paper were run with nodes randomly distributed in an area of 1000 m × 1000 m. The nodes moved following the random waypoint mobility model with a speed of 5 meters per second and a pause time of 100 seconds. The protocols that were studied in the simulation are: DSR, AODV, OLSR and TORA.
In this paper, two scenarios were modeled:

- **Scenario 1**: this scenario consists of a MANET with a size of 30 mobile nodes
- **Scenario 2:** this scenario consists of a MANET with a size of 60 mobile nodes

The nodes in the MANET modeled supported a data rate transmission of 11Mbps with a power of 0.005 Watts. The packet size used for modeling was 1024 bytes. The MAC

protocol used was the IEEE 802.11b and the transmission range was set to 250 meters. Each scenario created was applied to each of the protocols during the simulation.

V. RESULTS AND DISCUSSION

In this section, the experiments results are presented and discussed. The performance analysis of the routing protocols AODV, DSR, OLSR and TORA are done according to the performance metrics cited earlier; that is based on the delay and the throughput. In terms of delay, TORA experiences oscillations due to the slow route reconstruction after a connection has been lost between nodes. Also in terms of delay, AODV and DSR routing protocols start to generate traffic only after a certain amount of time (simulation time) on a relatively dense network (scenario2); that is due to their route discovery mechanisms of reactive protocols in MANETs.

A. Delay Analysis under Scenario 1 and Scenario 2

The performance in terms of delay of AODV, DSR, OLSR and TORA routing protocols scenario 1 and scenario 2 using Email traffic is respectively shown Figure 1 and Figure 2.

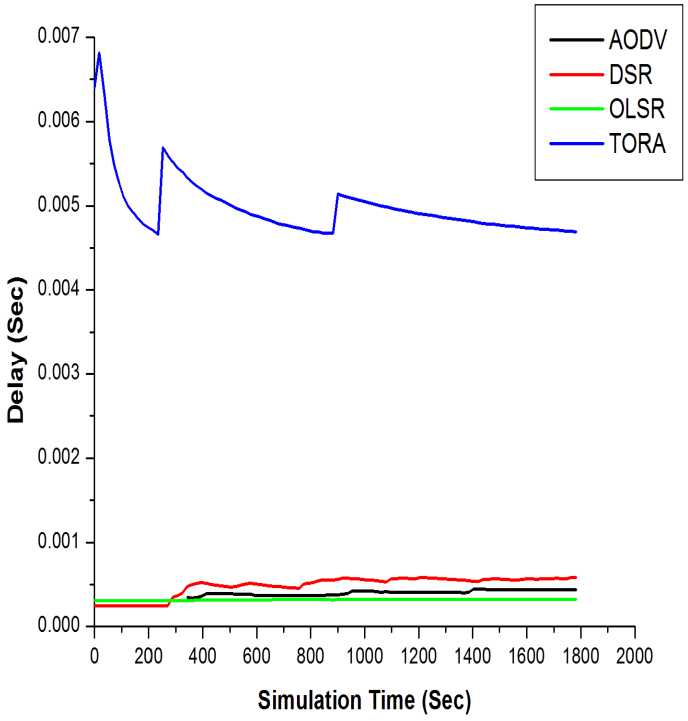


Figure 1: Delay of all the chosen routing protocols under scenario 1 and using email Traffic

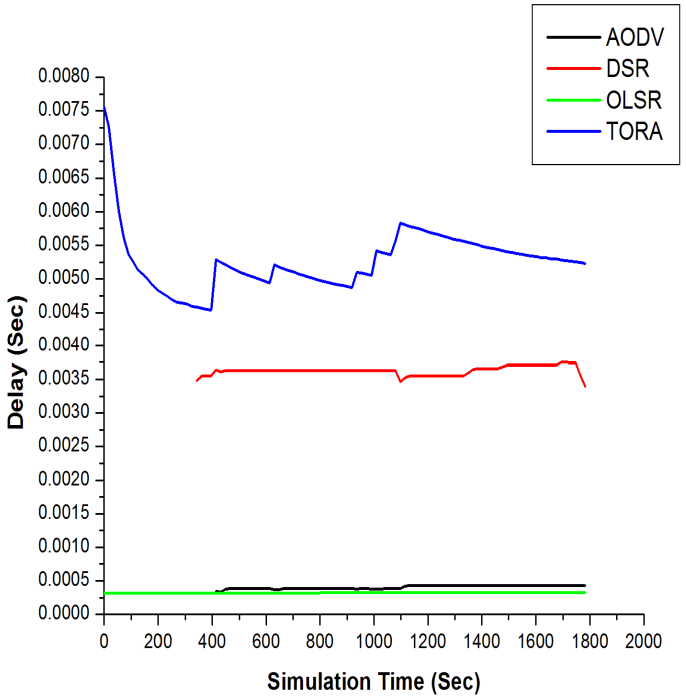


Figure 2: Delay of all the chosen routing protocols under scenario 2 using Email traffic

Figure 1 and Figure 2 indicate that under scenario1 and scenario 2 loads, the OLSR protocol have the shortest delay. The poor performance of TORA in terms of delay under scenario 1 and scenario 2 is due to fact that route rebuilding after a connection is lost may not occur as fast as in other reactive routing protocols [6] . This is due to the potential oscillations that may occur during this period. This is the basis behind the probable long delays encountered while waiting to determine the new routes. The DSR protocol has the second longer delay behind TORA under both scenarios; however the delay significantly increases with relatively dense MANET (scenario 2). The potential long delay experienced by DSR especially under the scenario 2 may be the result of wrong updates that could occur if its cache does not have the exact route to the destination node.

B. Throughput Analysis under Scenario 1 and Scenario 2

The performance in terms of throughput of the MANETs routing protocols AODV, DSR, OLSR and TORA under scenario 1 and scenario 2 using Email traffic is respectively shown in Figure 3 and Figure 4.

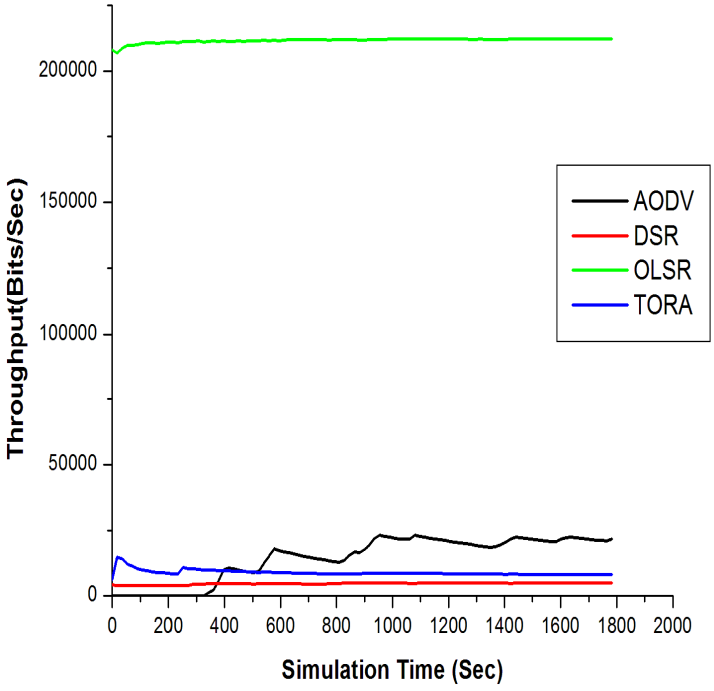


Figure 3: Throughput for all the chosen routing protocols scenario 1

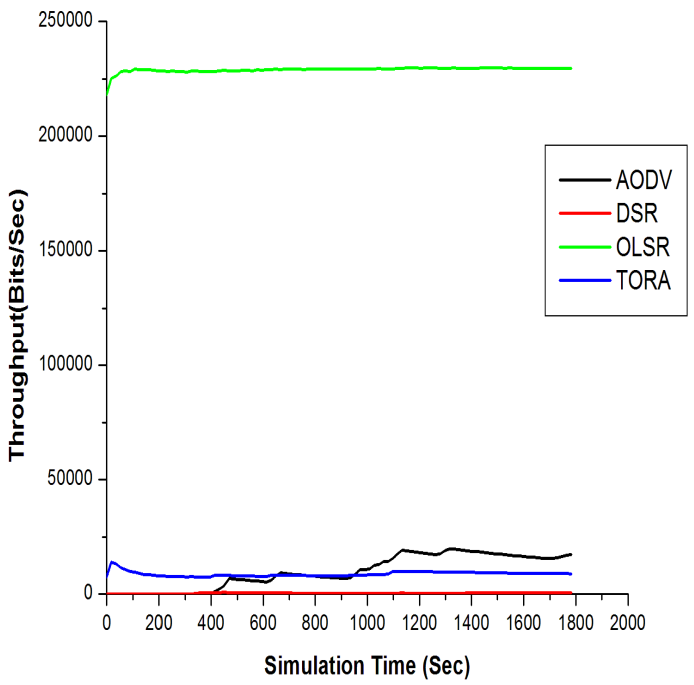


Figure 4: Throughput for all the chosen routing protocols under scenario 2

Figure 3 and Figure 4 show that the routing protocol OLSR outperforms the routing protocols AODV, DSR and TORA respectively under scenario 1 and scenario 2. This is due to the fact that OLSR does not need to find routes to the destination since all the paths are already available. Thus the source nodes are able to transmit more data packets when the OLSR routing algorithm is applied on the nodes. Under scenario 1 and scenario 2, Figure 3 and Figure 4 respectively show that the DSR routing protocol has the lowest throughput; that is due to its route discovery mechanism.

VI. CONCLUSIONS

From the results generated above, it can be concluded that:

- In terms of delay, OLSR has the shortest delay. DSR had the second longest delay behind TORA which had an extremely long delay. Still in terms of delay, it was observed that TORA oscillates and that was due to the time that TORA takes to rebuild the route after a link failure.
- In terms of throughput, OLSR outperformed AODV, DSR and TORA in all the scenarios. DSR had the lowest throughput. This is due to its route discovery process.

The overall results showed that the proactive routing protocol OLSR performed better than the reactive routing protocols AODV, DSR and TORA respectively under scenario 1 and scenario 2. One of the main reasons of the good performance of OLSR is that proactive routing protocols transmit control messages to all the nodes and update their routing information even if there is no actual routing request, hence the routes are always up to date. OLSR is therefore a routing protocol suitable for medium size MANETs.

VII. FUTURE WORK

The MANET modeled and designed in this paper uses the Random Waypoint as a mobility model. Further study could be done by modeling the Reference Group Point mobility model and using it as a mobility model under the same conditions as the ones used in this paper. Further study could also look at voice over IP traffic for the evaluation of MANETs under the same conditions as the ones used in this paper.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank the Department of Computer Science of the North West University, Mafikeng Campus and TELKOM CoE for making available the resources needed to conduct this research.

REFERENCES

[1] Irshad, E., Noshairwan, W., Usman, M., Irshad, A. and Gilani, M. 2008. Group Mobility in Mobile Ad hoc Networks. IADIS International Conference WWW/Internet 2008.

[2] Naumov, V. and Gross, T. 2005. Scalability of routing methods in ad hoc networks. *Performance Evaluation* 62 (2005) 193–209.

[3] Campos, C. A. V. and deMoraes, L. F.M. 2007. A Markovian Model Representation of Individual Mobility

Scenarios in Ad Hoc Networks and Its Evaluation. *EURASIP Journal on Wireless Communications and Networking* Volume 2007, Article ID 35946, 14 pages.

[4] Belding-Royer, E.M. and Perkins, C. E. 2003. Evolution and future directions of the ad hoc on-demand distance-vector routing protocol. *Ad Hoc Networks* 1 (2003) 125–150.

[5] De Rango, F., Cano, J. C., Fotino, M., Calafate, C. and Manzoni, P., Marano, S. 2008. OLSR vs DSR: A comparative analysis of proactive and reactive mechanisms from an energetic point of view in wireless ad hoc networks. *Computer Communications* 31 (2008) 3843–3854.

[6] Gupta, A. K., Sadawarti, H. and Verma, A. K. 2010. Performance analysis of AODV, DSR & TORA Routing Protocols. IACSIT *International Journal of Engineering and Technology*, Vol.2, No.2, April 2010, ISSN: 1793-8236.

[7] Bamis, A., Boukerche, A., Chatzigiannakis, I. and Nikoletseas, S. 2008. A mobility aware protocol synthesis for efficient routing in ad hoc mobile networks. *Computer Networks* 52 (2008) 130–154.

[8] Trung, H. D., Benjapolakul, W. and Duc, P. M. 2007. Performance evaluation and comparison of different ad hoc routing protocols. *Computer Communications* 30 (2007) 2478–2496.

[9] Bai, F., Sadagopan, N. and Helmy, A. 2003. The important framework for analyzing the impact of mobility on performance of routing protocols for Adhoc Networks. *Ad Hoc Networks* 1(2003) 383–403.

[10] Park, V. and Corson, S. 2001. Temporally-Ordered Routing Algorithm (TORA) Version 1 Functional Specification, IETF MANET Working Group, INTERNET-DRAFT, draft-ietf-manet-tora-spec-04.txt.

[11] S. Ahmed and M. S. Alam .Performance Evaluation of Important Ad Hoc Network Protocols, EURASIP Journal on Wireless Communications and Networking, Volume 2006, Article ID 78645, Pages 1–11

[12] Kulla, E., Ikeda, M., Barolli, L., Miho, R. and Kolic, V. 2010. Effects of Source and Destination Movement on MANET Performance Considering OLSR and AODV Protocols. 2010 13th International Conference on Network-Based Information Systems.

[13] Boukerche, A. 2004. Performance Evaluation of Routing Protocols for Ad Hoc Wireless Networks. *Mobile Networks and Applications* 9, 333–342, 2004.

[14] OPNET 14.0 Documentation.

Characterization and Evaluation : Temporal Properties of Real and Synthetic Datasets for DTN

Hemal Shah[1], Yogeshwar Kosta[2], and Vikrant Patel[1]

[1]Computer Engineering, Ganpat University, Mehsana, Gujarat, India

[2]Faculty of Technology, MEFGI, Rajkot, Gujarat, India

Abstract *- Node's movements play a significant role in disseminating messages in the intermittently connected mobile ad-hoc network. In such networks scenarios traditional end-to-end paths do not exist; mobility creates opportunities for nodes to connect and communicate when they are encountered. A series of encountering opportunities spread a message among many nodes and eventually deliver to the destination. Further improvements to the performance of message delivery can come from exploiting temporal properties of intermittent networks. It is modeled as time varying graphs, where, moving nodes are considered as vertices and contact opportunity to other nodes as an edge. The paper discusses about characterization and design of the temporal algorithm. Then, evaluating temporal distance and temporal centrality of real and synthetic data sets. Such, characterization can help in accurately understanding dynamic behavior and taking appropriate routing decision.*

Keywords: Temporal Graphs, Temporal Distance, Temporal Closeness Centrality, Temporal Path Length, Real Trace, Synthetic datasets.

1 Introduction

There are situations in mobile ad-hoc networks, where nodes are completely disconnected and may rely on relay nodes for contact opportunities to transfer the message. Such relay nodes create an opportunity for partial connectivity and carry the message until the next node or destination comes into contact[1]. In other networks, connectivity may exist, but only occasionally or intermittently. This intermittent connectivity is not failure or fault but, rather an integral part of dynamic networks. These networks are called Intermittently Connected Mobile Ad-hoc Networks (IC-MANET) also known as Delay Tolerant Networks (DTN)delay tolerant networks (DTNs)[2]. IC-MANET utilizes a Store-Carry-Forward[3] mechanism in which the intermediate node stores messages and forwards them to the nodes it encounters. In this manner, messages could be delivered to the destination hop-by-hop even if no stable end-to-end path exists. As network partition occurs frequently, if only one replica of the message[4] is kept, the message may reach the edge of partitioned network and be failed to be delivered at the destination. In order to increase the message delivery rate, each node can keep forwarded messages and copy them to other nodes it encounters. In this multiple-copy routing[5] manner, several replicas of the same message exist within the network.

In both single-copy routing and multiple-copy routing, the message delivery rate depends upon the node mobility, network connectivity and the intermediate node chosen strategy. Based on sufficient network connectivity, message delivery strategies should utilize node's mobility characteristics to increase the message delivery rate and reduce network overhead. Studies looked[6]at analyzing static networks, i.e., networks that do not change over time. Given the collections of measurements related to real network traces, authors[7] are quickly starting to realize that connections are inherently time varying and exhibit more dimensionality than aggregate analysis can capture.

In time varying graph or temporal graph[6][8] vertices represent the node and opportunistic contact between nodes represent edge or links. These links are changing over the time and raising interesting questions:

- Are there any metrics [9][10][11] evolved or proposed by researchers relating to temporal graphs in IC-MANET?
- If available, can they be used to analyze real and synthetic data sets?
- Can the time varying behavior of mobile ad-hoc network be used for designing IC-MANET routing algorithms?

This has motivated to contribute towards defining the metrics related to temporal graph. Then, designing the temporal algorithm to evaluate metrics from real trace and synthetic data sets. Author's contributions are:

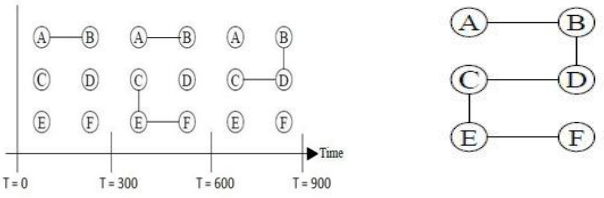
1. Modeling intermittent network as time varying graphs.
2. Defining temporal measurement: temporal distance and temporal centrality.
3. Design and development of temporal characterization algorithm.

4. Evaluation of real and synthetic trace datasets for temporal properties.

Section 2 discusses IC-MANET as time varying graphs and defining temporal metrics Section 3 discusses design temporal algorithm Section 4 presents application real and synthetic dataset evaluation of temporal properties. Section 5 discusses about conclusion and future work.

2 Temporal Graph and Temporal Metrics

In IC-MANET, it is observed[10] that the connections are inherently varying over time and exhibit more dimensionality[12] than static network analysis can capture. Static graphs treat all links as appearing at the same time. It is unable to capture key temporal characteristics, and gives an overestimate of potential paths, connection pairs of nodes which cannot provide any information on the delay associated with an information spreading process. Thus, to represent DTN as temporal graphs, the mobile nodes can be presented as vertices and opportunistic contact between nodes as an edge. It enables understanding duration of contact, inter-contact time, repeated contact, the time order of contact along a path on time interval basis. Temporal graph[6] is represented by sequence of time windows, for each window is considered a snapshot of the network at that time interval. The temporal distance and centrality metrics evolve over this view of the temporal graph retain the time ordering, repeated occurrences of contact between nodes, contact time and deletions of edges.



(a) Temporal Graph (b) Static Graph

Figure 1 Example of Temporal Graph with three time windows and six nodes

Consider the sequence of interaction as an example temporal graph (Figure 1(a)) and corresponding static aggregated graph (Figure 1(b)), where interactions between a pair of nodes defines an edge or, equivalently, generated from the union of all edges in the temporal graph[6]. Next, lets define the definition of temporal graph, path and distance.

Given a network trace starting at T_{min} and ending at T_{max}, a contact between nodes, i,j at time 's' is defined with the notation R_{ij}^{s}. A temporal graph[6] $\mathcal{G}_t^{\omega}(T_{min}, T_{max})$ with N nodes consists of a sequence of graphs $G_{t_{min}}$, $G_{T_{min}+w}$,…, $G_{T_{max}}$, where 'w' is the size of each time window unit e.g., seconds. Then, G_t consists of a set of nodes V and a set of edges E such that $i, j \in V$, if and only if, there exists R_{ij}^{s} with $t \leq s \leq t + w$. Our simplifying assumptions are :

We shall only consider unweighted graphs since, the datasets employed in this thesis (RollerNet, INFOCOM'06) contain only binary contact information. However, weighted temporal graphs would be a good candidate for future work. Secondly we shall refer to the set of nodes in a temporal graph as V = Vt, $\forall t \in [0, T)$ and N = $|V|$.

2.1 Temporal Metrics

The shortest path length on static graphs returns the number of hops from a source node to destination node; this does not retain temporal information and hence cannot capture the true duration or speed of dissemination. Instead, we now re-define a metric as the shortest temporal path length which gives an indication of the speed of message delivery from a source to destination. Before we can formalize this metric we first define the concepts of temporal paths. Following this, we then run through an example calculation of the temporal shortest path length and then define the algorithm that is used to compute the temporal distance.

2.1.1 Temporal Path

For given two nodes i and j temporal path defines as: $p_{ij}^{h}(T_{min}, T_{max})$To be the set of paths starting from i and finishing at j that passes through the nodes $n_1^{t_1} \ldots n_i^{t_i}$, where $t_{i-1} \leq t_i$ and $T_{min} \leq t_i \leq T_{max}$ is the time window, that node n is visited and h is the max hops within the same window t. There may be more than one shortest path.

2.1.2 Temporal Distance

Given two nodes i and j, the shortest temporal distance defines as $d_{ij}^{h}(T_{min}, T_{max})$ to be the shortest temporal path length, starting from time T_{min}, this can be thought as the number of time windows (or temporal hops) which takes for information to spread from a node i to node j. The horizon h indicates the maximum number of nodes within each window G_T through which information can be exchanged, or in practical terms, the speed that a message travels. In the case of temporally disconnected node pairs q,p i.e., information from q never reaches p, then set the temporal distance $d_{pq} = \infty$.

2.1.3 Temporal Betweeness Centrality

Betweenness is commonly used to discover nodes that are critical for mediating information flow[13]. To identify these mediating nodes, the static betweenness centrality of a node 'i' is defined as the proportion of shortest paths between all pairs of nodes that pass through 'i'. This proportion is important in that it gives a higher

weight to nodes which facilitate paths where there are no alternatives.

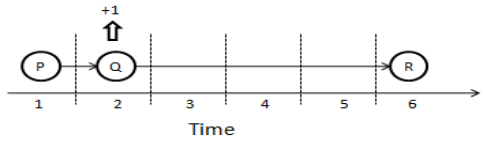


Figure 2 Using Temporal Path length betweenness centrality

To capture the notion of temporal betweenness it is important to take into account not only the proportion of shortest paths which pass through a node, but also the length of time for which a node along the shortest path retains a piece of information before forwarding it to the next node. For example, consider the 2-hop shortest temporal path from node 'P' to 'R', (P;Q;R) as shown in Figure 2.

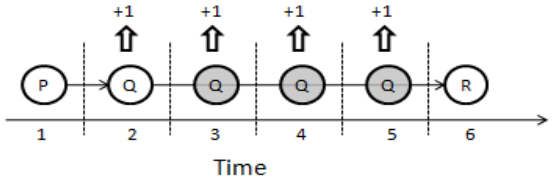


Figure 3 Betweenness centrality by consideration of time duration

In terms of time, this path could be represented as (P;Q;Q;Q;Q;R) since a piece of information resides on node 'Q' for 4 time windows, and so we want to assign a higher value as removing this node will have a greater impact in disrupting the network as shown in Figure 3

2.1.4 Temporal Closeness Centrality

The two nodes of a static graph are said to be close to each other if their geodesic distance is small. We can extend the definition of closeness to temporal graphs using the temporal shortest path length between nodes, which is a measure of how fast a source node can deliver a message to all the other nodes of the network.

Given the shortest temporal distance dij(T_{min}, T_{max}), temporal closeness centrality[7] can then be expressed as: $C_i^h = \frac{1}{W(N-1)} \sum_{j \neq i \in V} d_{i,j}^h$ So that, the nodes having, on average, shorter temporal distances to the other nodes are considered more central. Note that the subtraction from one is only required for a descending ranking.

3 Temporal Algorithm

Temporal distance $d_{ij}(T_{min}, T_{max})$, is computed in terms of number of time windows i.e.$d_{ij}(T_{min}, T_{max})$= $d_{ij}^t(T_{min}, T_{max})$. For each pair of i and j, algorithm computes $d_{ij}(T_{min}, T_{max})$ and then, takes the average of all values. This way temporal distance is computed in a number of time stamps. If average value multiplies with w, then result is the temporal distance in terms of time (in seconds). Eq. (1) gives the average temporal distance between T_{min} and T_{max} :

$$L(T_{min}, T_{max}) = \frac{\omega}{N(N-1)} \sum_{ij} d_{ij}(T_{min}, T_{max}) \quad (1)$$

3.1 Timewindow (w) Calculation

To understand the computation of Time window, refer Table 1 below showing calculation on dataset as an example, where each cell value represents the total contact time between a particular pair i,j divided by total the number of contact occurrences. For each node pair (i,j) compute a sum of all values. It returns the average meeting time per contact. The optimal value of time window is greater than average meeting time, because if time window<= average meeting time, then in most of the time windows, number of contact occurrence will be around one. That means, the information cannot be diffused efficiently into the network.

Table 1 Time window Calculation

Node ID	1	2	$\sum \frac{\text{Total Contact Time}}{\text{Total No. of Occurances}}$
1	0/0	480/2	480/2
2	500/2	0/0	500/2
$\frac{\sum T_{ij}(T_{min}, T_{max})}{\sum N_{ij}}$			$\frac{980}{4}$ = 245 Time Window size

From above calculations it is established that, for effective information diffusion process into the network optimal time window should be greater than $\frac{\sum T_{ij}(T_{min}, T_{max})}{\sum N_{ij}}$ .E.g. In Figure 1(a) total number of time window = (T_{max} – T_{min})/w = (900 – 0)/300 = 3 timestamps, assuming time window size = 300.

Let's find temporal distance $d_{ij}^t(t_0, t_{900})$ for the temporal graph shown in Figure 1(a). Here, T_{min} = 0 and T_{max} = 900. Time window size = 300. Thus, there are three time windows t1, t2 and t3.

3.2 Computation of Temporal Distance

Before starting calculation of temporal distance of each pair i,j, initialize number of empty lists equal to that of calculating number of time window. For each pair (i, j), i≠j, start scanning timestamps from 1 to 3. For each timestamp, add occurred node id into the respective list of timestamp. Pair of node (i,j) occurs whenever there is a contact edge between node pair (i,j).

3.2.1 Preconditions

Pair of node (i,j) occurs whenever there is a contact edge between node pair (i,j).

Case 1:

If I == j then, return 0, in computing matrix below, temporal distance (A, A) = (B, B) = (C, C) = (D, D) = 0.

Case 2:

If both i and j occurs in same timestamp then return (jth timestamp number – ith timestamp number) or return (0). In Figure 1(a), node A and node B occurs in same timestamp no. 1 , so the temporal distance between A and B is (B's timestamp no. – A's timestamp no.) = (1-1) = 0 timestamps.

Case 3:

If i occurs earlier than j, then search occurrences of j in consecutive timestamps by using other occurred nodes in same timestamp in which i has occurred; for each pair i,j it may give more than one path in terms of required timestamp, in such a case select the shortest timestamp. In Figure 1(a), for temporal distance (A,D), node A occurred in timestamp number 1 and node D occurred in timestamp number 3. Also, there is an intermediate node B which is common between node A and node D. So temporal distance (A,D) = (node D's timestamp number – node A's timestamp number) = (3 – 1) = 2 timestamps.

Case 4:

If i occurs and j do not occur during a consecutive timestamp till T_{max}, then the temporal path between a pair of i, j is not possible. So, return ∞. Figure 1 (a), for a temporal distance (D, E), node D occurred in timestamp number 3. But there are no occurrences of node E also by using other intermediate occurrences of other nodes. So temporal distance (D,E) = ∞. In continuation of the example shown in Figure 1(a) and successive computation of temporal distance matrix as shown in Figure 2, the sum of non-negative values of matrix = 10. Now, calculate the average temporal distance metric: 300 (10/ (6) (5)) =3000/ (30) =100. i.e., it takes average 100 seconds to reach from source 'i' to destination 'j'.

Temporal Distance Matrix =

```
[ [ 0,  0,  1,  1  -1,  -1 ],
  [ 0,  0,  1,  1,  -1,  -1 ],
    1,  1,  0,  1,  0,   0  ],
    1,  0,  0,  0,  -1,  -1 ],
    1,  1,  0,  1,  0,   0  ],
[ -1,  1,  0,  1,  0,   0  ]]
```

Figure 2 computed temporal distance matrix values

3.3 Algorithm

1. Input source and target, T_{min} and T_{max} time window Time Window Equation:

$$w(Timewindow) > \frac{\sum T_{ij}(T_{min}, T_{max})}{\sum N_{ij}} \quad (2)$$

Where, $\sum T_{ij}(T_{min}, T_{max})$ = Total contact time between all pairs of nodes I, j and $\sum N_{ij}$ = Total occurrences of all pairs of nodes i and j.

2. Number of times frames = T_{max} – T_{min} / Time window.

3. Initialize number of empty list equal to number of time frames. Each list shows node ids whose contact occurred in a respective time frame.

4. Read the dataset and perform a lookup for node contact in different time frames and generate a distance matrix for each node. Per contact frame, fill up the array / list with node ids in contact.

5. Compute the temporal distance as:

   a. If source and target ids are in the same list, return (target time frame number – source time frame number) as temporal distance.

   b. Otherwise, look up source and target in different time frames. If the source time frame < target time frame then return (target time frame number – source time frame number) as temporal distance.

   c. In case repeated occurrence of the source, target sets T_{min}= last target occurred +1 timestamp and repeat steps a and b.

6. Take average values of all pairs (source, target) temporal distance.

7. Repeat steps 4,5,6 and 7 for all pairs(source, target) and generate matrix. Minus one (-1) indicates no edge between a pair of nodes in the matrix.

4 Application of Temporal Algorithm

First network topology is generated from large real data sets using python custom made script. Python provides a module called networkX , which helps to generate network topology according to the dataset. For evaluation, we have downloaded the INFOCOM'06, RollerNet real trace data from CRAWDAD and Random Way Point (RWP) generated using ONE simulator. Real and Synthetic data set information are presented in Table 2 (a) and Table 2 (b).

Table 2 (a) RollerNet and RWP Dataset details

Datasets	RollerNet	RWP_63
Start Date	2/2/2009	NA
Duration	0.12 days	0.12 days
T_{min}-T_{max}	Day 1: 0-3096(51.6 min)	Day 1: 0- 3096
Number of Nodes	63	63
Contacts	80824	576

Table 2 (b) INFOCOM'06 and RWP Dataset details

Datasets	INFOCOM'06	RWP_98
Start Date	13/03/2005	NA
Duration	4 days	1 day
T_{min}-T_{max}	Day 1: 61260 - 86400 (6.98 hours) Day 2: 86400 - 172800 (24 hours) Day 3: 172800 - 259200 (24 hours) Day 4: 259200 – 345600 (24 hours)	Day 1: 0-342915
Number of Nodes	98	98
Contacts	118875(of all four days)	4412929

For any real trace or synthetic dataset, it is required to set the common format as shown in Table 3 in order to carry out time window calculations. The customized scripts are written taking care of converting datasets into desired format.

Table 3 Common format for data sets

Source Node ID	Destination Node ID	CONNECTION UP Time	CONNECTION DOWN Time	Occurrence Count	Inter Contact Time
1	3	51293	51293	1	0
1	3	60603	60603	2	9310
1	3	62363	62363	3	1760
1	3	79649	79649	4	17286

4.1 Time Window Calculation

The time window is calculated as discussed in Section 3.1 using customized script. For INFOCOM'06 day1 number of timestamps are 7, for day 2,3, and 4 are 26 ; timewindow size for all four days is 3240. For Rollernet timestamps are 207 and time window size is 15. For RWP_63, timestamps are 44 and time window size is 71.RWP_xx is used for comparison with real traces and xx denotes number of nodes. For RWP_98 time stamps are 2598 and window size is 132.It is observed that keeping too small size of the window results in an increase in timestamps. It means node pairs (source, target) are at a distance and requires considerable timestamps to reach a target.

4.2 Temporal Distance

Temporal algorithm uses the value of T_{min}, T_{max}, nodes, number of connections and timestamps, time window size as input computed and presented in Table 3. Then, it evaluates the temporal distance and centrality metrics as below.

Table 4 Temporal Distance evaluation for data sets

Dataset Details	Number of Timestamps	Time window (w) (seconds)	Static Distance	Average Temporal Distance
INFOCOM'06				
Day 1	7	3240	1.56	3.97
Day 2	26	3240	1.23	14.26
Day 3	26	3240	1.3	12.8
Day 4	26	3240	1.3	11.75
RollerNet				
Day 1	207	15	1.22	297
RWP_63				
Day 1	44	71	3.64	0.46
RWP_98				
Day 1	2598	132	1.81	14.94

Temporal distance values presented in Table 4 gives us a better understanding of the network. Since, it can provide us an accurate measure of the delay of the information diffusion process that is not possible with traditional static metrics. In particular, since static shortest paths ignore time-order of contacts, it over-estimate the availability of contacts and therefore under-estimates the true shortest path.

4.3 Centrality Evaluation

The existence of special nodes is vital due to their strong impact on message delivery. Degree centrality is measured as the number of direct ties that involves a given node. The node with highest degree centrality contacts the largest number of nodes in the network, so it is suitable to choose this node to act as a forwarder.Closeness centrality of nodes defines how long it takes information to spread from a given node to other nodes.The node with higher betweenness centrality has more chance to assist the communication on the link between two nodes. Table 5 shows the evaluated values of different centrality for real traces.

Table 5 Centrality evaluation for data sets

Dataset Details	Diameter	Degree Centrality	Betweeness Centrality	Closeness Centrality
INFOCOM'06				
Day 1	4	(27, 0.81)	(85, 0.04)	(40, 0.83)
Day 2	3	(56, 0.98)	(16, 0.02)	(56, 0.98)
Day 3	3	(48, 0.95)	(51, 0.01)	(48, 0.95)
Day 4	3	(44, 0.77)	(30, 0.05)	(44, 0.80)
RollerNet				
Day 1	2	(51, 0.1)	(51,0.0003)	(51, 0.1)
RWP_63				
Day 1	6	(14, 0.29)	(14, 0.65)	(15, 0.42)
RWP_98				
Day 1	2	(17, 0.99)	(17, 0.09)	(17, 0.99)

It is observed that the values of temporal distance and centrality for the INFOCOM'06, RollerNet and RWP presented in section 4.2 and 4.3 are difficult to compare. Rather, our objective is to pop out time varying properties for efficient IC-MANET routing decision.

4.4 Observations

- Optimal time window size varies as per number of connections between nodes, number of nodes and total duration of T_{min}, T_{max}. Keeping the value < derived through script may result in overlooking connection and keeping too high will result in wastage of network resources.

- It found that synthetic data set values for temporal distance is poorer than real trees. This is due to its characteristics moving towards the center and random nature of the movement. Average temporal distance values of real trace analysis enables better routing decision and it is more accurate than static analysis.

- Diameter values can be used for evaluating maximum hops per time frame basis or on an average.

- Different centrality values help in identifying the important nodes. Such nodes can assist in the efficient information dissemination process.

- Referring the readings of RollerNet and RWP_63 : It reveals that in RWP model the node movements are random and hence, the number of contacts and time stamps are less, resulting in lower average temporal distance value. It is seen that most of the time the nodes are moving around center due to which diameter, degree centrality, betweenness and closeness values are higher. These values clearly indicate the reasons (described above) behind not using the synthetic models for realistic scenarios. On the other hand, RollerNet data have comparatively higher contacts, and higher number of timestamps resulting better connectivity. Therefore, for efficient information dissemination these characteristics of dataset are being used by routing engines.

5 Conclusions

It reveals that the node mobility plays a vital role for the efficient diffusion of information in challenging environment. And while doing so one cannot ignore to understand the movement patterns and related properties such as time order, frequency, contact duration, inter contact time, etc. These dynamic properties of connection are first analyzed and understood by using time varying matrices: temporal distance, diameter and centrality. General framework has been to design carrying capability of evaluating temporal metrics from any synthetic and real trace data. Because such frameworks help in computing number of time frames and the size of time windows which in turn calculate temporal distance. These properties are very useful in designing the DTN routing protocol and understanding the dynamics of the network and thereby taking forwarding or replication decision.

6 Acknowledgment

We express our sincere gratitude to the management of Ganpat University – Mehsana and Marwadi Education Foundation - Rajkot; for providing us research opportunities and their wholehearted support for such activities. Finally, our acknowledgement cannot end without thanking the authors whose research papers helped us in making this research.

7 References

[1] W. Zhao, Y. Chen, M. Ammar, M. Corner, B. Levine, and E. Zegura, "Capacity Enhancement using Throwboxes in DTNs," in *2006 IEEE International Conference on Mobile Ad Hoc and Sensor Sysetems*, 2006, pp. 31–40.

[2] K. Fall, K. L. Scott, S. C. Burleigh, L. Torgerson, A. J. Hooke, H. S. Weiss, R. C. Durst, and V. Cerf, "Delay-Tolerant Networking Architecture," *IETF*, vol. 54, no. 4838, 2007.

[3] B. D. Davison, "Store-and-Forward Performance in a DTN," in *2006 IEEE 63rd Vehicular Technology Conference*, 2006, vol. 1, pp. 187–191.

[4] W. Zhao, M. Ammar, and E. Zegura, "A message ferrying approach for data delivery in sparse mobile ad hoc networks," in *Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing - MobiHoc '04*, 2004.

[5] Amin Vahdat and David Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," *Tech. Rep. CS-200006*, 2000.

[6] V. Kostakos, "Temporal graphs," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 6, pp. 1007–1023, Mar. 2009.

[7] N. Santoro, W. Quattrociocchi, P. Flocchini, A. Casteigts, and F. Amblard, "Time-Varying Graphs and Social Network Analysis: Temporal Indicators and Metrics," Feb. 2011.

[8] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro, "Time-Varying Graphs and Dynamic Networks," *CoRR*, Nov. 2010.

[9] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005.

[10] J. Tang, M. Musolesi, C. Mascolo, and V. Latora, "Temporal distance metrics for social network analysis," in *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*, 2009.

[11] A. Clauset and N. Eagle, "Persistence and periodicity in a dynamic proximity network," *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, Nov. 2012.

[12] J. Tang, M. Musolesi, C. Mascolo, and V. Latora, "Characterising temporal distance and reachability in mobile and online social networks," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, Jan. 2010.

[13] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, "Analysing information flows and key mediators through temporal centrality metrics," in *Proceedings of the 3rd Workshop on Social Network Systems - SNS '10*, 2010, pp. 1–6.

Evaluation of Modulo in a Multi-Channel 802.11 Wireless Network

Dr A. Paraskelidis, Dr Mo Adda.
Pervasive Computing Research Group, School of Computing, University of Portsmouth, Portsmouth, United Kingdom.

Abstract - *Since the introduction of the IEEE 802.11 standard, researchers have moved from the concept of deploying a single channel and proposed the utilisation of multiple channels within a wireless network. This new scheme posed a new problem, the ability to coordinate the various channels and the majority of the proposed works focus on mechanisms that would reduce the adjacent channel interference caused by the use of partially overlapping channels. The proposed idea in this paper borrows the concept of network segregation, firstly introduced for security purposes in wired networks, by dividing a wireless network into smaller independent subnetworks and in collaboration with a channel assignment, the Modulo. Modulo defines a set of rules that nodes should obey to when they transmit data. The utilization of multiple channels under the guidance of Modulo for each subnetwork, proves to improve the performance of an ad-hoc network even in noisy environments.*

Keywords: networks, ad-hoc, interference, segregate networks, modulo, throughput.

1 Introduction

Ad-hoc wireless networks provide a means of networking together groups of computing devices without the need for any existing infrastructure. Devices automatically form a network when within range of each other, and also act as routing nodes by forwarding any packets not intended for them.

A single channel for transmission is not always enough and in high traffic routes, a single channel device can create more problems than it can solve. Common problems with wireless networks are interference, multipath and attenuation. All these prevent the wireless networks from performing to their maximum capabilities. Places and environments, which accommodate all the above-mentioned problems, make the existence and deployment of wireless LANs highly restrictive.

In this paper we examine the impact of utilising multi-channel technology within a legacy 802.11g network. Our target is to investigate the performance of segregated multi-channel mesh network and a simple, single channel wireless network - WLAN. The term segregated means that the network is divided into smaller subnetworks and each one operates at different frequencies than others.

2 Literature Review

Node placement and deployment play a crucial role to the network stability and performance. During node placement, variable environment characteristics such as sources of interference and area morphology like physical obstacles and constructions should be taken seriously into consideration. This way it is easier to adjust the deployed wireless network to those needs, achieving maximum operability and performance.

2.1 Channel Assignment Algorithms

To reduce interference, neighbouring nodes should operate in different frequency channels. For example the IEEE 802.11b standard for wireless LANs can operate simultaneously in three non overlapping channels (1, 6 and 11) [1] without each node to interfere with each other. During our testing we used the multi-hop infrastructure which has been proved [2] to overcome many problems of the single-hop networks.

In the multi-hop infrastructure, a node may find many routes to access different access points, potentially operating on different channels. Kyasamur and Vaidya So et al. [3] proposed a routing and channel assignment protocol which is was based on traffic load information.

The proposed protocol successfully adapted to changing traffic conditions and improved performance over a single-channel protocol and one with random channel assignment

Bahl et al. [4] suggested a link-layer protocol called SSCH that increases the capacity of an IEEE 802.11 network by utilizing frequency diversity. Nodes are aware of each other's channel hopping schedules and are also free to change their schedule.

Raniwala et al. [5] developed a wireless mesh network architecture called Hyacinth. This architecture equips each node with multiple IEEE 802.11a NICs supporting distributed channel assignment/routing to increase the overall throughput of the network. Apart from that, there are other proposals [6] and [7] which in fact require proprietary MAC protocols. They propose something like a packet-by-packet channel switching which resulted in an increased time per transmission. More MAC modifications were proposed in [8] to support beamforming, whereas [9] and [10] required a separate radio to communicate firstly with the neighbours and then start transmission. These approaches are under utilizing a channel just for configuration set up whereas it could be used in a more efficient and useful way.

3 Systems Architecture & Evaluation

In the case of an industrial environment, the problems can be more persistent and result in really bad quality of service even of no service. The problem of broken links has been mainly encountered by the deployment of multi-channel networks.

Range is crucial during deployment and operation as it defines and the amount of wireless nodes that should be used for the full coverage of the required area. In wireless networks the number of the devices deployed can have advantages and disadvantages. The main advantage is the best signal coverage throughout the area. On the other hand the main disadvantage is the appearance of interference between the operating wireless nodes. Interference comes into two forms, the co-channel interference (CCI) for devices operating in the same frequency [11] and the adjacent channel interference (ACI) when nodes operate in different frequency spaces [12] but they are close enough to each other

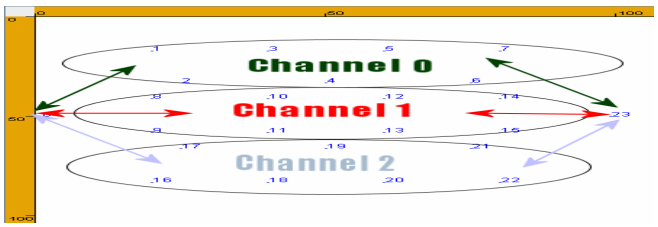


Fig. 1 A sample of a 24 node segregate network using three different channels.

Throughout the experiments that take place we assume that there is no limit to the number of channels that can be used. Although IEEE802.11 sets a limit to the available channels, in our case we emphasize on a more standard independent approach able to operate in all available technologies.

In previous approach [13], we showed that by segregating a network we can achieve better network performance. Current target was to improve further by using more channels inside the segregated network. There are three main steps to achieve that. The first step was to simulate a single channel network, then to divide the network into a variable number of subnetworks and use one different channel for each subnetwork and finally the multichannel approach by using more than one channel within each subnetwork.

3.1 Single channel network

This is the simplest form of a wireless network. A number of nodes able to relay data from one side to the other by using one channel only. This approach is used only for benchmark reasons in order to be able to decide if any improvement has been achieved. Routing protocol used is the Ad hoc On-Demand Distance Vector (AODV) [14] in a standard mode, no multichannel enabled.

3.2 Segregate network using single channel

The approach is the same as explained in figure (1) and figure (2). It should be made clear that nodes don't always follow the configuration given in figure (1) as they are usually placed randomly in the simulated area.

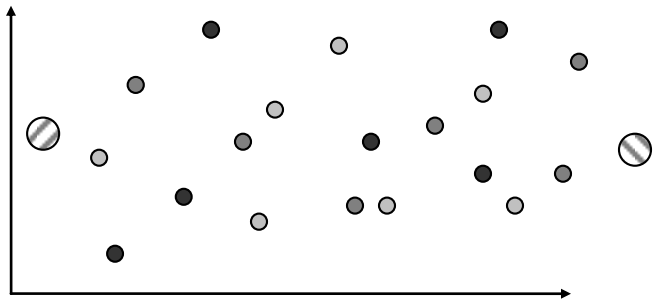


Fig. 2 A segregate network of 21 nodes. The side nodes operate in all the three channels available. All the rest nodes operate in different channels as separated from their colors.

We start dividing the network into smaller subnetworks and watch if there is improvement over this segregation. Channels are randomly chosen during transmission by the edge nodes, whilst inside each subnetwork since there is only one channel operating and the routing is done using AODV multichannel enabled [15] in both cases.

The best way to describe a segregated network is with the help of the parameters that affect it. First, we call *S* the segregated network, *n* the total number of nodes, *g* the number of subnetworks and finally *k* the number of channels for each subnetwork, which in this scenario is always equal to 1, then *S* would be expressed as:

$$S\ (n,g,1) \tag{1}$$

3.3 Segregate networks using modulo

In this case, each subnetwork is operating into more than one frequency channel. Again the frequencies in one subnetwork $\{k1, k3, k5 \ldots k_n\}$ differ from the frequencies operating in the other $\{k2, k4, k6 \ldots k_{n+1}\}$. Again, the number of channels existing in one subnetwork will be the same to all the rest. Based on equation (1), for the current case, the total number of channels T_k equals to,

$$T_k = g*k \tag{2}$$

and the number of available nodes within every subnetwork

$$T_n = n / g \tag{3}$$

The increase rate of the delay is reduced as the network is segregated into more subnetworks due to the smaller density λ of nodes that operate in the same channel. Take a single channel network where all nodes operate on the same frequency, when segregation is applied, the density λ of nodes operating on the same channel within a unit area is decreased. Let T_N be the number of nodes listening to the same channel and α the size of the simulated area then λ would be expressed as in equation (4)

$$\lambda = (T_N/g)/\alpha = T_N/(\alpha * g) \tag{4}$$

The density of a single node network λ_s with transmission range R_{tx} is

$$\lambda_S = 1/\pi R_{tx}^2 \tag{5}$$

From equations (4) and (5) we define the density and the number of segregate networks to maintain connectivity between the nodes of each segregate network

$$\lambda \geq \lambda_S \Rightarrow \lambda \geq 1/\pi R_{tx}^2 \Rightarrow T_N/(\alpha * g) \geq 1/\pi R_{tx}^2 \Rightarrow \Rightarrow g \leq T_N\ \pi\ R_{tx}^2 / \alpha \tag{6}$$

The limitations of density λ are demonstrated in figure (6).

With the introduction of multiple channels inside each subnetwork, modulo was utilised to coordinate the channel assignment decisions of each node. The switching technique is based on modulo algorithm [16] shown in figure (3).

A node, upon receiving a data packet on a channel k, transmits it on the next channel $k+1$, where $k+1$ is next channel greater than the current one in rank. In general, the channel that is in use at hop *h*, given a starting channel *k* and *e* channels available can be expressed as:

$$f_n = (n+k)\ mod\ c \tag{7}$$

A graphical representation of the modulo technique is shown below.

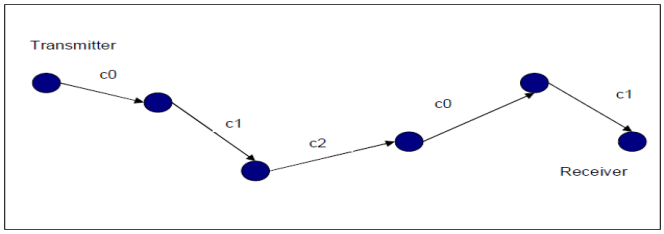


Fig. 3 Modulo channel allocation using three frequency channels.

Modulo adopts a store and forward packet transmission mechanism for every single packet that travels through the multi-hop path defined by AODV [12] and this mechanism is shown in figure (4).

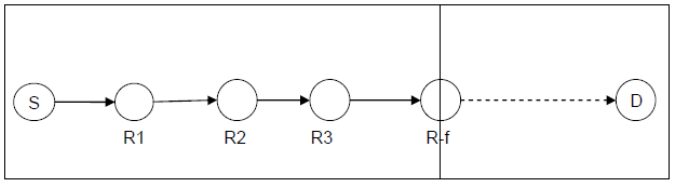


Fig. 4 Modulo channel allocation using four frequency channels.

S is the source node, *D* is the destination node and all the rest are the intermediate nodes between source and destination. *R-f* is the last node that interferes with the transmission of *S* and after the *R-f* node all remaining nodes can transmit using the same frequency with *S* without interfering. The position of *R-f* depends on the transmission range and the location of *S*.

Let denote T_h the transmission time between two adjacent nodes as *R1* and *R2* or *S* and *R1* and let assume that there are *m* chain nodes distributed randomly within the subnetwork of a segregated network *S* (*n*, *g*, *k*), where *g* is the number of segregated networks and *k* the number of channels in each subnetwork. The value of *m* is a number smaller or equal to the number of member nodes of a single subnetwork.

$$m \le n / g \tag{8}$$

The source station is sending N_p number of packets of length *L* (bytes). The packet may be segmented into fragments *F* with each fragment being acknowledged by an acknowledgement packet *A*. If no acknowledgment is required, then a fragmentation is not required and *L* is equal to *A*. With *S* being the only injection of traffic source, the end-to-end delay is,

$$T = (m+1) * T_h \tag{9}$$

The total transmission time T_S of N_p packets will equal to,

$$T_s = m * T_h + (N_p + f(\left\lceil \frac{N_p}{k} \right\rceil - 1)) * (T_h + T_a) \tag{10}$$

where T_h is the transmission delay for one packet within a single hop, T_a is the transmission delay of a single acknowledgment packet (34 bytes), *f* describes R-f as explained above, *k* is the number of channels utilised in the subnetwork. Equation (10) shows the dependency between the number of packets that have to be transmitted, the amount of channels utilised within each segregate network and finally the interference range. This equation applies to every segregate network separately and not to the whole network. The upper limit indicator ensures that the outcome of the division between N_p and *k* is always an integer. Since modulo technique is trying to achieve concurrent transmissions in a chain of nodes, the maximum achievable number of these concurrent transmissions are related to how many packets have to be sent. The number of channels which are available and how many of them will actually be used is related to the interference range *f*. Consider the scenario where four packets have to be transmitted, there are two channels available the interference range is equal to two and the total nodes in the chain equals to eight. Equation (10) shows that once the first two packets are transmitted, they should be two hops away with the aim of achieving another two concurrent transmissions for the next packets in the queue. Having eight nodes in the chain, modulo can achieve four concurrent transmissions of the four packets. If interference range was larger than two, then the concurrent transmissions for the whole length of the chain would be less.

Finally, the capacity C_s of the transmission measured in packets/second is calculated as,

$$C_s = N_p / T_S \tag{11}$$

Each time *S* transmits a packet to node *R1* on channel *k*, the packet is stored temporarily in the node and an acknowledgment (ACK) is sent to the source node. Once the ACK is received, the packet is transmitted to node *R2* on channel *k*+1 and at the same time node *S* sends the next packet to node *R1*. This way all nodes can transmit simultaneously only if there are enough available channels for utilisation. If there are only two channels available then only two nodes can communicate simultaneously. The transmissions of ACKs don't affect the network's performance as long as multiple channels are used.

4 Methodology

Some of the scenarios presented and investigated in this paper are difficult to investigate and deploy in the real world, thus the best way to gather information is through mathematical analysis simulations performed using one of the network simulators available. The simulator used is

GlomoSim v2.03 [17], a well known widely used and free to use tool able to simulate wireless and wired networks systems. It has been designed using the parallel discrete-event simulating capability provided by Parsec.

5 Results

First of all we start with the simulation results of a wireless network using just one channel, the most basic form of a wireless network, without any segregation. It should be made clear that only delay is presented and evaluated at the moment, due to the big variety of the scenarios. Next, there is a mathematical analysis and evaluation of the modulo approach based on equations (10) and (11). For given scenarios we test the validity of our mathematical model against previously published results that were based on simulations results. The following figures confirm our previous simulations based results [18] [19] [20] and satisfy the design purpose of modulo.

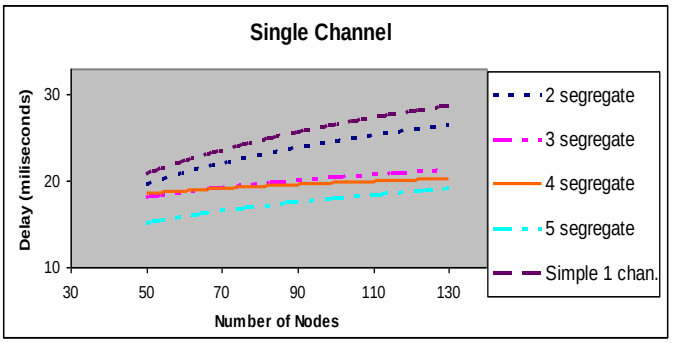


Fig. 5 The average delay of the networks for a variable number of nodes.

As we can see from figure (5), the segregate network operates quite well and overcomes in terms of delay the basic configuration. Something that was expected as it operates in a single channel, thus interference and the luck of multiple routes increases the delay. This first, figure (5), is the base for the comparisons for the segregate network using modulo.

The following results are based on scenarios trying to calculate the transmission time T_S and capacity C_s improvements that modulo offers within a segregated network utilising multiple channels. Consider the scenario where there is a chain of nodes for variable numbers of transmitted packets N_p and variable utilised channels k. The rate of transmission is set to 11Mbps, and initially m is set to 6 nodes and f equals to 4 nodes, although this values may change for comparison reasons. No ACKs are required and a single packet is 1375 Bytes long, resulting to a T_h of 1 millisecond and finally N_p gets values of 6000, 9500 and 13000 packets respectively.

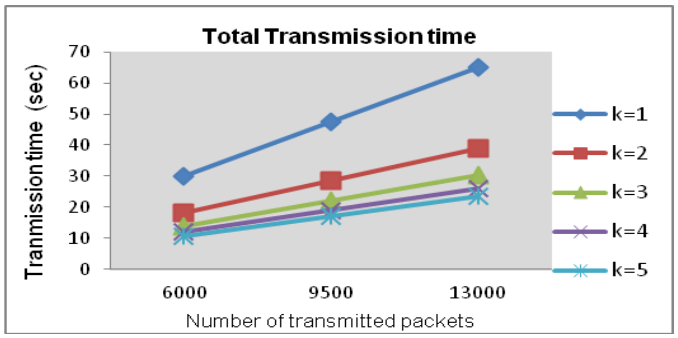


Fig. 6 Transmission time improvement over utilised channels for f=4 and m=6.

Figure (6) presents the improvement of the total transmission time of a single chain of nodes utilising variable numbers of channels while f is equal to 4 nodes and there are 6 nodes in the chain used for the transmission. With the utilisation of a second channel in the chain, the transmission time is improved significantly, and this improvement continues with the addition of extra channels, although with a smaller rate. At the end, with the use of 5 channels, T_S has achieved an improvement of 45 seconds over the single channel scenario when N_p = 13000.

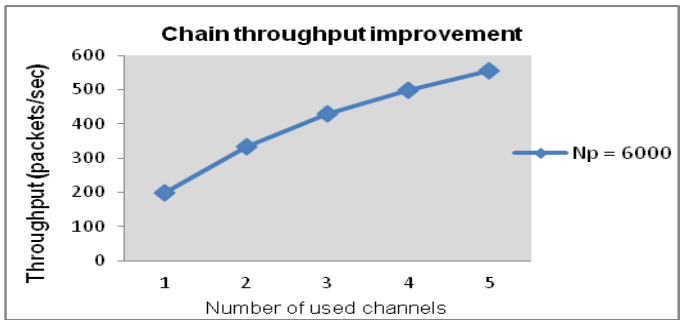


Fig. 7 Chain throughput improvement for f=4 and m=6.

Figure (7) presents the improvement of the throughput of a single chain of nodes utilising variable number of channels while f equals to 4 nodes and there are 6 nodes in the chain used for transmission. By adding extra channels the capacity of the chain is increased following the same rate as the transmission time. For = 6000, there is an increase of 255 packets/sec when 5 channels are utilised within the chain. The same is trend is followed for N_p = 9500 and N_p= 13000

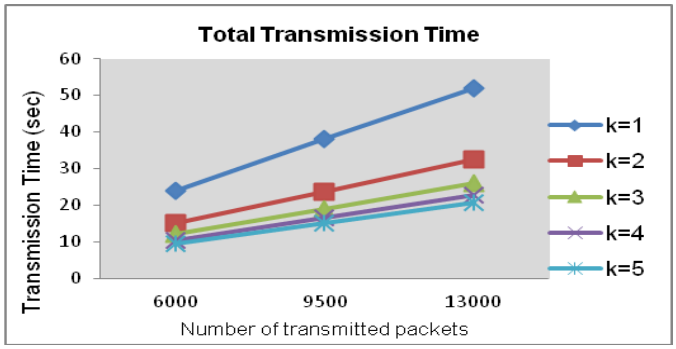


Fig. 8 Transmission time improvement for f=3 and m=6.

Figure (8) presents the improvement of the total transmission time of a single chain of nodes utilising variable numbers of channels while f is equal to 3 nodes and there are 6 nodes in the chain used for the transmission. With the utilisation of a second channel in the chain, the transmission time is improved significantly, and this improvement continues with the addition of extra channels, although with a smaller rate. At the end, with the use of 5 channels, T_S has achieved an improvement of 45 seconds over the single channel scenario when N_p = 13000.

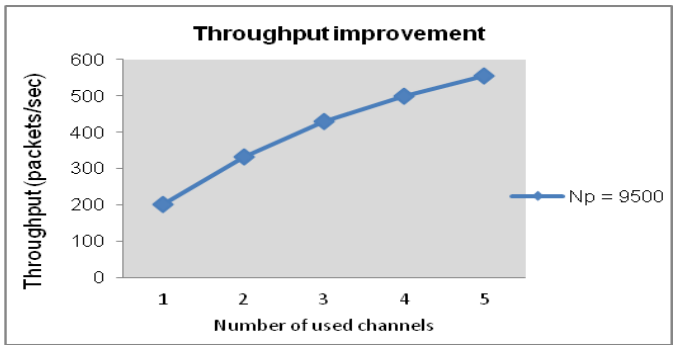


Fig. 9 Chain throughput improvement for f=3 and m=6.

Figure (9) presents the improvement of the throughput of a single chain of nodes utilising variable number of channels while f equals to 4 nodes and there are 6 nodes in the chain used for transmission. By adding extra channels the capacity of the chain is increased following the same rate as the transmission time. For N_p = 9500, there is an increase of 255 packets/sec when 5 channels are utilised within the chain. The same trend is followed for N_p = 6000 and N_p = 13000.

The next figure, figure (10) shows the improvement to the transmission time as f is further reduced to only two 2 nodes away. The reason behind this is the smaller amount of interference. If we deploy more than 3 channels within the same chain, the rate of improvement is reduced significantly and this indicates that any extra channels do not offer any great benefits.

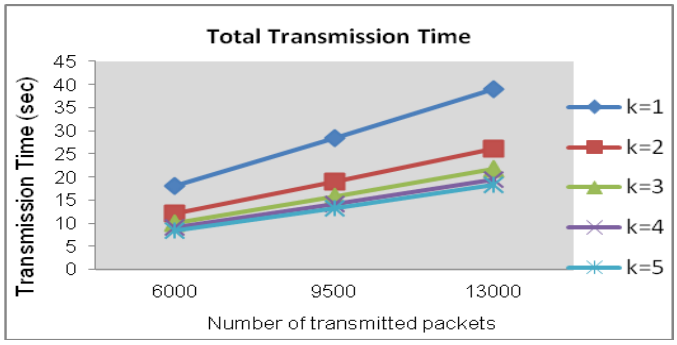


Fig. 10 Transmission time improvement for f=2 and m=6.

Throughput is further improved as it happened in the last two scenarios and modulo now achieves throughput of more than 710 packets/sec. This improvement is shown in figure (11) for N_p = 13000.

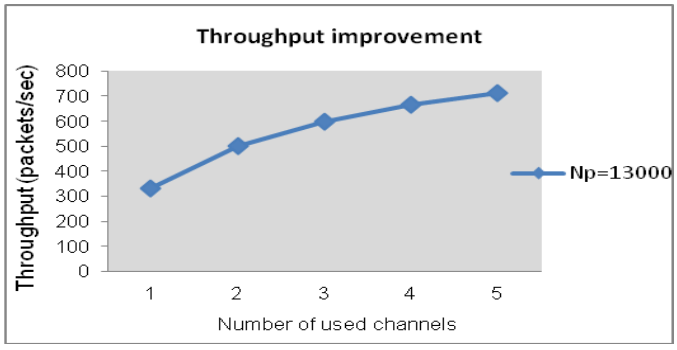


Fig. 11 Chain throughput improvement for f=2 and m=6.

6 Conclusion and Future Work

In this paper we evaluated the performance of a wireless network that is divided into smaller subnetworks and these utilize a variable number of frequency channels. The findings from the proposed theoretical approach show that when nodes are deployed in a chain topology, as it is performed in a segregated network, the use of extra channels for switching from hop to hop reduces the total transmission time for a number of packets N_p and consequently increases the throughput of the chain. When multiple chains are deployed using different channels then the improvement of the throughput is multiple. Apart from the utilised channels, the reduction in

the transmission range of the nodes improves significantly the chain's throughput. This reduction of the transmission range has a double positive impact, as less power is required for transmission and more packets can travel through the chain by using less energy.

Future work plans include the intention to move away from the legacy IEEE802.11 standards such as 802.11b and 8002.11g and start examining the efficient spectrum use of the new IEEE standards such as 802.11n [21] and 802.11ac [22]. When 802.11b/g were introduced there were no plans for any MIMO support by utilising multiple channels within the same network. The future of wireless communications is heavily depending on more competent spectrum management and utilisation of existing available frequencies.

8 References

[1] IEEE Standard for Wireless LAN-Medium Access Control and Physical Layer Specification, P802.11, 1999

[2] Tropos Networks, http://www.tropos.com

[3] Pradeep Kyasanur and Nitin H. Vaidya, Technical Report, October 2004.

[4] P. Bahl, R. Chandra, J. Dunagan, "SSCH: Slotted seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad-Hoc Wireless Networks", ACM Mobicom, 2004.

[5] A. Raniwala, T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-Based Multi-Channel Wireless Mesh Network", in IEEE Infocom, 2005.

[6] J. So, N. Vaidya, ".Multi-Channel MAC for Ad Hoc Networks: Handling Multi-Channel Hidden Terminals Using A Single Transceiver", in MobiHOC 2004.

[7] Y. Liu, E. Knightly; "Opportunistic Fair Scheduling over Multiple Wireless Channels", in IEEE INFOCOM '03.

[8] R. Karrer, A. Sabharwal, E. Knightly, "Enabling Large-scale Wireless Broadband: The Case for TAPs", in HotNets '03

[9] P. Hsiao, A. Hwang, H. Kung, D. Vlah, ".Load-Balancing Routing for Wireless Access Networks", in IEEE INFOCOM '01.

[10] BelAir Networks, http://www.belairnetworks.com

[11] S. Chhabra ,A. Zaghloul,O. Kilic, Co-channel Interference in satellite-based cellular communication Systems, The International Union of Radio Science 2005, New Delhi, India.

[12] J. Potman, F. Hoeksema, C. Slump, (2006) Adjacent Channel Interference in UMTS Networks. In: ProRISC 2006, 17th Workshop on Circuits, Systems and Signal Processing, 23-24 November 2006, Veldhoven, The Netherlands.

[13] A. Paraskelidis, M. Adda, "An evaluation of uniform multi-channel network versus multi-channel segregate network", in IADAT-tcn2006, September 2006.

[14] C. Perkins, E. Belding-Royer, S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing", in Ietf RFC 3561, July 2003.

[15] Owen, G., Adda, M., "Modulo and Grid Based Channel Selection in Ad Hoc Networks", in IADAT-tcn, September 2006.

[16] M. Adda, G. H. Owen, M. Kassasbeh, A. Paraskelidis, and A. Peart, "Communication issues in Large Scale Wireless Ad hoc Networks," in International Conference on Computer Science and Information Systems, Athens, Greece, 2005.

[17] UCLA Parallel Computing Laboratory, "Global Mobile Information Systems Simulation Library", Available: http://pcl.cs.ucla.edu/projects/glomosim

[18] A. Paraskelidis, M. Adda, "The Performance of a Segregate WLAN inside a Noisy Industrial Environment", in International Conference on Computer Science and Information Systems, ATINER, Athens, Greece, 2008.

[19] A. Paraskelidis, M. Adda, "Achieving 802.11 Wireless Networks Deployment in Noisy Environments", International Conference in Wireless Networks, ICWN '09, Las Vegas, USA, 2009.

[20] A. Paraskelidis, M. Adda, "Efficiency and Benefits of Wireless Network Segregation", in ICNS '09, Valencia Spain, 2009

[21] IEEE 802.11n-2009—Amendment 5: Enhancements for Higher Throughput, October, 2009.

[22] IEEE 80211ac, "IEEE P082.11ac/D1.1", Institute of Electronic Engineers, August, 2011.

Development of Outage Tolerant FSM Model for Fading Channels

Ms. Anjana Jain[1] **P. D. Vyavahare[1]** **L. D. Arya[2]**
[1]Department of Electronics and Telecomm. Engg. , Shri G. S. Institute of Technology and Science, Indore, M.P., India
[2]Department of Electrical Engineering, Shri G. S. Institute of Technology and Science, Indore, M.P., India

Abstract *- Finite State Markov (FSM) models for fading channels need to be revised for more realistic design of emerging mobile networks and their performance evaluation. In this paper a Outage Tolerant FSM Model (OTFSM) is proposed based on concept of certain tolerable outage times, which are defined as 'Tolerance time'. These are the short duration of outage time which is considered as satisfactory times over the channel. In this paper, a statistical approach is being presented for the development of OTFSM model and evaluation of its fading parameters such as Average Fade Duration (AFD), outage probability and outage frequency. Derived results may be used for higher layer performance evaluation and selection of physical layer parameters of wireless networks.*

Keywords: *FSM model, Tolerable outage time, Outage probability, Fading channel, Fade Duration Distribution*

1 Introduction

Various approaches for characterization of fading radio channels as Finite State Markov (FSM) model have appeared in the literature over last five decades. Initial channel models assumed fading as a Gaussian process and the resultant envelope with Gaussian probability density function (pdf). Mobile wireless channels suffer from multipath propagation and therefore the received signal envelope is approximated according to certain pdf's like Rayleigh, Rican, and Nakagami [1, 2]. For reliable characterization of the fading channels,various approaches were presented to model the channel.

Started with the pioneer work of Gilbert and Elliot [3, 4] for two state model of fading channel, the FSM channel models are classified as finite 'N' state and variable state Markov chain model. [5]. Wang and Mayori [6] proposed FSMC model with more than two states based on Signal to Noise Ratio (SNR) partitioning for Rayleigh channel. Binary Symmetric channel (BSC) is associated with each state and transitions with Markov property are assumed between states. It is pointed out that fading speed of the channel decides the SNR and its partitions to specify the states. Deterministic channel modeling and long range prediction of fast fading mobile radio channels has also appeared in [7]. The relationship between a physical channel and its FSM model for bit and packet error probabilities are demonstrated in [8, 9, 10, 11]. Babich demonstrated a technique to improve FSM model description based on Context Tree Pruning (CTP) algorithm [12]. Tan and Beaulieu proposed fading simulation via filtering methods and quantization of resulting process [13]. Further First order Markov chain analysis for short and long time duration of time is examined for the Rayleigh Channel by Chockolingam and Milstein [14]. Bai [15] presented various error modeling schemes for fading channels. Stochastic channel models were used to compute the fading parameters such as Level Crossing Rate (LCR) and Average Fade Duration (AFD) for fading channels [16, 17]. The Fade Duration Distribution (FDD) function and Minimum Duration Outage (MDO) in Weibull fading channel is described in [18]. Concept of repair time is elaborated in [19]. Reig and Rubio have shown the modeling of Fade Depth and the Fade Margin in UWB Channels applicable to emerging mobile networks [20].

It is revealed from the literature survey that a major contribution to channel characterization studies goes to FSM channel modeling and evaluation of traditional fading statistical parameters such as AFD, outage probability and level crossing rate. These parameters give the insight of mean behavior of the wireless systems. However in recent years mobile systems are emerging as 3G and 4G networks. For such systems the different frequency components contained in the transmitted bandwidth experience different propagation environments and operate at high data rate, ultra high frequency and in a hostile channel environment. Therefore mean behavior is not sufficient and more precise characterization of fading process is needed. Existing FSM models are therefore needed to be reformed with evaluation of additional channel characterization parameters. Apart from the traditional fading parameters outage time distribution is also identified as one of the important parameter for investigation.

This paper presents the methodology to develop Outage Tolerant FSM (OTFSM) channel model based on the

concept of tolerable outage time. Second-order fading parameters such as AFD (Average duration of outage time), outage probability (probability of being in outage) and outage frequency (number of times received signal crosses the threshold level and experience outage) are derived for proposed model. The proposed model considers the discrete value of the observation time of the channel. The concept is applied to Rayleigh and Nakagami channel with the variation in parameters like Doppler spread and fade margin. Paper is organized as follows. Section II presents formulation of tolerance time. Development of OTFSM model is discussed in section III. Fading parameters of the OTFSM model are derived in section IV with results and discussions in section V. Paper is concluded in section VI.

2 Formulation of tolerance time

3G and 4G wireless communication systems are supposed to offer variety of services such as voice, data, image transmission and internet browsing at high data rate in a relatively severe fading environment. The paper proposes a new FSM channel model as Outage Tolerant FSM (OTFSM) model while introducing 'tolerance time' as a new statistical fading parameter. In the proposed model the fading channel is described by discrete time and discrete amplitude Markov chain, *X (n)* with sample space

$$S_1 = \{1 \ldots\ldots k\} \quad (1)$$

$$S_2 = \{k + 1 \ldots\ldots N\} \quad (2)$$

Subset S_1 consists of outage states and subset S_2 consists of satisfactory states. $X(n)$ can be represented as sampled PSD of the channel with the sampling period denoted as slot duration Δt. A threshold value of received signal PSD is selected to decide the state of Markov process. When received signal during slot duration is more than the specified threshold, state of the process belongs to subset S_2, otherwise in subset S_1. The finite number of slot durations which are in outage and are tolerated by the channel with acceptable with acceptable Packet Error Rate (PER) target, is defined as the tolerable error event length and the time duration is denoted as tolerance time. In OTFSM model discrete time approach is proposed for the tolerance time and outage is considered when the received signal strength stays below the threshold value longer than tolerance time. Based on the concept, OTFSM model considers the states which are in tolerable outages as satisfactory states rather than outage states. OTFSM channel model is developed in this section with its state diagram, steady state probabilities of being in satisfactory, outage and tolerant states and fading parameters.

It is proposed that channel may tolerate maximum duration of time denoted as t_{tol} which is equal discrete number of time slot τ. Fading parameters for OTFSM channel model are denoted as

p_{out}'(Outage probability), $\overline{T_{out}'}$ (Average outage time) and f_{out}' (Frequency of outage). The choice of tolerance time depends on the past outage statistics, error correction time and acceptable PER. Tolerance time may be considered as a non-negative continuous random variable with certain distribution functions like Beta or Exponential [18]. The range of tolerance time τ is selected from τ_1 to τ_2 with $\tau_1 < \tau < \tau_2$ and tolerance time is considered as continuous variable. If

$\tau = 0$, The OTFSM model is the same as the original model

$\tau = \infty$, The OTFSM model will never be in fading state.

3 Development of outage tolerant FSM Channel model

In the proposed that in the OTFSM model short tolerable outages are considered as tolerant state along with satisfactory and outage state. States *1* to *N- k* are considered as outage states. States *N- k+ 1* to *N- k + m* is considered as satisfactory states and state *N – k + m +1* to *N* are considered as tolerant states. Figure one shows the plot for channel state versus time for OTFSM model. The model can be represented by three states.

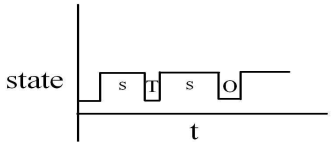


Figure 1- The timing diagram of proposed OTFSM model

O - Outage state, represented by subset S1
S - Satisfactory state, represented by subset S2
T- Tolerant state, represented by subset 3

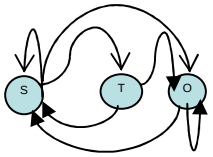


Figure 2- State diagram of OTFSM model with satisfactory outage and tolerate state

State diagram shows the probability of transition from satisfactory state to tolerant state p_{st} and from tolerant state

to satisfactory state is p_{ts} is assumed to be nonzero. Probability of transition from satisfactory state to outage state p_{so} is nonzero however probability of transition from an outage to tolerant state p_{ot} is assumed to be zero. Channel is modeled as OTFSM where during tolerance time channel is supposed to be nonfading state as described above and made available to user. Model is based on following assumptions -

i. This is a Markov model of fading channel subject to multiple outages.
ii. The duration between the satisfactory instant and occurrence of an outage event is the satisfactory time and the time for the channel being in outage is random variable.
iii. Figure two is the state diagram of this model.
iv. Tolerant and outage state cannot occur at the same time.

The state space of the fading process of OTFSM model is given as

$$S_1 = \{1, \ldots\ldots N-k\} \quad (3)$$
$$S_2 = \{N-k+1, \ldots\ldots \{N-k+m\} \quad (4)$$
$$S_3 = \{N-k+m+1, \text{-----}N\} \quad (5)$$

The one step transition probability matrix of the model is defined as

$$P = \begin{vmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{vmatrix} \quad (6)$$

For the proposed model p_{13} and p_{33} are assumed to be zero. The one step probability matrix for the OTFSM model can be represented as follows –

$$P = \begin{vmatrix} p_{11} & p_{12} & 0 \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & 0 \end{vmatrix} \quad (7)$$

Error probability matrix, P_B , in which all entries of being in satisfactory state is set to zero, is given as

$$P_B = \begin{vmatrix} p_{11} & 0 & 0 \\ p_{21} & 0 & 0 \\ p_{31} & 0 & 0 \end{vmatrix} \quad (8)$$

Steady state probability can be shown as

$$\Pi_0 = \Pi_1 = \Pi_1\ p_{11} + \Pi_2\ p_{21} + \Pi_3 p_{31} \quad (9)$$
$$\Pi_2 = \Pi_1\ p_{12} + \Pi_2\ p_{22} + \Pi_3 p_{32} \quad (10)$$
$$\Pi_3 = \Pi_1\ p_{13} + \Pi_2\ p_{23} + \Pi_3 p_{33} \quad (11)$$
$$\Pi_1 + \Pi_2 + \Pi_3 = 1 \quad (12)$$

As p_{13} and p_{33} are assumed to be zero, solving the equations for steady state probabilities

$$\Pi_1 = \frac{(p_{21} + p_{31} p_{23})}{(p_{12} + p_{21} + (p_{12} + p_{31}) p_{23}} \quad (13)$$
$$\Pi_2 = 1 - \frac{\Pi_1}{1 + p_{23}} \quad (14)$$
$$\Pi_3 = \frac{p_{23} p_{12}}{(p_{12} + p_{21} + (p_{12} + p_{31}) p_{23}} \quad (15)$$

For existing model tolerant states are considered as outage states, therefore

$$p_{23} = p_{21} \quad (16)$$
$$p_{31} = p_{11} \quad (17)$$
$$p_{11} + p_{12} = 1 \quad (18)$$

Substituting above equations in equation (9) and equation (10) , steady state probabilities would be same as two state model, shown as below

$$\Pi_1 = \frac{p_{21}}{p_{12} + p_{21}} \quad (19)$$
$$\Pi_2 = \frac{p_{12}}{p_{12} + p_{21}} \quad (20)$$

4 Estimation of fading parameters

For OTFSM model outage is considered when channel is in satisfactory state at time zero followed by $n\ \Delta t$ slots of outage state, equal to tolerance time, t_{tol}. AFD can be rewritten as below [17] –

$$\overline{T'_{out}} = t_{tol} + P_B (I - P_B)^{-1} \quad (21)$$

$$P_B = \begin{vmatrix} p_{11} & 0 & 0 \\ p_{21} & 0 & 0 \\ p_{31} & 0 & 0 \end{vmatrix} \quad (22)$$

$$(I - P_B) = \begin{vmatrix} 1 - p_{11} & 0 & 0 \\ 1 - p_{21} & 0 & 0 \\ 1 - p_{31} & 0 & 0 \end{vmatrix} \quad (23)$$

$$(I - P_B)^{-1} = \begin{vmatrix} 1 & 0 & 0 \\ -p_{21} & 1-p_{11} & 0 \\ p_{31} & 0 & 1-p_{11} \end{vmatrix} \quad (24)$$

$$P_B (I - P_B)^{-1} = \left(\begin{vmatrix} p_{11} & 0 & 0 \\ p_{21} & 0 & 0 \\ p_{31} & 0 & 0 \end{vmatrix} \right) \left(\begin{vmatrix} 1 & 0 & 0 \\ -p_{21} & 1-p_{11} & 0 \\ p_{31} & 0 & 1-p_{11} \end{vmatrix} \right) \quad (25)$$

$$= \frac{p_{11}}{1 - p_{11}} \quad (26)$$

Substituting in equation (21), AFD of OTFSM would be

$$\overline{T'_{out}} = t_{tol} + \frac{p_{11}}{1 - p_{11}} \quad (27)$$

Frequency of outage f'_{out} is given as [16]

$$= \Pi_3 (P_B^{k-1} P)\ e_B \quad (28)$$

For OTFSM model $\Pi_3 = (\Pi_3 + \Pi_1)$ and $e_B = p_{21}\ p_{31}$ are considered and hence equation (28) can be rewritten as

$$= (\Pi_3 + \Pi_1)\ p_{11}^{k-1}\ (p_{21}\ p_{31}) \quad (29)$$

p'_{out} can be given as

= Frequency of outage. AFD (30)

Result are in agreement with the earler results for two stage FSM model [16]. Probabilities of being in any of the states

can be computed either using simulation using MATLAB or mathematically for different values of Doppler shift and fade margin. The work can be extended for continuous and arbitrary distributed tolerance time [19].

5 Results and discussions

Observations were made for estimating AFD of OTFSM model. Further AFD of the proposed model is evaluated for the different values of one step transition probability. It has been demonstrated that AFD is the continuous increasing function of tolerance time. With tolerance time AFD decreases almost by ten percent as the value of ρ changes from one to two, as demonstrated in figure 3. Similarly, figure 4 demonstrates that the fading rate has an large impact on AFD. When fading rate is low i.e. low value of Doppler spread results in large value of AFD. The channel is likely to stay for larger time in the state once it crosses the threshold. Twenty percent fall in AFD results with the decrease in the Doppler spread from 120 Hz to 60 Hz.

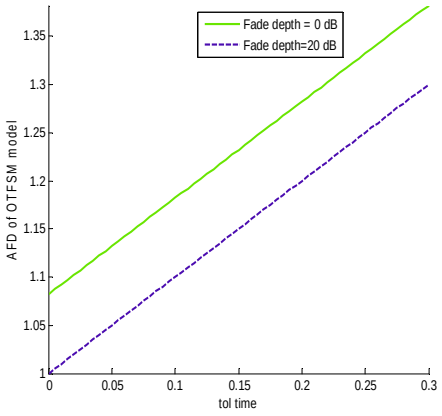


Figure 3: Variation in AFD of OTFSM with Tolerance time with various fade depth

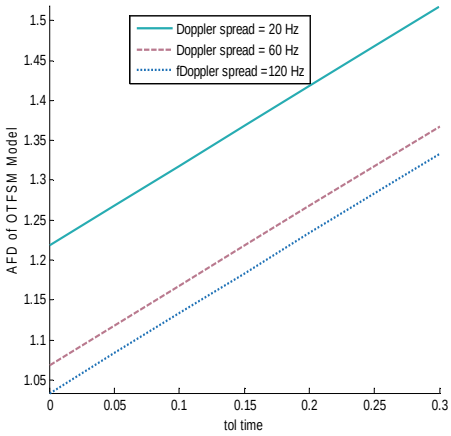


Figure 4: Variation in AFD of OTFSM with tolerance time for various Doppler spread in Hz.

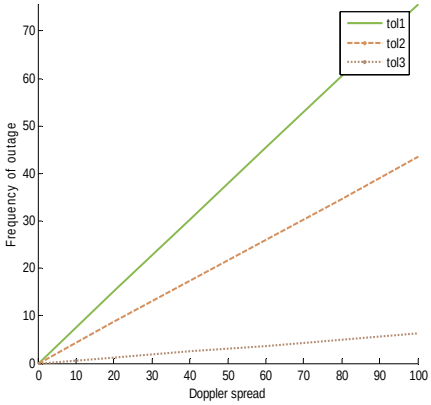


Figure 5: Variation in frequency of outage with Doppler frequency with tol1 < tol2 < tol3

Figure 5 shows the variations in frequency of outage with Doppler frequency with tol1 < tol2 < tol3. There is a slow change in frequency of outage with large value of tolerance time.

6 Conclusion

The paper presents the development of Outage Tolerant Finite State Markov model, which is based on the concept of tolerable outage time. Various fading parameters of the channel such as AFD, outage probability and outage frequency are evaluate and their behavior is studied using proposed OTFSM model . Simulation of the proposed model is carried out and the results are validated with the earlier literature. Channel can be made available to the user under the tolerable outage times and hence will result in higher spectral efficiency. Presented work may be extended for relaying fading channels and descrete sampled fading channels of emerging mobile networks [21]. Results may be used for higher layer Protocol performance evaluation and physical layer parameters selection of wireless networks.

7 References

[1] Rappport T. S., Wireless Communication, Pearson Education, Second Edition, 2002.

[2] Goldsmith A., Wireless Communication, Cambridge Press, 2005.

[3] Gilbert E., "Capacity of a burst-noise channel", Bell System Technical Journal, vol. 39, pp. 1253-1266, Sept. 1960.

[4] Elliott E. O., "Estimates of error rates for codes on burst-error channels", Bell System Technical Journal, vol. 42, pp. 1977-1997, September 1963.

[5] Aulin T., "A modified Markov Model for the fading signal at a mobile radio channel", IEEE Transactions on

Vehicular Technology, vol.28, no. 3, pp. 182-203, August 1979.
[6] Wang H.S. and Mouyeri N, "Finite State Markov Channel- A useful model for radio communication channels", IEEE Transactions of Vehicular Technology, vol. 44, no. 1, pp. 163-171, Feb. 1995.
[7] Eyceoz T., Hallen A., and Hallen H., "Deterministic channel modeling and long range prediction of fast fading mobile radio channels", IEEE Communication Letter, vol. 2, no. 9, pp. 254–256, Sept. 1998.
[8] Wang H.S. and Chang P.C., "On Verifying the First Order Markovian Assumption for a Rayleigh Fading Channel Model", IEEE Transactions on Vehicular Technology, vol. 45, no. 2, pp. 353-357, May 1996.
[9] Ross S. M., Introduction to Probability Models, Elsevier Publication, IX edition, 2007.
[10] Michele Zorzi, "On the Statistics of Block Errors in Bursty Channel", IEEE Transactions on Communication, vol. 45, no.6, pp. 660-666, June1997.
[11] Zhang Q. and Kassam S. A., "Finite State Markov Model for Rayleigh Fading Channels", IEEE Transactions on Communication, vol. 47, no. 11, pp. 1688-1692, Nov. 1999.
[12] Babich F., Kelly O. E, Lombardi G, "Generalized Markov Modeling for Flat Fading", IEEE Transactions on Communication, vol. 48, no.4, pp. 547-551, April 2000.
[13] Tan C.C and Beaulieu N.C, "On First Order Markov Modeling for Rayleigh Fading Channel", IEEE Transactions on Communication, vol. 48, no. 12, pp. 2032-2040, Dec. 2000.
[14] Ramesh A. Chockolingam, and Laurence B. Milstein, "SNR Estimation in Macadam-m Fading With Diversity Combining and Its Application to Turbo Decoding", IEEE Transactions on Communication, vol. 50, no. 11, pp. 1719-1724, Nov. 2002.
[15] Bai H and Atiquzzaman M., "Error modeling schemes for fading channel in Wireless Communication – A Survey", IEEE Electronics Magazine on Communication, vol. 5, no.2, pp. 2-8, Fourth Quarter 2003.
[16] Zorzi M., "Minimum duration outage in Markov channel", IEEE Transactions on Communication, vol. 54, no. 6, pp. 2102 - 09, 1998.
[17] Beaulieu N. C. and Dong X., "Level Crossing Rate & Average Fade Duration of MRC (maximum ratio combiner) and EGC (Equal gain combiner) Diversity in Rician Fading", IEEE Transactions on Communication, vol. 51, no. 5 , pp. 722-726, May 2003.
[18] Wang Miao Liu et al., "Fade Duration Distribution and Minimum Duration Outage in Weibull Fading Channels", Vehicular Technology Conference, 2011.
[19] Zheng Z. Cui L., and Hawakes A. G., "A Study on a Single-Unit Markov Repairable System with Repair Time Omission", IEEE Transactions on Reliability, vol. 55, no.2, pp. 182 – 188, June 2006.
[20] Lano G. Reig and J. Rubio L.,"Analytical Approach to Model the Fade Depth and the Fade Margin in UWB Channels", IEEE Transactions on Vehicular Technology, vol. 59, no. 9, pp. 4214 – 4221, 2010.
[21] Javier L Martinez F. et al.,"Higher order statistics of sampled fading channels with applications", IEEE Transactions of Vehicular Technology, vol. 61, no. 7, pp. 3342-45, Sept. 2012.

SESSION

WIRELESS NETWORKS AND ENERGY EFFICIENCY ISSUES

Chair(s)

TBA

An intelligent approach for improving energy efficiently in smart grids

Geraldo P. R. Filho[1], Jó Ueyama[1], Leandro A. Villas[2], A. R. Pinto[3], Vinícius P. Gonçalves[1] and Sibelius Seraphini[1]
[1]Institute of Mathematics and Computer Science, University of Sao Paulo - Brazil
[2]Institute of Computing, University of Campinas - Brazil
[3]Sao Paulo State University, Sao Jose do Rio Preto - Brazil
{geraldop, joueyama, vpg}@icmc.usp.br, leandro@ic.unicamp.br,
arpinto@ibilce.unesp.br, sibelius@grad.icmc.usp.br

Abstract— *This article proposes a platform that integrates wireless sensor networks and cloud computing for remote monitoring of electric energy consumption by means of any device (e.g., smartphones, tablets and notebooks) that has access to the Internet. The solution proposed is different from other available solutions in at least three respects: it employs an intelligent method for monitoring electric energy consumption; it uses Machine Learning techniques to analyze the behavior of electronic equipment; and it sends intelligent alerts to the device when an anomaly arises. The results of the experiments showed the efficiency of the method in detecting novelties in electronic equipment which has made their use viable in our platform.*

Keywords: Smart grid, intelligence, novelties and energy.

1. Introduction

In recent years, there has been a growing demand for electricity on the part of industry, businesses and residential dwellings. This has been the situation in both Brazil and throughout the world because the per capita consumption of electricity in Brazil and the rest of the world increased by 20% and 22% respectively in the period between 1999 and 2009 [4]. These increases in demand and supply require a more intelligent electric system that can allow the consumption of electricity to be reduced in every electronic appliance by encouraging consumers to adopt efficient strategies for the reduction of energy consumption in electrical equipment.

The information technology for electric power systems integrated with electronic communications network and infrastructure, known as smart grid [6], allows the electric power system to be monitored and managed at any place and any time. As a result, the use of smart grids has become increasingly important in the urban scenario since it provides integration for various sources of energy such as: hydroelectric, solar and atomic energy and wind power.

It is expected that smart grid will become a reality in the next few years since industry, universities and governments throughout the world are devoting financial resources just to the development of smart grids. This can be confirmed through different national and international schemes and measures by governments, industry and the academic world that are devoted to intelligent networks [6], [7] and [9].

Besides the advances in smart grid techniques, electrical companies do not offer support for final consumers (for example, the remote management of energy consumption). Economical cost is the main reason that prevents companies to offer these services. Therefore, total consumption is the only metric available for final consumers. In this way, the detection of anomaly energy consumption in each electrical equipment is not trivial.

Thus, the integration of Wireless Sensor Networks (WSN) in a smart grid is an interesting and cheap alternative to deal with this issue [2], [5], [11] and [16]. One of the main advantages of a WSN is its capabilities, to timely monitoring each electronic equipment. Therefore, consumers could be able to detect their electrical consumptions patterns. Based on the consumption patterns that were detected, consumers can change their habits arming a decrease in energy consumption.

This study proposes an architecture that integrates WSN and cloud computing for remote monitoring of energy consumption by means of any device that has access to the Internet. Our system sends data consumption of the electronic equipments in a distributed way of accordance with each sensor (see Figure 1), however the novelty detection for each electronic equipment is individual. Some of the solutions in the literature explore the question of monitoring energy consumption [2], [5], [7], [11], [12] and [16]. However, these solutions do not classify and detect novelties in electronic equipment. Our solution makes use of an intelligent method to monitor the consumption of electricity that is incorporated in the platform, where Machine Learning (ML) techniques are employed to analyze the behavior of the electronic equipment and in this way, send intelligent alerts to the mobile device (e.g., smartphone) when an anomaly arises.

The modeling used to detect novelties in the consumption of electrical energy consumption employs Instance Based Learning (IBL) concepts, Markov chains and entropy. The aim is to carry out a behavioral study of the electronic equip-

ment by defining reduction and/or re-education strategies for energy consumption. This is because the feedback on the monitoring of energy consumption, (both quantitative and qualitative), is formed in an efficient instrument that can allow assessment for decision-making.

The results of the experiments show that the proposed method for this study achieves good results, achieving satisfactory performance in the novelty detection in a decentralized environment via WSN (see Section 4). Thus, the integration of the method with our platform has made it possible for the users to receive alerts when an anomaly arises in their equipment.

Our solution differs from those that already exist in at least three respects: it makes use of an intelligent method for monitoring electric energy consumption; it employs ML techniques to analyze the behavior of the electronic equipment and it sends alerts in an intelligent way to the mobile device when some anomaly arises.

The article is structured in the following way: Section 2 outlines a general review of the approaches that can be used to explore the monitoring of energy consumption; the approach used to provide intelligence in a smart grid as shown in Section 3, the results are analyzed in Section 4 and finally Section 5 describes the conclusions and makes suggestions for future work.

2. Related Work

Several studies have been published in the area of smart grid in recent years and the aim of this section is to provide a review of scientific work [2], [5], [7], [11] and [16] where the focal point is largely on the monitoring of electric energy consumption. However, in spite of all the advances in this field, there still remain a number of challenges and problems in the area. For example, the lack of a methodology for novelty detection in a decentralized environment monitored via WSN.

One of the oldest smart grid models is the Telegestore Project [2]. Telegestore is a system that remotely manages residential and commercial meters with a view to exploring the low voltage distribution network between the transformers and the meters. The main disadvantages with regard to this study are: (i) the lack a method to detect anomalies (e.g., blackout) on the low voltage network; (ii) remote monitoring by sectors is not explored; and (iii) due to its centralized nature, the system is vulnerable to data overload.

Erol-Kantarci and Mouftah [7] propose the use of WSN to manage the energy of a residence through the smart grid. To achieve this, the researchers proposed a Coordination Appliance (ACCORD) to reduce the cost of electricity at peak periods. The disadvantages of the ACCORD are: (i) not avail the benefits of the WSNs (e.g., monitoring by sectors); (ii) The lack of a method for novelty detection in the monitored environment; and (iii) experiments are simulated.

In Brazil, there are initiatives with regard to smart grids; for example, the Monitoring Center for End-users (CMUF) [11]. The aim of this project is to help people manage their electric energy consumption; it is a low-cost system and easy to install. However the disadvantages of the CMUF are: (i) data collection is centralized; (ii) The system does not use a distributed WSN for data collection; and (iii) no method of novelty detection.

The study that most closely resembles to this paper is of the [5]. The authors propose a smart meter to measure electric energy in a residency. The meter consists of manage electric energy consumption by enabling people within a private residence or business to detect which equipment consumes most. Despite the similarity, the disadvantages of the smart meter are: (i) platform does not use a method to novelty detection in electronic equipment; and (ii) server cloud is not used to manage information.

3. Our Approach to Provide a Smarter Smart Grid

This section sets out a model that is used to provide a more intelligent smart grid. This intelligence is provided by means of the Machine Learning technical system which was implemented and integrated in the platform for monitoring electric energy consumption. The aim is to send intelligent alerts so that the users can make changes in their habitual use and in this way, be more aware of the way they use energy.

3.1 Discussion of the Techniques Used to Provide Intelligence

This section discusses the best practice techniques for ML such as IBL, the Markov chain and entropy. The purpose is to show how these techniques are used in smart grid to make it become more intelligent.

The IBL was employed to provide this autonomy [14], since it is an incremental ML technique that generalizes the information on the basis of examples in training sessions. This technique has a method known as K-Nearest Neighbors (KNN) in which the nearest K training instances are analyzed to determine the class of an unknown element.

In this way, the Markov chains [13] are coupled in the method to capture the dynamics of the behavior of the electronic equipment in accordance with the KNN classification. The Markov chain was used because the Markov decision processes can be used to represent the relationship between the different states of a process.

Since a set of Markov chains is generated as a result of the interaction of the user with the electronic equipment, it is possible to calculate the entropy variation [17] and thus measure the degree of disorder in a system. In carrying this out, the use of entropy was one of the essential techniques for detecting novelties in our platform and letting it become

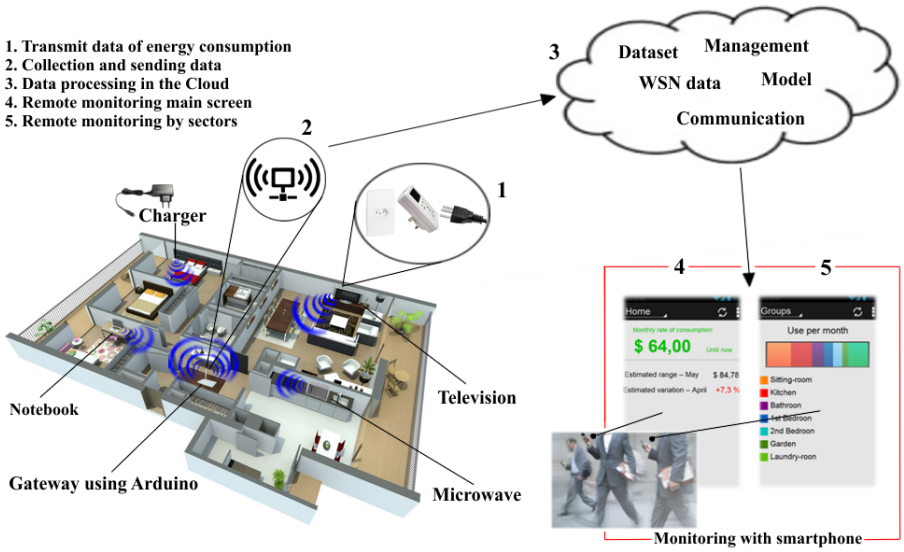


Fig. 1: Operating scenario of the plataform.

more intelligent. In this way, autonomous alerts of an intelligent kind, are sent to a smartphone when an entropy variation occurs.

Hereafter is described the operation of our platform constructed/implemented, Subsection 3.2, and the model used to provide more intelligence in our platform, Subsection 3.3.

3.2 Our Hardware and Software Prototyping

The proposed platform to carry out the monitoring of the electric energy consumption comprises three stages:

- Stage 1 is responsible for the acquisition of data on the consumption of energy by electronic equipment
- Stage 2 is related to an application that receives information and processes it.
- The final stage is responsible for making available the information about the consumption of electric energy by the interested parties

Figure 1 shows, in a general way, how the energy consumption remote monitoring platform functions. This was carried out by creating some WSN prototypes (see diagram 1 in Figure 1) on the basis of mounting Wattmeters (devices that measure electric energy consumption), and Kill-a-Watt marketed by P3[1] Company, which are connected to electronic equipment so that these are able to communicate with the server. As Wattmeters not have access to any means of communication was added to them a XBee[2] module, in order to transmit the info about energy consumption for a server. It was essential to construct an infrastructure for this (diagram 2, Figure 1) to carry out the monitoring via the WNSs and transmitting the data that was read for Cloud (diagram 3 in Figure 1). Thus makes it possible to carry out the remote monitoring through an application for the cell phone, smartphone, (diagram 4 and 5 in Figure 1).

Thus, the users receive the alerts and access to important information of energy consumption on the diagram 4 in Figure 1. It should be stressed that in the "monthly rate of consumption" field there are color variations with the color becoming more red as the electricity bill passes beyond the threshold. The screen on the diagram 5 in Figure 1, carries out the monitoring through the sectors or in other words, by means of the consumption measurements for areas such as the living-room, the kitchen, the bedroom, the bathroom and the whole surrounding area, only possible because of the WSN.

3.3 Collecting, Classifying and Detecting Novelties

It is necessary to go through three stages to carry out a behavioral study of the electronic equipment (the fridge, television, modem, washing-machine, among others) with the aim of detecting novelties in a smart grid:

- Data collection in the interactions between the users using equipment;
- Processing the data collected in an IBL to classify and obtain behavior equipment standards by means of Markov chains;
- Calculating the degree of uncertainty of the electronic equipment through the entropy variation of the Markov chain.

The stages described earlier will be rearranged hereafter to improve the presentation of this work.

[1]P3 International innovative electronic solutions, http://www.p3international.com

[2]Xbee-pro module datasheet, http://ftp1.digi.com/support/documentation/90000976G.pdf

3.3.1 Collecting Data

This stage addresses the question of preparing the database or in other words, the data collection relative to the problem, as well as its separation both for a set of training instances and for a set of unknown instances.

According to Morais & Mondaini [15] the data collection requires very careful analysis to tackle the problem in a significant way and reduce the risk of ambiguity and errors. This is because the collected data must not only deal with routine problems but also exceptions which broaden the range of the problem.

The energy consumption information is stored in a database, where real information about the devices is provided. This database has four attributes: (i) device id; (ii) power in watts; (iii) date of equipment utilization; (iv) and period of functioning time.

3.3.2 Classifying and Obtaining the Behavior

This stage outlines the method that has been chosen, KNN, to classify the data and incorporate them in the Markov chain. The aim is to obtain standards of behavior for the electronic equipment.

The choice of KNN was based on two main reasons: (i) KNN is a simple and efficient classification technique; (ii) and it can be used in several problems with a low overhead [18].

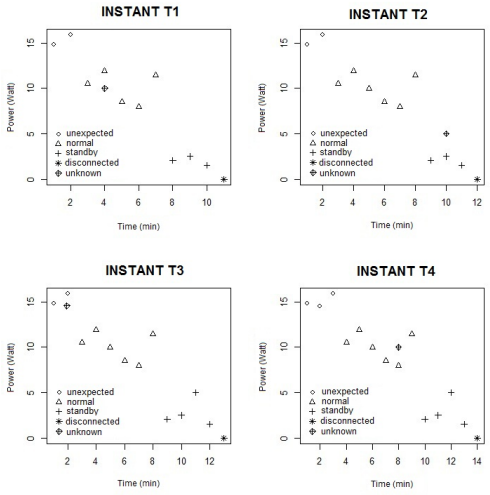


Fig. 2: Example of the KNN classification.

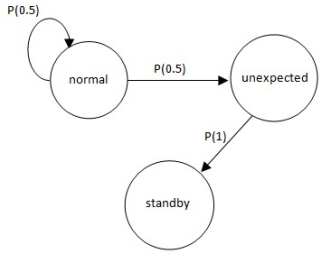
Figure 2 illustrates how the input data are classified to provide a better understanding of the method that is used for this study.

The graphs in Figure 2 represent the power of an electronic appliance (at each moment of time) as measured in time and it is possible to find four distinct kinds of behavior for energy consumption (unexpected, normal, standby and disconnected).

As the KNN classifies the input standards for unknown instances, a new Markov chain is generated to group these instances to a determined behavior. Thus there are distinct Markov chains for each time instant of the electronic equipment because at each time instant, the instants can have a different behavior.

from \ to	unexpected	normal	standby	disconnected
unexpected	0	0	0	0
normal	0	0.5	0.5	0
standby	1	0	0	0
disconnected	0	0	0	0

(a) Transtion matrix.



(b) Markov chain.

Fig. 3: Example of a transition matrix and Markov chain, in the T4 Instant of Figure 2.

In Figure 3, there is an example of a transition matrix (Figure 3a) and a Markov chain (Figure 3b) for a T4 INSTANT is shown in Figure 2. This is a transition matrix and a Markov chain which are updated at each time instant in accordance with the KNN classification. In addition, the Markov chain (Figure 3b) has arcs that represent a transition between the states (unexpected, normal and standby) with their respective probability, P(σ). The objective is to employ the probability, P(σ), to measure the degree of uncertainty in the system by detecting novelties/anomalies of an autonomous and intelligent kind.

3.3.3 Detecting Novelty with the Entropy Variation of Markov Chain

With the set of Markov chains generated, as previously explained, it is possible to calculate the entropy variation [17] through the following formula:

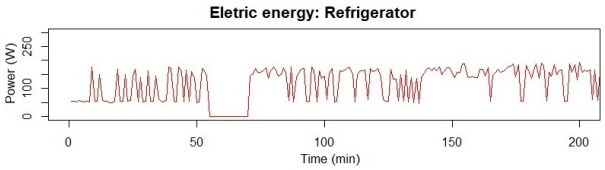
$$H(X) = -\sum_{i=1} p(x_i) \log_b p(x_i)$$

Where $\log_b$ denotes the logarithm of x at base 2 and the $p(x_i)$ is the probability of an event that will be for another state. Thus, each user interaction with the electronic equipment generates an energy curve, which represents behavioral changes of the equipment. In section 4 experiments are realized of the types of novelties found in the equipments.

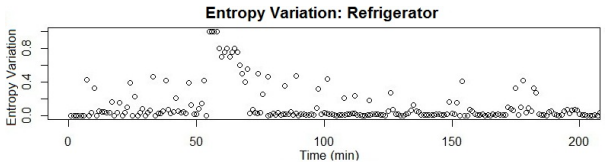
4. Experimental Results

The aim of this section is to conduct experiments and analyze results to validate the proposed method. This was carried out by mounting a real environment to monitor the consumption of the electronics equipment in a residence. In this scenario, the appliances were located in different rooms and for reasons of convenience, the Arduino gateway was joined to the router to communicate with the Internet. The data collection was carried out for the period of three months and during this time it was possible to observe the consumption profile of the user.

Figure 4 shows the experiment that was conducted of the electronic equipment (refrigerator). Figure 4 shows the entropy variation, where it is possible to notice the standard behavior (without novelty) between periods 0 to 54 and 71 to 200 minutes and the unexpected behavior (with novelty) between period 55 to 70 minutes.



(a) Electric energy consumption of the refrigerator during a period something unexpected happening between 55 to 70 minutes.

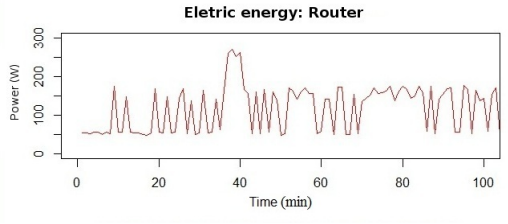


(b) Detecting novelty with an unexpected change in the behavior of the refrigerator according to the Graph 4a.

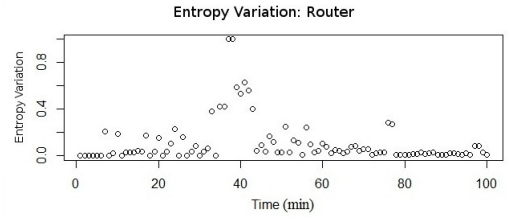
Fig. 4: Detection of novelties caused by changes in the behavior of the appliance.

As the Graph 4a show the power consumption with regard to time, it is essential to observe the behavior of the refrigerator at each period. Thus, in the time interval between 55 and 77 (Graph 4b), a sudden alteration in the entropy can be noted which is caused by a change in the standard behavior of the appliance (Graph 4a). This abrupt change came about due to the shut off of the refrigerator for 15 minutes (see Graph 4a in the between periods 55 to 70 minutes), or in other words, the equipment came out of its usual state which was not expected. In spite of this, the entropy variation begins to stabilize after 66 minutes. This stabilization can occur in two ways: the first is due to a change in the habits of the user (a question for feedback) and the second is that in the course of time, the method considers that the novelty is no longer something unexpected.

Another experiment that was conducted (Figure 5) is related to the high electric energy consumption of a device (a router) during a fixed period of time. The graph in Figure 5a is a good representation of this behavior and displays the power used by the router over a period of time. In this graph a high energy peak is noticeable in the period from 36 to 43. Thus the Graph 5b which represents the entropy variation in a time setting. This finds the occurrence of a novelty at the same time instant as the Graph 5a due to an abrupt change in the entropy. In this experiment the stabilization of entropy variation also occurs as described previously.



(a) Consumption of energy during a fixed period of time entropy variation: router.

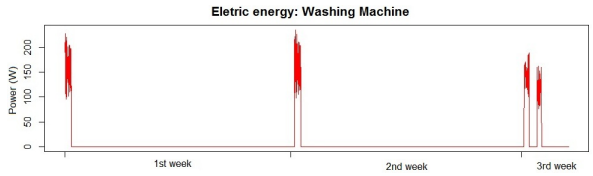


(b) Detecting novelty in eletric energy consumption.

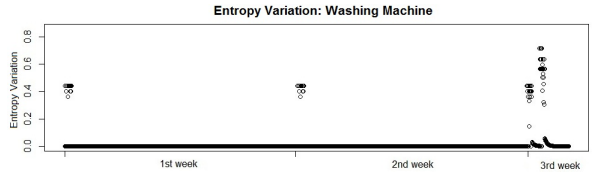
Fig. 5: Detection of novelties with an unexpected increase of energy consumption.

Figure 6 shows the eletric energy consumption of the washing machine. This appliance is not turned on constantly in the outlet, hence consumption data of this equipment are not behaved, unlike the refrigerator, router and/or freezer, which are constantly connected. Equipments that have similar characteristics to the washing machine, such as microwaves, coffee and/or televisions, despite having data no behaved it is possible to detect the novelties described previously. Graphs 6a and 6b illustrate this scenario. Note that at the beginning of 3rd week (Graph 6b) there was an increase in the entropy variation. This increase is explained by two reasons: (i) increased consumption of electricity energy in 3rd week, compared to the previous weeks; (ii) and changes in the habits of the user at the beginning of the 3rd week.

An analysis of the obtained data was also conducted with the purpose of validating the experiments. The aim is to estimate the power of classification and/or detection of the method employed for a correct state with regard to the behavior of the electronic equipment. This required statistical measurements such as : sensitivity/recall, precision, specificity and accuracy, which are used to estimate the performance of the proposed solution. The ultimate goal is to investigate the feasibility of employing entropy variation to detect novelties in mobile devices.



(a) Electrical energy consumption of the washing machine, occurring on the 3rd week a significant increase in energy consumption and/or a change in the habit of the user.



(b) Detecting novelty in the same time instant as the Graph 6a.

Fig. 6: Detecting novelty in electronic equipment, washing machine, which data not behaved.

According to Fawcett [8] these statistical measurements have inherent features since the sensitivity of the total number of samples is really positive (true positive); the precision is the the total number of examples classified as positive but which are not always so, or in other words, true negative; specificity is the opposite of sensitivity and in its negative classifications, only imports those that are de facto negative (false positive) and accuracy makes it possible to analyze how precise the method is when appropriately classifying the behavior of an appliance. These statistical measures are calculated from a confusion matrix shown in Figure 7, which evaluates the results based on the loss caused.

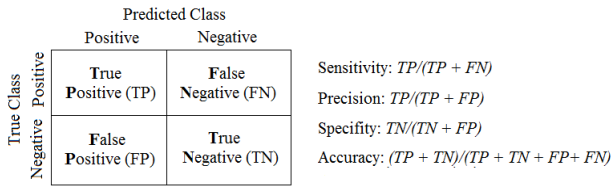
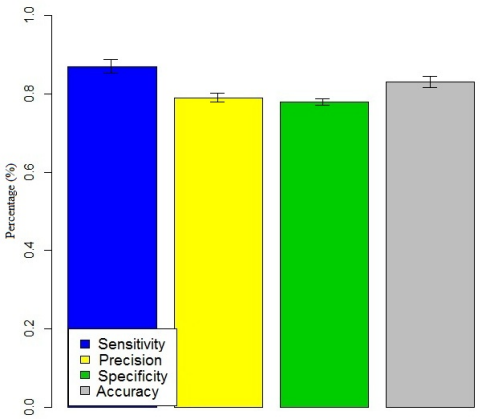


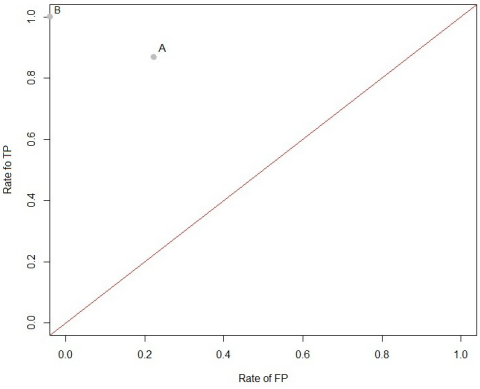
Fig. 7: Statistical measures calculated from the confusion matrix.

Were carried eight partitions of the samples of the dataset collection of three months. Each partition of the dataset was divided in 25% of training instances and 75% of the tests instances. This technique is known as hold-out, in which the database is divided in two sets (training and test set) [14]. For does not exist degree of dependence in the experiments, the division of training and test set was randomly replicated 8 times. The replications in the same subset have the aim to contemplate the various behaviors that users have (for example, different behavior at the end of the week compared to other days).

The Results shown in Figure 8a (using 95% of confidence interval) were very satisfactory since the sensitivity, precision, specificity and accuracy obtained 87%, 78%, 79%, and 83% respectively. In addition, Figure 8b displays a Receiver Operating Characteristic (ROC) graph where the axes y and x represent the rate of true positive and the rate of false positive respectively, allowing intuitive visualization of the efficiency of the method. Since Point A (Figure 8b) is located in the region that shows good results, it can be found at a point that approximates to a perfect classification known as ROC Heaven (Point B).



(a) Statistical measurements.



(b) ROC Graph.

Fig. 8: The results obtained through statistical measurements.

In this way, on the basis of the entropy variation it is possible to send alerts to the user of the equipment in a smartphone.

In the next section, there is a discussion of the experiments that were conducted

4.1 Discussion of the Results

In this section, there is a discussion of the results obtained with a view to investigating the influence of the method in our platform.

On the basis of the experiments carried out, two types of novelty (qualitative and quantitative) could be seen when the

entropy variation of the Markov chain was employed:

- The first, shown in Figure 4, occurs when there is an abrupt change in the standard behavior of the electronic equipment or rather there is a difference in its habitual state and it passes from State x to State y, which is unexpected. (qualitative).
- The second, shown in Figure 5, occurs when the electronic appliance begins to consume more energy than expected during a determined period. (quantitative).

Although they have two types of novelty, both are able to be stabilized in the error, as can be observed in figures 4b and 5b. From the user standpoint, this can be regarded as an advantage. The reason for this is that if the user does not want to change his behavior when faced with defective equipment or an excessive consumption of electricity, the method used will not cause inconvenience with regard to his alerts.

Another advantage concerns the messages sent to the users since the alerts are essential to allow people to make changes in their habits and even achieve a reduction in their electric energy consumption. The reason for this is that when they measure their consumption quantitatively, they can make intelligent choices with regard to their use of energy.

In the light of this analysis, it can be claimed that the method has a satisfactory performance and means that the solution for this study is viable for use in our platform. Moreover, the statistical measurements obtained results that were similar to those of the authors [1], [3] and [19] in the area of machine learning.

5. Conclusion and Future Work

In undertaking this study, a striking feature was the importance of using a classifier system to detect novelties in electronic equipment, in view of the growing demand for electricity on the part of industry, businesses and residential dwellings in recent years. accordingly, it is essential to explore intelligent ways of allowing a reduction in electric energy consumption in every electronic appliance and encourage consumers to adopt efficient measures.

Through an analysis of the results, it was possible to know the degree of efficiency of the method employed because the statistical measurements yielded good results in the experiments. Hence it has become feasible to apply the solution of the problem in our plataform so that those concerned can receive alerts in the mobile devices (in our application, with the Android platform) when something unexpected occurs in their equipment.

The main contributions of this paper are as follows: (i) We use a WSN for the construction of a prototype, where it monitors the energy consumption of electronic equipment individually; and (ii) Intelligent method (absolutely nonintrusive) based on the concept of ML to detect novelty in a monitored environment.

As a means of giving continuity to the work that has been carried out here, the implementation of a flexible interfaces for the end-users as done in [10] can be cited for future studies.

6. Acknowledgements

The authors of this paper would like to thank FAPESP for funding their research project (Process ID 2008/05346-4).

References

[1] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 1991.

[2] Brunello Botte, Vincenzo Cannatelli, and Sergio Rogai. The telegestore project in enelt's metering system. *International Conference on Electricity Distribution*, 18th, 2005.

[3] ANDREW P. BRADLEY. The use of the area under the roc curve in the evaluation of machine learning algorithms. 1997.

[4] PUBLIC DATA. Indicadores do desenvolvimento mundial. http://www.google.com.br/publicdata/, 2012. visitado em 02 de julho de 2012.

[5] Luís F. C. Duarte, José D. Zambiaco, Douglas Airoldi, Elnatan C. Ferreira, and José A. Siqueira Dias. Characterization and breakdown of the electricity bill using custom smart meters: a tool for energy-efficiency programs. *International journal of circuits, system and signal processing*, 2011.

[6] ENERGY. Smart grid | department of energy. http://energy.gov/oe/technology-development/smart-grid, 2012. visitado em 05 de março de 2012.

[7] Melike Erol-Kantarci and Hussein T. Mouftah. Wireless sensor networks for domestic energy management in smart grids. *Biennial Symposium on Communications*, 2010.

[8] Tom Fawcett. An introduction to roc analysis. *Elsevier*, 2006.

[9] Geraldo P. R. Filho, Jó Ueyama, Leandro Villas, Alex Pinto, and Sibelius Seraphini. Nodepm: Um sistema de monitoramento remoto do consumo de energia elétrica via redes de sensores sem fio. In Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), editor, *Sociedade Brasileira de Computação (SBC)*, volume 31, pages 17–30, may 2013.

[10] V. P. Gonçalves, S. Seraphini, V. P. A. Neris, and J. Ueyama. Flexinterface: a framework to provide flexible mobile phone user interfaces. *15th International Conference on Enterprise Information Systems (ICEIS 2013)*, 2013. In Proc. ICEIS 2013. No prelo (2013).

[11] Fábio Gonçalves Jota, Patricia Romeiro Silva Jota, and Eduardo Carvalhaes Nobre. Gerenciamento efetivo de energia por uso final: Um sistema de monitoramento de baixo custo via internet. *Seminário Nacional de Distribuição de Energia Elétrica*, XVII, 2006.

[12] Younghun Kim, Thomas Schmid, Zainul M. Charbiwala, Jonathan Friedman, and Mani B. Srivastava. Nawms: Nonintrusive autonomous water monitoring system, 2008.

[13] Andrei Andreyevich Markov. *Extension of the Limit Theorems of Probability Theorty to a Sum of Variables Connected in a Chain*. In Dynamic Probabilistic System, 1971.

[14] Tom Michael Mitchell. *Machine Learning*. McGraw-Hill Science, 1997.

[15] Emerson Cordeiro Morais. *Reconhecimento de padrões e redes neurais artificiais em predição de estruturas secundárias de proteínas.* PhD thesis, Universidade Federal do Rio de Janeiro, 2010.

[16] Power-Meter. Google powermeter: A google.org project. http://www.google.com/powermeter, 2012. visitado em 14 de Março de 2012.

[17] Claude Shannon. *A Mathematical Theory of Communication*. The Bell System Technical Journal, 1948.

[18] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

[19] Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, 2005.

An Energy Efficient and Minimum Latency Routing Protocol for Multihop WSNs

Changyan Yi and Ken Ferens
Electrical and Computer Engineering,
University of Manitoba, Winnipeg, MB, Canada
yic3@cc.umanitoba.ca, Ken.Ferens@ad.umanitoba.ca

Abstract— This paper presents a novel multihop routing protocol, which aims to simultaneously minimize data aggregation latency and maximize network lifetime. The protocol constructs a dynamic greedy growing tree (GGT) from sink to all sensor nodes. Latency and energy consumption minimization priority rules are applied to node selection at each step in the tree construction process. Latency minimization rules are applied first, and then energy consumption priority rules are applied to break any ties. The tree is constructed periodically to balance the energy consumption of nodes across the network. Priority Rules consist of the number of interfering neighbors, residual energy of senders and receivers, link quality, and load balancing. MATLAB simulations show that the proposed Energy Efficient Greedy Growing Tree (EEGGT) has the same latency performance as basic GGT. However, EEGGT significantly outperforms basic GGT, Static GGT (SGGT) and Dynamic GGT (DGGT) in terms of network lifetime.

Keywords— minimum delay routing, maximizing network lifetime, minimum delay data aggregation scheduling.

I. INTRODUCTION

A common difficulty encountered in the design of multi-hop routing protocols for wireless sensor networks (WSNs) is to minimize the delay and to simultaneously maximize network lifetime. These two goals seem to oppose one another. A routing protocol which aims to minimize delay usually achieves that goal by trading off network lifetime, and vice versa. On the one hand, a delay oriented routing protocol will typically chose the same subset of nodes in a portion of a route for data relaying purposes, simply because they represent distribution points in a least delay path. A such, these nodes will lose their energy before other nodes, since they have more work to do in terms of relaying radio messages and data processing. Consequently, the network suffers premature death. On the other hand, a network lifetime oriented protocol will typically choose a longer delay path, suitably because the nodes along that path have more energy than others. But, this network lifetime resultant path will consist of more hops than the least delay path; this results in a longer delay route. As such, more energy will be used, simply because there are more hops along the path. More nodes will be forced to perform radio relaying and data processing tasks, which means a higher total energy consumption amount, consequently leading to premature network death. A common goal in multi-hop routing protocols for wireless sensor networks (WSNs) is to achieve a compromise between minimizing the delay and maximizing network lifetime.

An approach of achieving both goals simultaneously is to tie together the decision making rules of both goals with a common parameter that constrains the design of each. One such parameter is data aggregation. Data aggregation can be used to significantly reduce the amount of consumed energy in a network, since, in an aggregation enabled network, the nodes along a routing path work together to reduce the representation of sensor readings by performing averaging, min/max, data compression, data merging, etc. This significantly reduces the size of radio transmissions, albeit, at the cost of increased data processing, but, it is well known that the cost of radio transmission far exceeds that of data processing. By decreasing the amount of radio transmissions, this increases the network lifetime.

A bridge connecting data aggregation and maximizing network lifetime with minimizing delay is to choose a minimum delay path such that the opportunity of parallel transmissions are maximized. When a protocol forms a minimum delay path, it should do so by choosing nodes in the path, which allow other nodes the opportunity to transmit simultaneously without interfering with the chosen nodes' transmissions. In this way, there is more opportunity to perform data aggregation, resulting in decreasing the amount of consumed energy, while at the same time, minimizing the delay. Another benefit of maximizing parallel transmission is that the total amount of interference is reduced, thus saving the network even more energy by reducing the cost of retransmission due to collisions.

Other parameters that can be used to unify the goals of minimizing delay in a network and maximizing network lifetime are choosing nodes to take part on routing based on their residual energy; choosing destination nodes based on link quality; and choosing nodes to maximize load balancing. This paper presents a novel routing protocol for WSNs, which aims to minimize the delay while simultaneously maximizing the network lifetime, by unifying the two goals with parameter constraining decision making processes based on data aggregation, residual energy, link quality, and load balancing.

The remaining sections of this paper are organised as follows. Section 2 presents related work in routing protocols for minimizing the delay and maximizing network lifetime. Section 3 gives the system model. Section 4 discusses the proposed protocol, which incorporates node residual energy, edge link quality, and network wide load balancing into the decision making process of a modified minimum latency aggregation scheduling (MLAS) protocol [1], [2]. Section 5 describes the experiments and simulations which compare the proposed protocol with basic MLAS. Section 6 concludes the paper and gives future work.

II. Related Work

Minimum latency aggregation scheduling (MLAS) is a well-known problem of constructing minimum delay routes by minimizing the number of time slots in routes and incorporating data aggregation [1]. Most of the previous MLAS work use two consecutive independent phases, i.e., tree construction phase and edge scheduling phase (time slot scheduling phase) [1], [3], [4], [5]. Two notable approaches are Chen et al.'s approach [4], which is based on the shortest path tree (SPT), and Huang et al.'s [3] and Wan et al.'s [1] approach, which is based on the dominating set tree (DST). However, two-phase approaches have some significant problems: First, the degree of latency varies greatly, even within the same algorithm, depending on the tree constructed initially. Second, the opportunity of parallel transmissions among nodes would decrease because the role of leaf nodes and non-leaf nodes are predefined. Therefore, Tian et al. [2] proposed a new aggregation scheduling by defining a greedy growing tree that only depends on one phase approach. This method easily assigns a minimum number of time slots for all the sender-receiver pairs. In all of the aforementioned works, the energy consumption of the network is largely ignored, and consequently, the network, while having minimum latency routes, dies prematurely, due to a large imbalance in node residual energy. This paper presents a novel routing protocol for WSNs, which starts with the work of MLAS [1], [2], and modifies this work by incorporating node residual energy, edge link quality, and network wide load balancing into the route formation decision making process.

III. System Model

Consider a network $G = (V, E)$ with V sensor nodes and E edges. There is only one sink node $b \in V$. Moreover, the communication graph of the network should be a digraph obtained from G by replacing each link in G with two oppositely directed edges $u \rightarrow v$ and $v \rightarrow u$. That means all the sensor nodes are located in Euclidean plane and are equipped with omnidirectional antennas. Furthermore:

1. All nodes have the same fixed transmission radius r.
2. For simplify the problem, the interference range equals to the transmission radius, $\rho = r$.
3. Each node works in half-duplex mode, so that it can either send or receive at one time slot.

Fig. 1 shows an example network. Two pairs of communication edges S1→D1 and S2→D2 are interference free because the two line segments (S1, D2) and (S2, D1) are both longer than ρ. However, S1 → D1 and S2 → D4 are exclusive interference edges that cannot be scheduled in the same time slot. In the subset of edges that belongs to interference free, multiple transmissions can be assigned in a same time slot. This model is referred to as the protocol interference model [6].

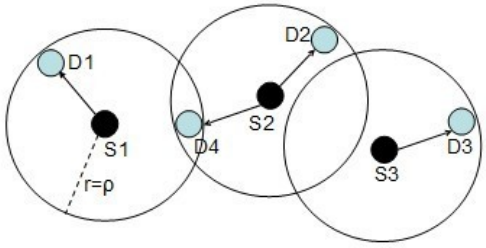


Fig. 1 Protocol interference model: Each node has a unit transmission radius and an interference radius.

Energy Model

The energy model used in this work is the same as in [7]:

$$\begin{aligned} E_{total} &= \sum_{i \in V} E_i \\ E_i &= E_{T_x} + E_{R_x} \\ E_{T_x} &= E_{elec} \times k + \epsilon_{amp} \times k \times r^2 \\ E_{R_x} &= E_{elec} \times k \end{aligned} \quad (1)$$

E_{total} is the total energy consumption of the whole network. k is the number of bits received and sent by node i, r is the transmission distance. The E_{R_x} and E_{T_x} are the energy consumption in reception and transmission, respectively. E_{elec} is the electrical energy required for the circuits in receiver and transmitter. The amplification energy for transmission is denoted by ϵ_{amp}.

Latency Model

Consider the network $G = (V, E)$ with V sensors. Define C_i as a subset of V. Therefore, the data aggregation process is a sequence of subsets $C_1, C_2 \ldots C_K, C_{K+1}$ that satisfy the following conditions.

1. $C_{k1} \supset C_{k2}, \ \forall 1 \leq k_1 < k_2 \leq K+1$
2. $C_1 = V$, and $C_{K+1} = Sink$
3. *Data packets are gathered from all the sensors to the sink from C_K to C_{K+1} in time slot K.*

Generally, the value of K is defined as the data aggregation latency.

IV. ENERGY EFFICIENT GREEDY GROWING TREE

The proposed algorithm modifies and extends the work done by Tian et al. [2]. In their paper, a Greedy Growing Tree (GGT) is proposed to schedule the aggregation process in a reverse order. In this algorithm, the spanning tree is constructed step by step, and the allocated time slot for each link is assigned along with the tree construction. In this way, the tree formation process starts from the sink. In the 1st step, the sink acts as the parent (receiver), and the algorithm searches for a child (sender). A sender is chosen based on a set of priority rules. This sender is connected to the sink and assigned a specific time slot. Any other node that is out of the interference range of the chosen node, but in the transmission range of the sink, is assigned the same time slot. In the 2nd step, the chosen node and any other nodes chosen (which was out of the interference range) in the 1st step now act as parents, and the algorithm searches and chooses the next sender (child). Similarly, any other node that is out of the interference range of the chosen node, but in the transmission range of the previously chosen parent(s), is assigned the same time slot. This continues until all of the sensors are connected to the spanning tree and every link has been assigned a specific time slot.

The basic idea of GGT is to construct larger and larger spanning trees rooted at the sink: It starts by considering the sink as the only node in the tree and tries to find the sender; in each subsequent round, all non-leaf nodes of the temporary spanning tree are the candidates of receivers; and all leaf nodes are the candidate of senders [2]. However, the GGT is only proposed to address the MLAS problem. In our system model, we also have to consider about the energy consumption by all the sensors. In order to expand the lifetime of the network, we have to balance the energy consumed by each node. Obviously, in the spanning tree of the network, non-leaf nodes consume more energy than the leaf nodes. A leaf node has only one link, i.e., for transmitting its own data packet. Thus, the energy consumed by leaf nodes can be represented as:

$$E_{leaf} = E_{T_x} = E_{elec} \times k + \epsilon_{amp} \times k \times r^2 \quad (2)$$

A non-leaf node has more than one link. There is one link for transmission, and i links for receiving data packets from all of its children nodes (e.g. the number of children is i, where ($i \geq 1$ and i is an integer)).

$$\begin{aligned} E_{non-leaf} &= E_{T_x} + i \times E_{R_x} \\ &= E_{elec} \times k + \epsilon_{amp} \times k \times r^2 + i \\ &\quad \times E_{elec} \times k \\ &where\ i \geq 1\ and\ i\ is\ an\ integer. \end{aligned} \quad (3)$$

The next step describes how to select the <*sender, receiver*> pairs in each round. The goal is to maximize the opportunity of parallel transmissions and balance the energy consumption by deciding which node should have larger value of i and which one should has smaller value of i. The proposed algorithm favors the minimization of latency by applying the priority rules of [2] firstly, and then, secondly, it applies the node residual energy, link quality, and load balancing constraints to the <*sender, receiver*> selection in each step of the tree construction process.

Basic Greedy Growing Tree (GGT)

The MLAS problem can be solved by using the priority rules with the GGT of [2]. This work asserts that the selection of <*sender, receiver*> pairs should be considered as beneficial for both current and later rounds, and the choices that benefit later rounds have priority. Thus, they define a Priority Rule for selecting proper <*sender, receiver*> pairs. To facilitate the discussion, the following definitions and notations are used:

Neighbor (I, V')	The set of neighbors of a node i in a node set V' of G.
NumNeigbor (I, V')	The number of neighbors of a node i in a node set V' of G.
NumNeigbor (I, $V' \backslash Z$)	The number of neighbors of a node i in a node set V' but not in Z of G.
Sizeof(V')	the number of nodes in a node set V' of G.
EnergyOf(i)	the remaining energy of sender.
EnergyOf(j)	the remaining energy of receiver.

Now, consider the network G = (V, E), let r as round index and T_r to be the temporary tree. The Priority Rules of GGT can be presented as follows:

Priority Rule 1:	First, sort all the nodes based on the increasing order of *NumNeigbor (I, T_r)*.
Priority Rule 2:	For nodes with the same order by Rule 1, sort them based on the increasing order of *NumNeigbor (I, $T_r \backslash Z$)*.
Priority Rule 3:	For nodes with the same order by Rule 1 and Rule 2, sort them based on lexicographic order.

The basic GGT algorithm uses he above rules to select <*sender, receiver*> pairs. As mentioned, this paper modifies and extends these rules to incorporate maximization of network lifetime. To better understand the extensions and modifications, the motivation and effect of the basic priority rules are explained first.

Fig. 2 shows an example wireless sensor network with 20 nodes. In this network, there are only two nodes in the temporary tree which are represented by color black. Now, consider the order of nodes a; b; c; d with Priority Rule to decide the sequence of adding them in the tree.

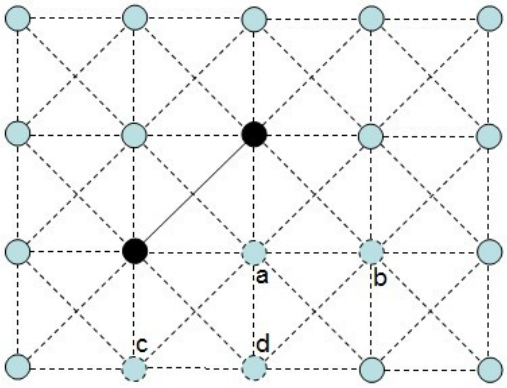


Fig. 2 Example network for Priority Rules.

Based on the properties of node a; b; c; d as shown in the Table, we can use the Priority Rules to determine the order of adding in the tree for nodes a; b; c; d.

Step1. Increasing order by Rule1: b = c = d < a
Step2. Increasing order by Rule2: c = d < b
Step3. Lexicographic order by Rule3: c < d
Step4. Final order: c < d < b < a

Therefore, node *c* has the highest priority and node *a* has the lowest priority among these four nodes. Accordingly, the algorithm would choose node *c* as the child of the parents, and node c becomes the parent for the next step. In next section, this example will be used again to explain the rationality of Priority Rule.

Rationale of GGT

According to the example network shown in Fig. 2, the GGT algorithm may be analyzed as follows:

For Rule 1 (Comparing nodes a and b):

Node *a* has two parents in the tree, but *b* only has one. For node *a*, all the neighbors of its two parents might be the sender in subsequent rounds. All of them would interfere with *a*, because two parents are definitely in *a's* transmission range. However, only neighbors of one parent would interfere with node *b*. Accordingly, we can say that *b* has more opportunity in parallel transmission than node *a*. Thus, node *b* has priority over node *a*.

For Rule 2 (Comparing nodes b and c):

Node *b* and *c* have the same number of neighbors in the tree. That is the reason why we need Rule 2. Node *c* has less number of neighbors outside of the tree than node *b*. That means if *b* or *c* acts as receiver after they join in the tree, choosing *c* may increase the opportunity of parallel transmission because it only has four potential senders. Only four potential transmissions would interfere with *c* in this way, but for *b*, the number of potential transmissions is seven.

Rule 3 is actually only used for braking the ties; so, it provides no help in addressing the MLAS problem or the maximizing the network lifetime problem.

With the above analysis of the Priority Rule, we find that each rule is reasonable for its particular situation. However, the reason why Rule 1 has higher priority than Rule 2 is not quite obvious. In fact, the following discussion shows that Rule 2 might have higher priority in certain circumstances. Let's consider a special case as shown in Fig. 3.

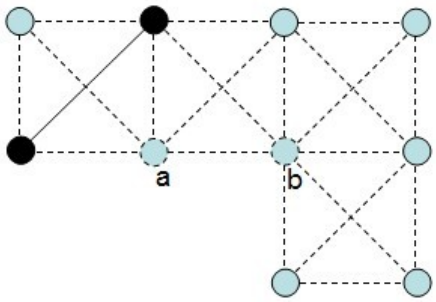


Fig. 3 Special case for Priority Rule.

Considering the scenario indicated by Fig. 3, node *b* would be given higher priority than node *a*, with the help of Rule 1. But, it is apparent that more collisions might occur if node *b* was chosen.

Special Node Processing in Basic GGT

The basic GGT algorithm proposed in [2] applies a higher priority to special nodes, such as the so called Articulation, Pilot and Critical nodes. Their rationale behind the special treatment stems from their concern that the network might become partitioned if these special nodes were not treated with higher priority than Priority Rule 1. They define an *Articulation node* as a node through which the traffic of a subset of other nodes $T \supset G$ must pass in order to be received by the sink. And, if this articulation node is not given higher priority, then the subset *T* would be cut-off

from the sink. However, our simulations show that the Articulation, Pilot, and Critical nodes are not necessary to consider first. Therefore, all the nodes can be treated equally and ordered simply by the Priority Rule if we only consider addressing the MLAS problem. So in our new routing protocol, the special treatment of nodes has been removed and the simulation shows no ill effects on the aggregation latency.

Energy Efficient Extension to Basic GGT

Since the network lifetime is significant in WSNs, we now take the energy consumption into account. There are two limitations in GGT algorithm: First, Rule 3 has no contribution in minimizing latency. Choosing a lexicographic order means that the order of the nodes would remain the same all the time. However, the order made by Rule 3 can be more useful if it is based on the remaining energy of the nodes. Second, GGT does not mention anything about the link selection; it only focuses on node selection. That means even if a node is chosen in a specific step, the link selection among all the possible links for this node is random. This process can also be modified if the quality of the link can be taken into account. Notice that if we do not change Rule 1 and Rule 2, the latency would not be changed (This latency is denoted as τ).

Energy Efficient Extensions

1. The static GGT can be easily changed to dynamic if the tree is reconstructed periodically. The minimum time of this periodic interval is set to the time required to completely gather the data from all the sensor nodes in the network; as such, this interval is the latency for one round τ. After every interval τ, the GGT is reconstructed. The added benefit of this modification is that the sensor load and energy consumption may be more balanced, different nodes will generally take part in the routes, since the link selection is random in the Priority Rules.
2. Rule 3 can be replaced by sorting the nodes in decreasing order of the remaining energy of each node. In this case, the node which has higher remaining energy would be chosen first. Since a node entering the tree has a higher chance to be a non-leaf node than the remaining nodes, with this extension, better energy balancing can be achieved, since a non-leaf node requires more energy in its role. (The number of links on a specific node is proportional to the priority of the node.)
3. Alternatively, Rule 3 can be replaced by sorting the nodes in decreasing order of link quality. The link quality is defined as *min{EnergyOf(i), EnergyOf(j)}*. The rationale behind this extension is that it would make more energy conserving sense to choose the node with the higher link quality than other contenders. For instance, if a node had higher residual energy, but its receiver had very low residual energy, then choosing that node would not be a good choice. Instead of choosing the node with higher remaining energy, this extension proposes to choose the best link between sender and receiver nodes.
4. Moreover, Rule 3 can be replaced by sorting the nodes in decreasing order of receiver residual energy. This extension proposes to sort the nodes in decreasing order of *max{EnergyOf(j)}*, where *j* is the index of nodes in the tree which can act as the receiver of this specific new sender.
5. Finally, the link quality between the sender and receiver may not be the most energy efficient choice, particularly when the receiver is over loaded. The basic GGT does not consider the number of links on the existing nodes in the growing tree. When a node is chosen by the Priority Rules in a certain step of the GGT algorithm, an arbitrary parent of the growing tree is chosen as the receiver of the chosen node's traffic. Perhaps, this is the closest parent. Many such nodes may also be sending their sensor data to this same arbitrarily chosen parent, and this parent would consume more energy than other parents in the range of the chosen node, thus leading to premature network death. It would be better to incorporate the number of links currently being handled by the parent into the decision making process. This extension proposes to use set the priority of parent node selection to be proportional to its remaining energy and inversely proportional to its connections. More details would be discussed in simulation part based on Fig. 13.

Algorithm for Energy Efficient GGT

Table 1 gives the one-round of the tree construction process proposed in the paper.

Table 1 Modified basic GGT construction algorithm.

Step	Action
0	**Initialization:** Start from the sink node by regarding it as the only node in the T_r, $r \leftarrow 1$.
1	**Adopt Rule 1:** Choose the node which has highest priority in Rule 1 as the next member.
2	**Single Node Case:** If only one node is chosen by Rule 1, go to Step 6. Otherwise, continue.
3	**Adopt Rule 2:** Choose the node which has highest priority in Rule 2 as the next member.
4	**Single Node Case:** If only one node is chosen by Rule 2, go to Step 6. Otherwise, continue.
5	**Remaining Energy judgement:** Choose the node along with the best link by choosing the one which has higher *min{EnergyOf (i), EnergyOf (j)}*. If more than one node is chosen, determine the selection based on *max{EnergyOf (i)}*. If the final decision is still uncertain, choose the link randomly. Goto Step 7.
6	**Link selection:** Connect the best link which has *max{EnergyOf (j)}* or *max{EnergyOf (j)/(number of j's childs + 1)}*. If more than 1 link has the highest remaining energy, then choose it randomly.
7	**Iteration Rule:** $r \leftarrow r + 1$, if $r \neq V$, go to Step 1. Otherwise Stop.

The algorithm shown in Table 1 indicates one-round tree construction. In simulation, the tree would be reconstructed every time period τ. The details of the algorithm for Rule 1, Rule 2 and interference judgement can be found in [2]. So next, only the detailed algorithm for energy judgement is listed in the following.

Table 2 Energy balancing algorithm.

Step	Action
0	**Initialization:** The initial energy of each node is set as E0 and energy of sink is assumed to infinity.
1	**Attribute update of each potential candidate:** After finding out the potential candidate for the tree (based on the transmission range), update their attribute of energy from its own remaining energy to *min{EnergyOf (i), EnergyOf (j)}*.
2	**New Rule:** If a winner cannot be obtained after Rule 1 and 2, then compare their attributes of energy derived from step 1. If multiple nodes still have the same attribute of energy, then perform a random selection.
3	**Decide its parent in the Tree:** Two different ways: 1: choose the parent which has max*{EnergyOf (j)}*. 2: choose the parent which has *max{EnergyOf (j)/(number of j's childs + 1)}*.
4	**Complete the tree construction and start transmitting:** Data aggregation starts after the tree is constructed. At the end of each round, calculate the remaining energy of each node.
5	**Renew status:** After each round of transmission, remove the dead ones and renew the status of each node for next tree construction process. And go back to Step 1.

There would be 2 main phases in the whole process. In the 1st phase, the tree construction process is performed time4-slot-by-timeslot until all the nodes are added in the tree. In the 2nd phase, data transmission is performed. After each round of data transmission, the tree is reconstructed.

EEGGT for Example Network

Consider an example network with 12 sensor nodes and one sink. Small letters indicate the index of all the nodes and their remaining energy are represented by different numerical numbers. Assume that the remaining energy of the sink is positive infinite and energy consumed in transmission and reception are 2 units and 1 unit respectively. The initial network is shown in Fig. 4.

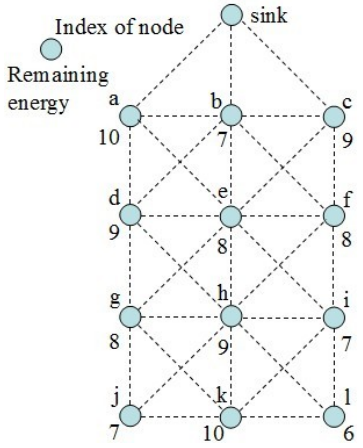


Fig. 4 Initial network for tree construction.

In the first round, the tree construction starts from the sink. Obviously, node a; b; c are the next potential members in the tree. By adopting Rule 1, three nodes get the same order. Then we need to use Rule 2, *a* and *c* have higher priority because they have less number of neighbors not in the temporary tree. In order to determine the selection between node *a* and *c*, the remaining energy judgement in step 5 is necessary. With all the criteria illustrated above, '*a* → sink' is chosen in the first time slot *s1*. Next, the node *b, c, d*, and *e* become the candidates of potential members. With Rule 1, we get an order $c = d = e < b$. Since some nodes have the same order, Rule 2 is applied to figure out a new order, thus we have $c < d < e < b$. Apparently, the next member *c* can be chosen and d can be chosen simultaneously in time slot *s2* because of the interference free judgement. This tree construction process continues regarding to the Energy Efficient GGT algorithm until all the nodes are connected in the spanning tree. This process is completely shown in Fig. 5 (A).

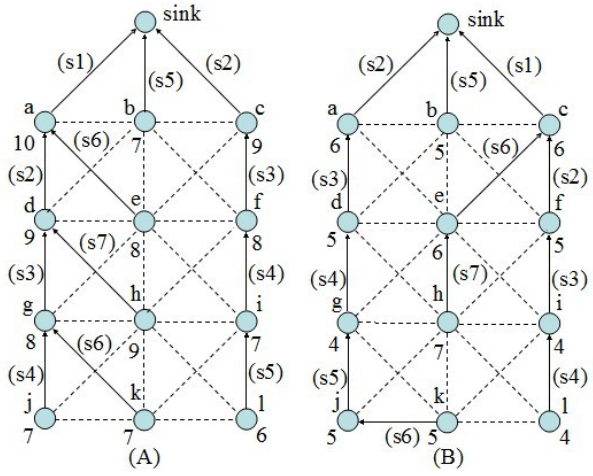


Fig. 5 (A)Tree constructed in 1st round. (B)Tree constructed in 2nd round.

However, after one round data aggregation process, the remaining energy of the nodes changes to the situation in Fig. 5 (B). We can easily find out the difference of the output of (A) and (B). The tree reconstructs in a different way because the factor of remaining energy affect the node selection and link selection. For instance, node *h* is chosen in the last order in both (A) and (B), but they select different link for h. In (A), node d has the highest remaining energy of 9, so it is more suitable to be the receiving node of h. In (B), node e has the highest remaining energy of 6 because d consumed more energy in the first round than e, which means h → e becomes a better choice.

This example shows that the Energy Efficient GGT algorithm does really balance the energy consumption among all the nodes by periodically reconstructing the data gathering tree. Moreover, the latency caused in (A) and (B) is the same in 7 time slots. In other words, Energy Efficient GGT algorithm maintains the latency of traditional GGT algorithm in solving MLAS problem, but, it improves the GGT algorithm by increasing the network lifetime.

V. SIMULATION

First, randomly deploy N sensors into a square region with edge length L; the density of node is determined by $O(N/L^2)$. Besides, the sink and all the sensors have a same transmission range λ. In multi-hop communications, the networks are considered to be fully connected when all the nodes are reachable. The connectivity depends on the radio range; thus, the radio range of the nodes should be configured optimally. In the simulation, we choose $N = 50$, $L = 50$ so that $O(N/L^2) = 0.02$. In order to ensure the network connectivity, λ has 3 different values 15; 20; 35. Since the network deployment and the location of sink have great impact on the performance, two different sensor networks with sink located at the corner and center, respectively, were setup. The remaining simulations are all based on the networks shown in Fig. 6 and Fig. 7.

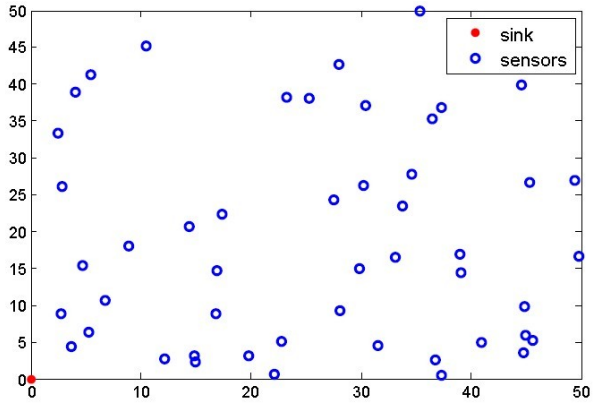


Fig. 6 Random deployed sensors network with sink located at the corner.

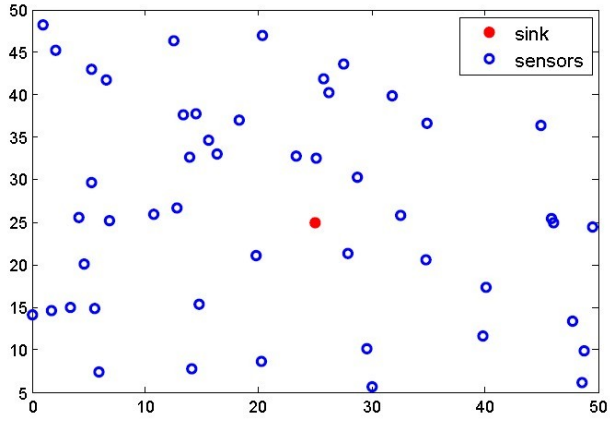


Fig. 7 Random deployed sensors network with sink located at the center.

The location of the sink does not only affect the time scheduling for data aggregation, but also leads a great difference in performance of network lifetime.

Figure 8 examines the correctness of my idea that considering special nodes is almost useless. Modified GGT shows almost the same performance in aggregation latency with traditional GGT; however it simplifies the time scheduling process by removing the step for considering local optimization which may be caused by some special nodes.

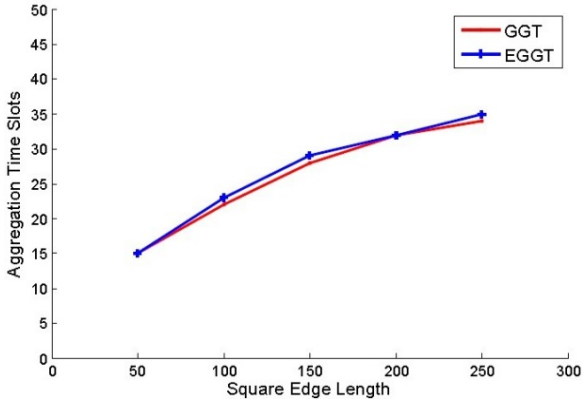


Fig. 8 With sink at the center, the performance of aggregation latency.

Traditional GGT is considered as a static transmission process which means that the tree would never be changed, it is denoted as SGGT. On the other hand, the tree construction can be re-built every round based on the remaining energy of the sensors, this process is called DGGT. EEGGT is an energy efficient GGT which is based on our proposed algorithm; the MLAS problem, remaining energy of each child node and parent node, the link quality and energy balancing are all taken into account. Simulation

ends when there is no path to the sink, meaning that all neighbors of sink are out of energy.

In Fig. 9, we can easily find that EEGGT has longer network lifetime that it runs more rounds based on the same random sensor network. And in this situation, only EEGGT makes all sensors exhausted; this implies the energy balancing performance of the protocol. The difference would be more obvious for lower network nodes density.

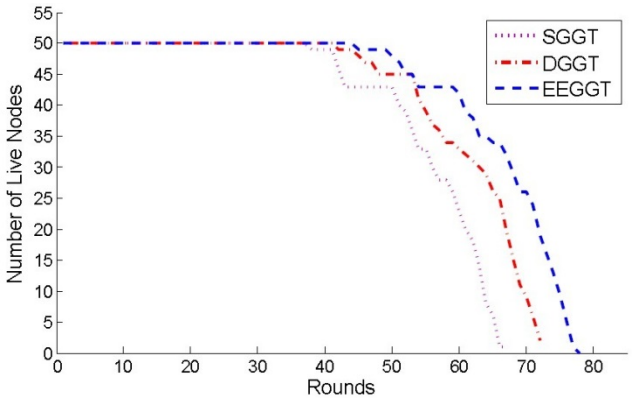


Fig. 9 With sink at the corner, the performance of network lifetime.

Consider the sink located at the center of the field in Figure 10. Using EEGGT still has advantage in expanding network lifetime and we can find that SGGT has more live nodes remaining when the network has been already collapsed (neighbors of sink are all running out of energy). However, the EEGGT method best balances the energy consumption.

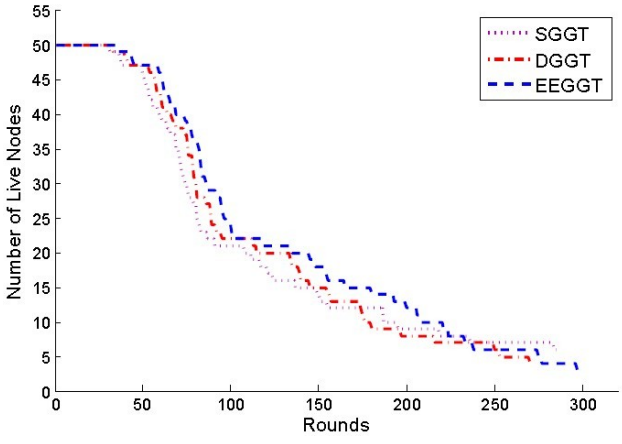


Fig. 10 With sink at the corner, the performance of network lifetime.

Keep the sink located at the corner and change the transmission range λ = 15 and 35 to see the difference. The longer transmission range leads to more choices for selecting new < *sender, receiver* > pair. In Fig. 11, the choices of choosing new node and adding new < *sender, receiver* > pairs are highly limited by the transmission range λ = 15. If the tree construction process is static, the network breaks down very fast. However, for DGGT and EEGGT, since they will reconstruct the tree in every round, they can choose different new < *sender, receiver* > based on current remaining energy of nodes. In this case, SGGT shows significant inferiority in network stability and lifetime. Whereas, EEGGT still performs better.

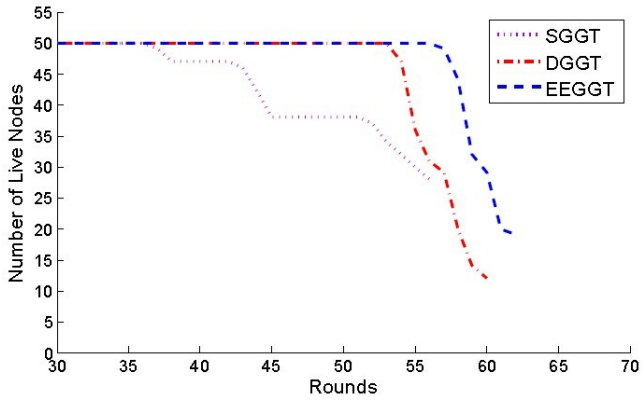


Fig. 11 With λ = 15, the performance of network lifetime.

Figure 12 indicates another problem of the network. When λ is relatively high, the simulation can last for many rounds even the number of live nodes is very small. Since the sink can reach most of the nodes in this extreme case, the network turns to be very similar to direct transmission after some long distance nodes died.

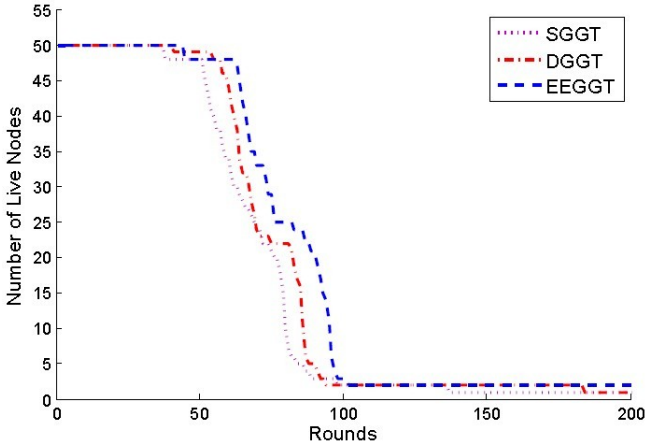


Fig. 12 With λ = 35, the performance of network lifetime.

Actually, when we have already made the decision on which new node to add in the tree, it is not optimal for energy consumption if we choose to connect the parent which has higher remaining energy. Because, the tree is reconstructed at the beginning of next round, the data transmission and tree construction is not working simultaneously, which means that maybe a lot of nodes would choose the same parent in

one tree construction process. So, some nodes would deplete its energy very fast. We can also consider the number of child nodes of these potential parents, which should be inversely proportional to the priority of parent selection. As in 2- EEGGT, each new added node chooses its receiver which has *max{EnergyOf (j)/(number of j's childs + 1)}*, the performance is enhanced in a certain extent as shown in Fig. 13.

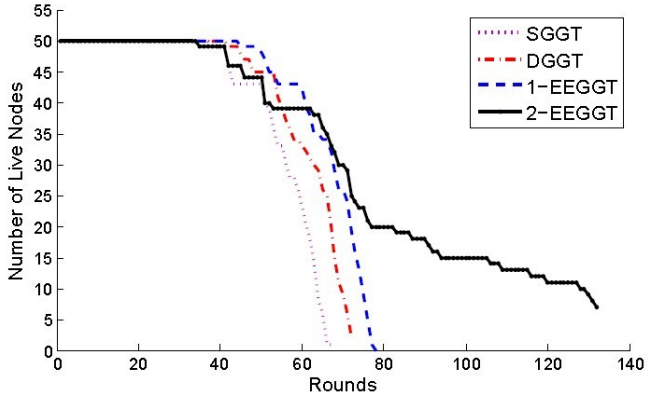


Fig. 13 The performance with different parent searching method.

VI. Conclusion and Future Work

In this paper, a modified greedy growing tree EEGGT is proposed. It simplifies the original GGT by removing the process of differentiating special nodes. And EEGGT does not only aim to minimize the aggregation latency in multi-hop wireless sensor networks, but also balance the energy consumption so that expanding the lifetime of the network. Though minimizing data aggregation time is still considered in the first place, the proposed novel protocol shows good energy efficiency. Simulation demonstrates that EEGGT has nearly the same performance in minimizing latency as traditional GGT. Furthermore, this modified protocol has much better energy saving and balancing. Moreover, numerical experiment also indicates that different parent searching method would lead to a great difference in the performance of network lifetime. However, the Minimum Aggregation Latency Problem and Energy Efficiency Problem are considered separately with different priority in this work. Future work can consider to set up joint rules for finding new <*sender, receiver*> pairs for these two goals. Besides, the transmission range is assumed fixed and same for all the sensors in this paper. For a more general case that sensors could adjust their transmission radius, the tree construction process would be much more complicated for the difficulty in interference judgement. And it needs to be considered as a Minimum Spanning Tree.

References

[1] P. J. Wan, C. H. Huang, L. Wang, Z. Wan and x. Jia, "Minimum latency aggregation scheduling in multihop wireless," in *Proceedings of the tenth ACM international symposium on Mobile adhoc networking and computing*, New York, 2009.

[2] C. Tian, H. Jiang, C. Wang, Z. Wu, J. Chen and W. Liu, "Neither shortest path nor dominating set: Aggregation scheduling by greedy growing tree in multihop wireless sensor networks," *IEEE Transactions on Vehicular Technology,* vol. 60, pp. 3462-3472, 2011.

[3] S. H. Huang, P. J. Wan, C. Vu, Y. Li and F. Yao, "Nearly constant approximation for data aggregation scheduling in wireless sensor networks," in *26th IEEE International Conference on Computer Communications*, 2007.

[4] X. Chen, X. Hu and J. Zhu, "Minimum data aggregation time problem in wireless sensor networks," in *Proceedings of the First international conference on Mobile Ad-hoc and Sensor Networks*, Heidelberg, Berlin, 2005.

[5] V. Annamalai, S. Gupta and L. Schwiebert, "On tree-based convergecasting," vol. 3, pp. 1942-1947, March 2003.

[6] P. Gupta and P. Kumar, "The capacity of wireless networks," *Transactions on Information Theory,* vol. 46, pp. 388-404, March 2000.

[7] W. Heinzelman,, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference*, 2000.

Enhancing Productivity of Costs and Energy through VNC and SAN

Sheida Shirazi[1], Mohammad Heidari Reyhani [2], Mohammad Ganji[3], Marjan Abdyazdan[4]

[1]Department of Computer Engineering, Mahshahr branch, Islamic Azad University, Mahshahr, Iran.
e-mail: shirazi85@gmail.com
[2]Department of Computer Engineering, Mahshahr branch, Islamic Azad University, Mahshahr, Iran.
e-mail: mohammad.hr2010@gmail.com
[3]Department of Computer Engineering, Tarbiat modares University, Tehran, Iran.
e-mail: m_ganji2011@yahoo.com
[4]Department of Computer Engineering, Mahshahr branch, Islamic Azad University, Mahshahr, Iran.
e-mail: m.abdeyazdan@mahshahriau.ac.ir , abdeyazdan87@yahoo.com

Abstract
SAN, which stands for Storage Area Network, is a data sharing architecture widely used for its great security and access as well as its compatibility with a variety of operating systems. In view of the importance of networks optimization, this system has recently shifted from a sole data sharing to resource sharing. As the reduction of energy consumption has become a vitally important issue worldwide, Iran has also taken appropriate actions in this regard which subsequently embraced by different organizations. A great deal of energy consumption reduction and costs cutbacks could be achieved through utilizing Virtual Network Computing in the system. The present article presents procedures for reducing energy consumption.
Key words: Virtual Network Computing, Resource sharing, SAN architecture, optimization, energy consumption reduction

Introduction

In the past, it was believed that every user would need separate and allocated resources; nowadays, however, we need to make offices smaller and reduce expenses as much as possible due to economic reasons, such as the rent of a place, cost of electricity, and air-conditioning. On the other hand, technical issues may also arise that necessitate the centralization and making offices smaller of which maintenance, imposing restrictions, having access to hardware, updating essential software programs on a server, to name but a few. In recent years, great innovations and technological advancement have emerged in this area that could generally be considered as solutions one of which is resource sharing services.

As the article continues, concepts of resource sharing, its subcategories, and implementation will be introduced, then a study of energy consumption and the conclusion will follow.

Resource sharing

Resource sharing familiarizes us with the concept that multiple users could utilize a piece of hardware simultaneously. For instance, a single processor may process multiple systems or a Grid Computing central system, which is a special type of resource sharing, may distribute input processing to other systems having lower amount of load.

These actions enjoy multiple advantages: firstly, the speed of processors will be boosted. Secondly, systems with higher load could be assisted by idle systems. Lastly, maximum productivity will be achieved by means of keeping time and hardware to a minimum. The upcoming sections will explain more area of resource sharing.

Virtual Network Computing (VNC)

Virtual Network Computing could be considered as a subcategory of resource sharing. This technology is currently very sophisticated and important as a representation of which could be found in every large computer network and server room. In virtualization as well as centralizing processing in one place, it is also possible to meet the demands of a particular server as nothing more than a display, keyboard and mouse would be needed and an interface or thin client can link them to the server. By help of this technology, the need for multiple computers in a workstation is eliminated and the heat produced will be reduced compared to before. The consumption of energy is reduced and there is no need for physical activity to upgrade systems. A practical example of this type of network will follow.

Virtual Machine (VM)

After understanding how hardware is used, it is time to learn about network software, that is, the operating system. Virtual operating systems or virtual machine make it possible for mother or principal systems to be

implemented on a server and clients to use their own separate operating system. In this case, installing multiple operating systems on a virtual machine, we enable users to use their OS to meet their needs without any change in clients. An advantage of these systems is that the need to install and run clients and OSs that serve all clients will be removed once they have been installed for the first time. Memory space will be optimized due to centralizing and eliminating the needs for multiple OS and software installation.

SAN architecture

SAN is a data sharing architecture that enables saving of all system data through substrate network. Enjoying high determination coefficient, data sustainability and security, SAN is employed by a large number of network users. One of its foremost characteristics, namely being centralized, it could be protected in the event of fire and earthquake and could be the best place for backup servers. SAN is best coordinated with and consolidated into modern virtual technologies to produce the best efficiency and productivity for users. SAN, per se, could also be used for regular systems and under different models with a capacity of 200 terabytes.

Employing SAN and VNC together

By combining the above-mentioned technology and architecture, SAN is employed to meet the requirements of storing and maintaining data. On the other hand, VNC is employed for centralized system processing and as a substrate operating system in a way that clients are connected to main VNC server through the Ethernet or optical fiber and SAN is usually linke to VNC through optical fiber.

Considering the implementation of the network above, we expect to meet the following objectives:

- Centralizing of the core of all systems in a unitary space
- Removing computer case from users surroundings
- Increasing data coefficient of determination
  - All data are kept on a server in a separate room so it is much easier to protect one room against natural disasters and human infiltration than is the case with several hard disks of a large number of systems.
- Reducing maintenance costs
  - Whenever there are a small number of physical parts, damages will be significantly decreased because all critical hardware is centralized in a VNC server.
  - It is easier to effect change and repair software when there are VM and a central operating system.
- Reducing computer accessories procurement costs
  - Due to not having to buy systems of different usability and capacity, it will be possible to purchase all needed parts uniformly and in bulk i.e. monitors and mouse. As a result, the power of these systems could be defined m means of VNC according to users' needs.
- Reducing ambient temperature
  - Through removing computer cases from workstations and replacing 10 computers with one server, the ambient temperature will be reduced and so will the air-conditioning costs.
- Increasing productivity through decreasing costs
- Reducing labor force in hardware department
- Upgrading systems through improving access
- The ability to use multiple OS on a server simultaneously
- Energy consumption optimization
  - Every system, whether working or standby, consumes energy. When it becomes virtual the energy consumption will be decreased by more than 1/5.
- Imposing restriction and monitoring fairly easily

Presenting a proposal for Masjed Soleiman Petroleum and Gas headquarter

The hardware and software requirements of headquarter having been examined and estimated, the following system was proposed which is currently being assessed.

The capabilities of the proposed server to serve 30 clients are listed as shown in the table below:

Main server 7G 380 HP DL

Product description	DL 380G& X5670 12MB (2P)
Processor family	Intel® Xeon® 5600 series
Processor Core available	6 or 4 or 2
Number of Processors	3
Maximum memory	384 GB
Memory type	PC3-10600R RDIMMs DDR3
Memory slots	18 DIMM slots
Storage controller	(1) Smart Array P411/1GB
Power supply	(2) 750 Watt hot plug CS HE

The parts that should be installed corresponding to the requirements of the head quarter are listed in the following table:

description	Qty	Model	Sum
CPU	3	X5670 (6 core, 2.93 GHz, 12MB L3, 95W)	18 core ,52.74 GHz
Memory	18	HP-16 GB, DIMM 240-pin, DDR3	288 GB

We need three servers like the one described above to meet our needs accordingly while having better efficiency. Consequently, in case of any problem for one of the servers, the other two servers will compensate.

The space needed for storage on SAN whose server enjoys these specifications:

P2000 G3 FC Controllers

Capacity	SFF: 21.6 TB SAS
Controller Cache	2 GB per controller
Protocol (host connect)	8 Gb Fibre Channel
Sequential Reads MB/s	1,572
Sequential Writes MB/s	790
OS Support Fibre Channel ports	• Microsoft Windows Server 2008 IA32, x64, IA64 (Standard, Enterprise, Datacenter) • Microsoft Windows 2003 SP1, SP2, and R2 and 2003 R2 IA32, x64 • Red Hat Linux (32/64) SuSE SLES (32/64) • Hyper-V VMware OpenVMS • Apple Mac OS • Solaris 10 (x86)
SAS & SATA Drives (SFF 2.5-inch)	HP 900GB 6G SAS 10K rpm SFF (2.5-inch) Enterprise 3yr Warranty Hard Drive
Input Power Requirements	110VAC 3.32A, 344-390 W; 220VAC 1.61A,374-432W

We need two of these severs each of which having a volume of 16 TB. The above-mentioned server is linked to the main server, to which clients connect, through optical fiber. At the site of headquarter a built-in wiring is used for making a network so RG45 cables are needed. The other equipment needed for running a centralized system including monitor, keyboard, mouse, and Thin Client- HP 2020.

To manage operating systems, VM Ware Base software is installed on the server and all clients will have access to Windows 7, Ultimate installed on VM.

Energy consumption

In a discussion of energy consumption, a comparison will be made between the consumption in traditional networks and the networks with new approaches such as the model of National Petroleum Company:

1. Consumption in traditional network serving 30 clients with one server: The consumption for each system is as the following:

A. Monitor 120W
B. Case power 400 W
Sum= 520W
The amount of consumption for one server
Server 1000W= (30*520) +1000=16600W

2. Consumption in networks like National Petroleum Company:

A. consumption per client

Monitor 150W+ Thin Client= 4500= 30*150= Sum

B. consumption per VNC server

W285W= 95*3 CPU= W1500= 2*750 Server= 5355W= (3*1785) Sum=

C. consumption per SAN server

864W= 2W*432Sum=

Total consumption:

10719= 864+5355+4500Sum=

As a result, it would be apparent that the new system, in addition to all abovementioned advantages, consumes about 600 W less which consequently saves up more money.

Conclusion

It could be possible to save up more energy by means of resource sharing in SAN and virtualization. Moreover, instead of using several computers, one server could be used to serve all computers, which may otherwise work without any load or may be on standby and consume energy, so that much less energy is used by far. Using software, virtualization makes managing resource possible and consolidates purchasing hardware that in turn results in cutbacks. Additionally, due to being more centralized, a significant reduction in security and maintenance costs are expected.

References:

[1] Hewlett-packard,Hp Development Company, L.P, HP P2000 G3 Modular Smart Array System, 2011
[2] www.novell.com
[3] h10010.www1.hp.com
[4] en.wikipedia.org/wiki/Xeon
[5] russell- peters, Hp Development Company, L.P, The HP 6400/8400 Enterprise Virtual Array, 2011
[6] Hp Development Company, The HP ProLiant DL380 G7 Server.PDF, 2010
[7] Mohammadi, Gh. (2010) ,ghasem,

Enhancement Energy Efficient Routing in WSN

Rushdi Hamamreh
Computer Engineering Department
Al-Quds University
Jerusalem – Palestine
Email: rhamamreh@eng.alquds.edu

Mahmoud Arda
Computer Engineering Department
Al-Quds University
Jerusalem – Palestine
Email: Mahmoud_arda@jbs.com.jo

Abstract: **the emergence of wireless sensor networks (WSNs) is essentially toward the miniaturization and ubiquity of computing devices. Sensor networks are composed of thousands of resource constrained sensor nodes and also some resourced base stations are there. The route of each message destined to the base station is really crucial in terms network lifetime. This paper introduces a new routing algorithm based on minimum energy and residual battery algorithms. Using energy threshold to switch routing path.**

***Keywords-*: WSN; Energy-aware routing; routing protocols; meta-data, Negotiation, network lifetime, energy threshold.**

1. INTRODUCTION:

A wireless sensor network (WSN) [1, 2] in its simplest form could be defined as a network of (possibly low-size and low-complex) devices denoted as nodes that can sense the environment and communicate the information gathered from the monitored field through wireless links; the data is forwarded, possibly via multiple hops relaying, to a sink that can use it locally, or is connected to other networks (e.g., the Internet) through a gateway [1]. . Each node has three basic components:

1. Sensing unit
2. Processing unit
3. Transmission unit

The node senses the data from the environment processes it and sends it to the base station. These nodes can either route the data to the base station (BS) or to other sensor nodes such that the data eventually reaches the base station. In most applications, sensor nodes suffer from limited energy supply and communication bandwidth. These nodes are powered by irreplaceable batteries and hence network lifetime depends on the battery consumption. Innovative techniques are developed to efficiently use the limited energy and bandwidth resource to maximize the lifetime of the network. These techniques work by careful design and management at all layers of the networking protocol. For example, at the network layer, it is highly desirable to find methods for energy efficient route discovery and relaying of data from the sensor nodes to the base station.

The route of each message destined to the base station is really crucial in terms network lifetime. On the other hand there are many factors that affect the network life time such as topology of the network, the transmission rate, transmission range and routing protocol.

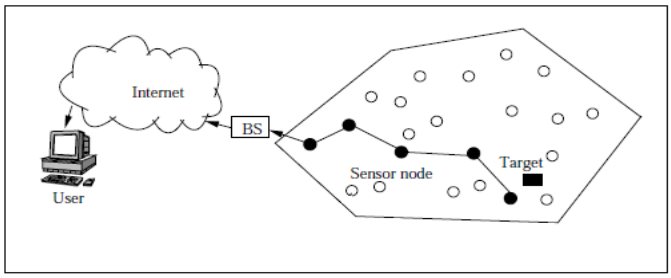


Figure 1 : WSN structure

The simplest forwarding rule is to flood [3] the network: Send an incoming packet to all neighbors. As long as source and destination node are in the same connected component of the network, the packet is sure to arrive at the destination. To avoid packets circulating endlessly, a node should only forward packets it has not yet seen (necessitating, for example, unique source identifier and sequence numbers in the packet). Also, packets usually carry some form of expiration date (time to live, maximum number of hops) to avoid needless propagation of the packet (e.g. if the destination node is not reachable at all). While these forwarding rules are simple, their performance in terms of number of sent packets or delay. Determining these routing tables is the task of the routing algorithm with the help of the routing protocol. In wired networks, these protocols are usually based on link state or distance vector algorithms (Dijkstra's or Bellman–Ford [4], [5]). In a wireless, possibly mobile, multi hop network, different approaches are required.

Routing protocols here should be distributed, have low overhead, be self-configuring, and be able to cope with frequently changing network topologies. This question of ad hoc routing has received a considerable amount of attention in the research literature and a large number of ad hoc routing protocols have been developed.

A commonly used taxonomy [6] classifies these protocols as either (i) **table-driven** or proactive protocols, which are "conservative" protocols in that they do try to keep accurate information in their routing tables, or (ii) **on-demand** protocols, which do not attempt to maintain routing tables at all times but only construct them when a packet is to be sent to a destination for which no routing information is available. In addition to energy efficiency, resiliency also can be an important consideration for WSNs. For example, when nodes rely on energy scavenging for their operation, they might have to power off at unforeseeable points in time until enough energy has been harvested again. Consequently, it may be desirable to use not only a single path between a sender and receiver but to at least explore multiple paths. Such multiple paths provide not only redundancy in the path selection but can also be used for load balancing, for example, to evenly spread the energy consumption required for forwarding.

2. RELATED WORK

AODV [7] is the widely used algorithm for both wired and wireless network. Ad-Hoc On-Demand Distance Vector is known as one of the most efficient routing protocols in terms of using the shortest path and lowest power consumption. AODV is a reactive protocol that builds routes between nodes on-demand i.e. only as needed. Messages to other nodes in the network do not depend on network-wide periodic advertisements of identification messages to other nodes in the network.

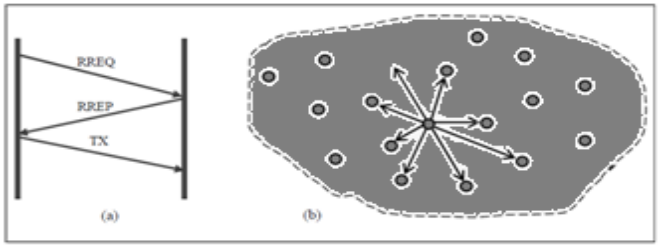


Figure 2: a) Timing diagram b) Hello packet

It broadcasts "HELLO" messages to the neighboring nodes. It then uses these neighbors in routing. Whenever any node (Source) wants to send a message to another node (Destination) that is not its neighbor, the source node initiates a Path Discovery in which the source would send a Route Request (RREQ) message to its neighbors. Nodes that receive the Route Request could update their information about the sending node. The RREQ should contain the IP address of source node. On the other hand the RREQ contains a broadcast ID that necessary to identify that RREQ. The RREQ has to have a current sequence number that determines the freshness of the message. Finally, the RREQ should keep track of the number of nodes that visited through path discovery in a variable of Hop Count. When a node receives a RREQ it would check wither it has received the same RREQ earlier (using IP, ID, and Sequence number), if so, it would discard it. On the other hand, if the recipient of the RREQ was an intermediate node that doesn't have any information about the path to the final destination, the node increases the hop count and rebroadcasts the RREQ to its neighbors. If the node that received the RREQ was the final destination, or an intermediate node that knows the path to the final destination, it sends back the Route Reply (RREP). This RREP should keep track of traverse path of the RREQ but from destination to source. As shown in figure 2, when the source node receives the RREP, it should then start sending data.

Heinzelman et.al.[8] Proposed a family of adaptive protocols called Sensor Protocols for Information via Negotiation (SPIN) that passes all the information at each node to every node in the network assuming that all nodes in the network to be a potential base-stations(BS). In this algorithm the user has the ability to query any node and get the required information or data immediately. These algorithms make assumes that nodes in close proximity have similar data, and hence there is a need to only distribute the data that other nodes do not posses. The SPIN family of protocols uses data negotiation and resource-adaptive algorithms. Nodes running SPIN assign a high-level name to completely describe their collected data (called meta-data or Meta content). Meta data in its simplest definitions is describes as Data of Data. That's it, Meta data should provide data about one or more aspects of the original data, for example Meta data aspects may be the Mean of creation of that data, Purpose of the data, Time and date of creation, Creator or author of data, and Location on a computer network where the data was created) and perform meta-data negotiations before any data (we means here original data) is

transmitted. Its importance arises from the fact that we have used to make sure that there is no redundant data sent throughout the network. That's it to reduce the overhead on the network and to save power. The semantics of the meta-data format is application-specific and is not specified in SPIN. For example, when sensors want to send meta-data for an event in certain area, it would use its ID. On the other hand, SPIN algorithm has the ability to access to the energy level of the node and monitor the protocol it is running according to how much energy is remaining in a certain node. These protocols are known as a time-driven fashion and broadcast the information all over the wireless sensor network, despite the fact that the user does not request any data at that moment. SPIN's **meta-data** negotiation approach solved the traditional problems of flooding, and thus achieving a lot of energy efficiency because you send meta-data, not all data as used to in flooding. In SPIN, there are three stages in which sensor nodes use three different types of messages ADV (advertise) REQ (request) and DATA to communicate with other nodes. ADV is used to advertise new data, REQ to request data by the node or sink or user itself and DATA is the actual message itself. The protocol starts when a node gets new data that it is willing to share with other nodes, after that it broadcasts an ADV message containing meta-data. If any nodes that receive ADV was interested in that data, it sends a REQ message for the DATA and the DATA is sent to this neighbor node. The neighbor sensor node then repeats this process with its neighbors. As a result, the entire sensor area will receive a copy of the data.

3. POWER AWARE ROUTING

Several algorithms had been developed for routing in wireless sensor network, some of these algorithms and protocols are energy based algorithms. In these algorithms we take the network graph, assign to each link a cost value that reflects the energy consumption across this link, and pick any algorithm that computes least-cost paths in a graph. An early paper along these lines is reference [10], which modified Dijkstra's shortest path algorithm to obtain routes with minimal total transmission power.

One of the most important algorithms used is known as minimum energy per packet or per bit. The most straightforward formulation is to look at the total energy required to transport a packet over a multi hop path from source to destination (including all overheads). The goal is then to minimize, for each packet, this total amount of energy by selecting a good route. Minimizing the hop count will typically not achieve this goal as routes with few hops might include hops with large transmission power to cover large distances – but be aware of distance-independent, constant offsets in the energy-consumption model. Nonetheless, this cost metric can be easily included in standard routing algorithms. It can lead to widely differing energy consumption on different nodes [11].

Some researches went to routing considering available battery energy, as the finite energy supply in nodes' batteries is the limiting factor to network lifetime, it stands to reason to use information about battery status in routing decisions. Some of the possibilities are **Maximum Total Available Battery Capacity** Choose that route where the sum of the available battery capacity is maximized, without taking needless detours (called, slightly incorrectly, "maximum available power" in reference [12]). **Minimum battery cost routing** Instead of looking directly at the sum of available battery capacities along a given path, MBCR instead looks at the "reluctance" of a node to route traffic [11, 13]. This reluctance increases as its battery is drained; for example, reluctance or routing cost can be measured as the reciprocal of the battery capacity. Then, the cost of a path is the sum of this reciprocals and the rule is to pick that path with the *smallest* cost. Since the reciprocal function assigns high costs to nodes with low battery capacity, this will automatically shift traffic away from routes with nodes about to run out of energy.

Min–Max Battery Cost Routing (MMBCR) This scheme [11, 13] follows a similar intention, to protect nodes with low energy battery resources. Instead of using the sum of reciprocal battery levels, simply the largest reciprocal level of all nodes along a path is used as the cost for this path. Then, again the path with the smallest cost is used. In this sense, the optimal path is chosen by minimizing over a maximum. The same effect is achieved by using the smallest battery level along a path and then maximizing over these path values [12]. This is then a maximum/minimum formulation of the problem. **Minimize variance in power levels** To ensure a long network lifetime, one strategy is to use up all the batteries uniformly to avoid some nodes prematurely running out of energy and disrupting the network. Hence, routes should be chosen such that the variance in battery levels between different routes is reduced.

Minimum Total Transmission Power Routing (MTPR) Without actually considering routing as such, Bambos [15] looked at the situation of several nodes transmitting directly to their destination, mutually causing interference with each other. A given transmission is successful if its SINR exceeds a given threshold. The goal is to find an assignment of transmission power values for each

transmitter (given the channel attenuation metric) such that all transmissions are successful and that the sum of all power values is minimized. **MTPR** is of course also applicable to multihop networks.

Archan Misra and Suman Banerjee [14] used to Maximize Network Lifetime for Reliable Routing in Wireless Environments (**MRPC**), they depended on the fact that selecting the path with the least transmission energy for reliable communication may not always maximize the lifetime of the ad-hoc network. On the other hand since the actual drain on a node's battery power will depend on the number of packets forwarded by that node, it is difficult to predict the optimal routing path unless the total size of the packet stream is known during path-setup. **MRPC** works on selecting a path, given the current battery power levels at the constituent nodes, that maximizes the total number of packets that may be ideally transmitted over that path, assuming that all other flows sharing that path do not transmit any further traffic.

4. PROPOSED WORK

MRPC algorithm has a problem in that it uses a path that consumes much power. Simulation results shown that the transmission power per packet was higher than that of minimum energy algorithm.

Figure 4 below shows that **MRPC** algorithm would take path *P1* (A -- C -- F -- H) because it would send 3 packets from A to H while it would send only 2 packets through *P2* (A –B – E -- H) despite the fact that sending a packet through *P1* (6 units) consumes much more power than *P2* (only 3 units). We proposed a new algorithm called **NEER** (Normalized Energy Efficient Routing).

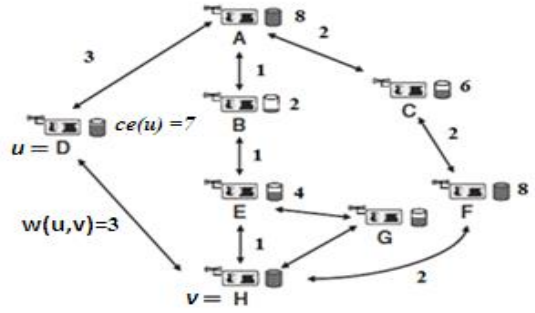


Figure 3: Graph G and its components

Our algorithm could be summarized as following:

- *Let G represent sensor network graph*
- *u,v represents nodes.*
- *Edge (u,v) is the link between u and v*
- *ce(u) :residual battery of node u*
- *w(u,v) is the weighted cost of edge(u,v)*
- *c(u,v) is the total number of packets that could be sent from u to v. this value is defined as ce(u)/w(u,v).*

Step 1: [Initialize]

Eliminate from G every edge *(u, v)* for which
*ce(u) < w(u, v).*this condition is used to ensure we could send at least one packet through this path.
For every remaining *edge (u, v)* let
c(u, v) = ce(u)/w(u, v).

Let *L* be the list of distinct *c(u, v)* values.

Step 2: [Binary Search]

Do a binary search in *L* to find the *maximum* value max for which there is a path *P* from Source to destination that uses no edge with
c(u, v) < max.
For this, when testing a value *q* from *L*, we perform a depth- or breadth-first search beginning at the source. The search is not permitted to use edges with c*(u, v) < q*.
Let *P* be the source-to-destination path with lifetime max.

Simultaneously we should find minimum energy path using Dijkstra's algorithm as following:

$$x = \sum_{i=1}^{i=n} w(u,v) \forall\, w(u,v) \epsilon\, P$$

Step 3: [Wrap Up]

If no path is found in Step 2, the route isn't possible. Otherwise, use P for the route.
Also find *Min*(x) , $\forall$ x ϵ *P*
Our new algorithm **power aware routing** needs to use a new hybrid algorithm that takes the advantages of both. Here we use the following equation

$$Z = \mu * (x) + (1 - \mu)\,(y)$$

Where μ, *(1- μ) are* the weight of that of the factors. (μ<*1*).

X: Minimum energy value

Y: P candidate value

If *Z* was above a certain value (threshold) we would use MRPC, if *Z* was less than that threshold, then we use minimum energy approach. In figure 5 below we could see our proposed algorithm flow chart that explain our algorithm steps in details.

In this case we took two factors in consideration. The total power consumed through that path and the residual battery in all nodes of that path. But we should note that we use weight in our new algorithm. The higher weight is for minimum energy factor. In such case we guarantee that we use minimum energy algorithm as long as possible but not to power off these nodes.

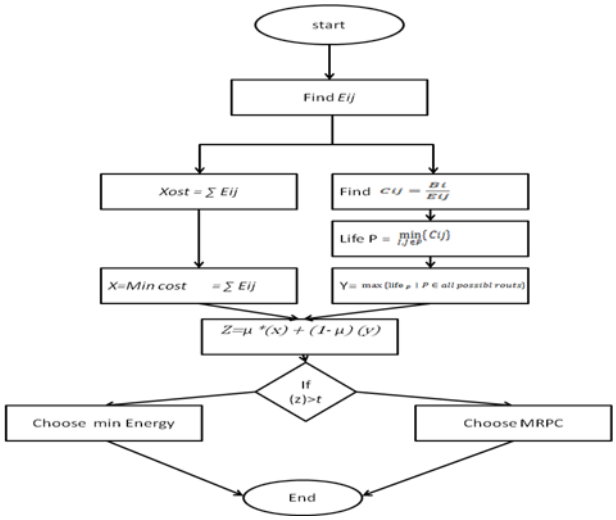


Figure 4: NEER flow chart

5. ANALYSIS:

There are some parameters used in simulation that could be summarized as following:

Table 1: Simulation Parameters

Parameter	Description
Channel type	Wireless channel
Mac protocol	Mac/802_11
number of nodes	40
routing protocol	Proposed Algorithm, MRPC, ME
grid size	800 X 800
packet size	64
simulation time	To die
Topology	Random , Flat
Initial energy	7 joule
Source node	7
Destination node	7

Here we would study the behavior of **NEER** algorithm in comparison with MRPC and Min-Energy Protocol in three fields. First factor is the total number of Dead nodes according to time. In this factor we expect that nodes are died slowly at the beginning of running for **NEER** algorithm and would die suddenly at the end of execution, since it takes the features of both of (MRPC and Min-Energy). The expected behavior of **NEER** algorithm is shown in figure (4).

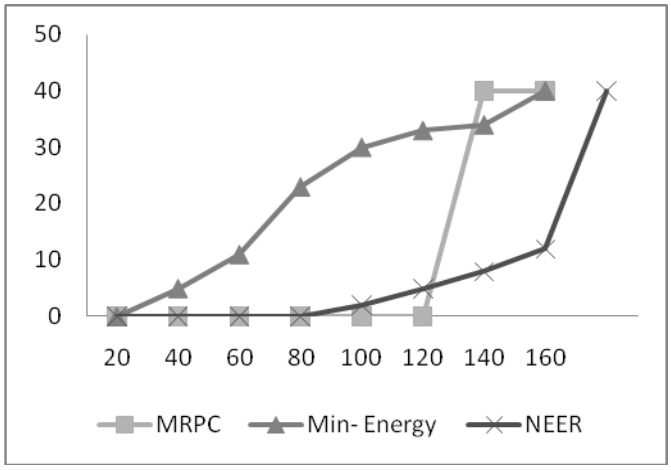


Figure 5: Expiration sequence

If we want to compare between these algorithms according to total sent packets by the network nodes, we expect that our algorithm would send packets more than Min-energy and less than MRPC as shown in Figure (5).

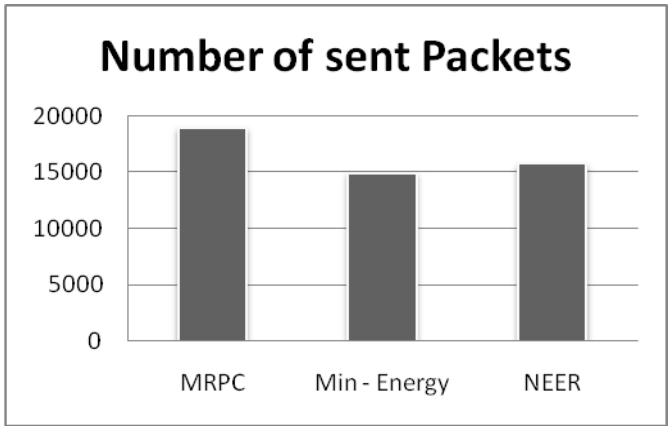


Figure 6: Number of sent Packets

Finally we would study energy per packet; here we also expect that energy per packet would be also between

MRPC and Min-Energy, more than Min-Energy, less than MRPC. Figure (6) explains the idea.

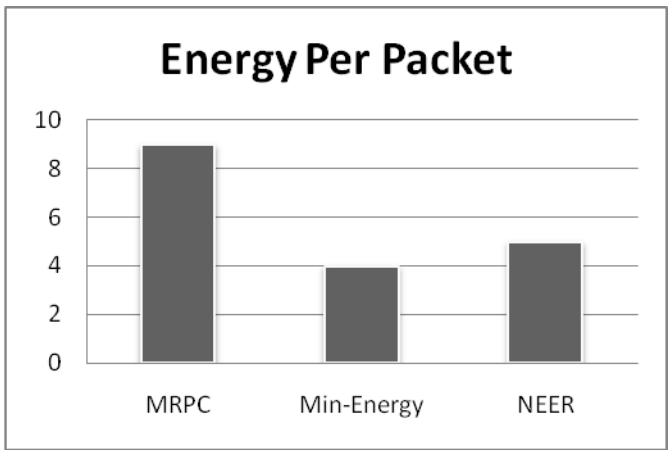


Figure 7: Energy per Packet

6. CONCLUSION

In this paper we aim to introduce a new hybrid algorithm that could increase network life time as long as possible using minimum energy and residual battery. In this paper our algorithm depends mainly on power consumption algorithms that are used in WSN. This algorithm is based in the fact that most power is consumed during transmission not during computations. This algorithm in fact takes advantages of two important protocols. It takes the advantage of consuming least power through minimum energy protocol. On the other hand the presence of MRPC protocol would increase network life time as long as possible. In this way, the proposed algorithm would use minimum energy protocol as long as the residual power is over a known threshold.

7. REFERENCES

[1] Guillermo Rodriguez-Navas, Miquel A. Ribot, Bartomeu Alorda, Understanding the Role of Transmission Power in Component-Based Architectures for Adaptive WSN, *Proceedings of the 2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*, July 2012.

[2] Fuu-Cheng Jiang, Der-Chen Huang, Chao-Tung Yang, Fang-Yi Leu , Lifetime elongation for wireless sensor network using queue-based approaches , March 2012.

[3] I. Stojmenovic and X. Lin. Loop-Free Hybrid Single-path/Flooding Routing Algorithms with Guaranteed Delivery for Wireless Networks. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1023–1032, 2001.

[4] X. Chen and J. Wu. Chapter Multicasting Techniques in Mobile Ad Hoc Networks. *The Handbook of Ad Hoc Wireless Networks*, pages 2-1–2-16. CRC Press, 2003.

[5] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides. On the Construction of Energy-Efficient Broadcast and Multicast Trees in Wireless Networks. In *Proceedings of IEEE Infocom*, Tel-Aviv, Israel, March 2000.

[6] E. M. Royer and C.-K. Toh. A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks. *IEEE Prersonal Communications*, 6(2): 46–55, 1999.

[7] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Mi-crosensor Networks," Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS '00), January 2000.

[8] W. Heinzelman, J. Kulik, H. Balakrishnan, Adaptive protocols for information dissemination in wireless sensor networks, in: Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom99), Seattle, WA, August 1999.

[9] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless sensor networks, in: Proceeding of the Hawaii International Conference System Sciences, Hawaii, January 2000.

[10] K. Scott and N. Bambos. Routing and Channel Assignment for Low Power Transmission in PCS. In Proceedings International Conference on Universal Personal Communications, pages 469–502, Cambridge, MA, September 1996.

[11] C. K. Toh. Maximum Battery Life Routing to Support Ubiquitous Mobile Computing in Wireless Ad Hoc Networks. IEEE Communications Magazine, 39: 138–147, 2001.

[12] I. F. Akyildiz, W. Su, Y. Sankasubramaniam, and E. Cayirci. Wireless Sensor Networks: A Survey. Computer Networks, 38: 393–422, 2002.

[13] S. Singh, M. Woo, and C. S. Raghavendra. Power-Aware Routing in Mobile Ad Hoc Networks. In *Proceedings of the 4th ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM'98)*, Dallas, TX, October 1998.

[14] Archan Misra and Suman Banerjee Maximizing Network Lifetime for Reliable Routing inWireless Environments, Department of Computer Science University of Maryland College Park, MD 20742, USA.2002.

[15] N. Bambos. Toward Power-Sensitive Network Architectures in Wireless Communications: Concepts, Issues, and Design Aspects. *IEEE Personal Communications*, 5: 50–59, 1998.

A Fuzzy Approach in Single Relay Selection for Cooperative Communication Based on Signal Strength and Residual Energy

Rafi Belal, Ken Ferens and Witold Kinsner
Electrical and Computer Engineering,
University of Manitoba, Winnipeg, MB, Canada
umbelal@cc.umanitoba.ca, Ken.Ferens|Witold.Kinsner@ad.umanitoba.ca

Abstract— Cooperative communication is a vital area of research in wireless sensor networks due to its advantages for spatial diversity. Specifically, selective single relay cooperative communication is an area of vast research interest due to its simplicity and practicality. In this paper, we propose a selective single relay cooperative communication scheme for wireless sensor networks which adapts power control and considers residual energy when electing potential relay nodes. Based on the received signal strength of the RTS/CTS messages, potential relay nodes compute the required transmission energy needed for data transmission from the source to itself and from itself to the destination. The required transmission energy from source to destination is also obtained through the MAC layer signaling. Node with the minimum transmission energy and highest residual energy is then elected as a relay for cooperative communication based on a fuzzy inference system. Our simulation results confirm that our scheme achieves significant energy savings in data transmission and enhances network lifetime.

Keywords— selective single relay routing, maximizing network lifetime, minimum transmission energy, received signal strength, residual energy, cooperative communication

I. INTRODUCTION

In recent years, wireless sensor networks have emerged as an important research area in the field of wireless communication for information acquisition, processing and transmission. In this technology, multiple independent nodes communicate with each other in a wireless medium to acquire, monitor and process data in various scenarios. In military research, wildlife monitoring, natural disaster warning system, etc. wireless sensor networks enables researchers an exclusive access for data acquisition and transmission. In typical ad hoc wireless networks, nodes communicate with each other without the need of any central base station and thus acquire and process information independently.

One of the main design challenges for wireless sensor networks is the finite energy of the nodes due to short battery life. Therefore, the minimization of transmission energy for these nodes poses a big challenge for a longer lifespan of the network.

Cooperative communication achieves the spatial diversity of a traditional antenna array due to its ability to share and transmit information jointly between multiple nodes [1], [2]. It has been an interesting research area for a number of researchers and the diversity performance and outage behavior of cooperative communication with different protocols have been studied [1], [3]. Distributed coding strategies have been applied and its feasibility studied for cooperative communication protocol in [4], [5].

Cooperative communication for cluster based networks has been studied in [6] and it has been shown that it greatly reduces the overall energy consumption among nodes and enhances network lifetime. Even in multi-hop wireless sensor networks, cooperative communication achieves significant energy consumption and reduces delay by enabling a cross layer approach [7]. A bandwidth-power trade-off analysis was conducted to analyze the energy efficiency for a low power low spectral environment [8] in [9] by the authors and it was proven that a simple decode and forward relay strategy achieves significant energy savings compared to the direct transmission.

In [10], the authors show that when the transmission distance between two nodes is large, collaboration among nodes in transmission and receive presents significant energy savings in a clustered network. This was done considering all the overheads for collaboration among the nodes. Cooperation of different number of nodes with distributed space time coding for inter-cluster communication was

investigated in [11] and it was shown that as the cooperation between nodes increased, the performance was degraded. This is because as the nodes cooperate more, the cooperation overhead also increases and increases the energy cost.

However, power control at the transmitting nodes can greatly enhance the energy savings of a cooperative communication scheme based on single relay cooperation. Minimization of link outage probability was achieved by [12] by employing optimal energy distribution among the cooperating nodes. A power allocation scheme for a simple decode and forward cooperation protocol, was also analyzed based on the symbol error rate (SER) analysis. However, in this case, the source and relays need the channel state information on all links.

In [13], the authors compute the energy efficiency for a selective single relay based cooperative communication. A novel scheme was proposed which jointly considers the MAC layer design and the physical layer power control. This scheme selects a "best" relay from a set of potential relay nodes; this relay is used in the cooperative communication between the source and the destination. The authors derived power control solution corresponding to two policies: one is to minimize the overall energy consumption per packet and the other is to maximize the network lifetime. The network lifetime is defined as the network operation time until the first node completely drains out its energy. Their scheme achieves significant energy conservation and enhances the network lifetime considerably. However, this schemes employs channel state information for determining the link quality. In wireless sensor networks, computing the channel state information is not a trivial task. Because of the unpredictable and fast changing nature of the wireless medium, obtaining the channel state information is highly prone to errors. In general, sensing of the channel at the time of data reception is one way of obtaining the instantaneous channel quality. For a fast fading environment, statistical characterization of the channel is more feasible. In [13], the authors adopt two important assumptions among others. The fading channels between the two nodes are flat in frequency and remain constant during one data transmission. The second one is the reciprocal channel from node A to node B is the same as the channel from node B to node A. However, when the node senses the channel to obtain the average channel gain, it may be prone to sensing error and noise from the medium. Channel state information corrupted with noise or sensing error might produce wrong estimation of channel quality and degrade the system performance.

In this paper, we propose a selective single relay based cooperative communication scheme based on estimation of link quality by the received signal strength of messages. We employ power allocation for source and relays to minimize the overall energy consumption per packet. Moreover, we consider the signal strength and residual energy of each potential relay nodes to select the "best" relay for cooperative communication. A novel fuzzy inference system was investigated to select the relay. Our scheme finds the best relay with minimum energy consumption from the received signal strength of MAC layer RTS/CTS signaling, without the need of error prone channel state estimation. Additionally, it considers the residual energy of the potential relay nodes when electing the best relay to enhance network lifetime. Our simulations confirm that our novel scheme achieves significant performance gain when compared to the techniques employed in [13] with minimum computational complexity and cost.

This paper is organized as follows. Section 3 describes the system model, different phases of the protocol involving relay selection, data transmission and sleeping strategy. Section 4 provides the performance evaluation of the simulation results. Section 5 describes some of the limitations of this scheme and Section 6 concludes the paper.

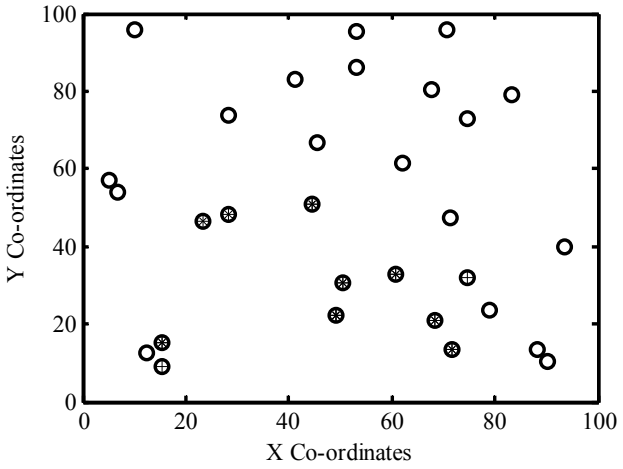


Fig. 1 Snapshot of the proposed network: * nodes are Potential Relay nodes and + nodes are source and destination nodes.

II. SYSTEM MODEL

As shown in Fig. 1, we deploy nodes in a decentralized fashion in our network. Nodes communicate with each other in an ad hoc fashion. We assume one hop transmission range in our network. All the nodes are within the maximum transmission range of each other. The instantaneous transmission power of the nodes can be adjusted tailored to the transmission distance. We assume flat fading in the network and the fading is constant during one data burst transmission. Moreover, reciprocal channel from node B to node A is the same as the channel from node A to node B.

Minimum Transmission Energy

Let us consider the following scenario. A node *t* transmits a frame with transmit power P_t. Node *r* receives this frame with power level P_r. It can be related to the transmit power as (1):

$$P_r = \mathrm{k} \times \frac{P_t}{D^\alpha} \quad (1)$$

We assume omnidirectional antenna and homogeneous receivers at all nodes. k is a proportionality constant and is assumed to be 1. α is the path attenuation coefficient. This is assumed to be 2 for short distance links. We employ the same principle as in [14], [15] where nodes choose the transmit power in a way so that the received power is above a certain threshold. When the received power minimally exceeds this threshold value, minimum energy consumption is achieved.

$$P_t = P_{th} \times D^\alpha \quad (2)$$

In search for the optimal transmission power level, we created a theoretical model for it from an information theoretic point of view. The study shows the relationship between optimal transmission power and link distance. In a slow fading channel model, the maximum rate of reliable data transmission supported by the channel is given by (3):

$$C = log_2\,(1 + |h|^2 \times SNR) \quad (3)$$

Here, *h* is the mean channel gain and SNR is the signal to noise ratio. Since the received signal strength is related to the transmission power as $P_r = \frac{P_t}{D^\alpha}$ and signal to noise ratio is given by $\frac{P_r}{\eta \times W}$, the relationship between the maximum possible rate of reliable data transfer and transmission power is given by (4):

$$C = log_2\,(1 + |h|^2 \times \frac{\frac{P_t}{D^\alpha}}{\eta \times W}) \quad (4)$$

We assume the noise variance to be $\eta = 1$ and $W = 1$ for simplicity. As the node is transmitting at a power level of P_t, energy needed to reliably transfer per bit is given by (5):

$$E_{P_t} = \frac{P_t}{log_2\,(1 + |h|^2 \times \frac{\frac{P_t}{D^\alpha}}{\eta \times W})} \quad (5)$$

It can be derived from (5) that the transmission energy needed to successfully transfer one bit in a slow fading environment is [16]:

$$E_{P_t} = \ln(2) \times \eta \times D^\alpha \quad (6)$$

To get minimum energy consumption given the link distance, the minimum transmission energy for successful data transfer is given by (7):

$$E_{min} \propto D^\alpha \quad (7)$$

This confirms that the minimum transmission energy for successfully transmitting data is directly proportional to link distance. If we assume that all the packets are of constant size then we can find from (2):

$$E_t \geq E_{th} \times D^\alpha \quad (8)$$

Relay Selection

When a node has data to send, it sends out an RTS message according to the IEEE 802.11 protocol with energy E_{max}. All the nodes hearing the RTS message wakes up from sleep (the sleep strategy will be discussed in section 3). The destination node receives this RTS and sends CTS accordingly. All nodes hear this CTS message. Based on the received signal strengths of both the RTS and CTS messages, the overhearing nodes estimate the link distances from the source and destination by equation (8) and is denoted by $d_{i,\,s}$ and $d_{i,\,d}$ (for node *i*). The destination node estimates the link distance between source and itself by the RTS message and sends this information with CTS. This is denoted by $d_{s,\,d}$. At this stage all the nodes know the link distance between itself and the source and destination and also have information about the link distance between the source and destination nodes.

Overhearing nodes compete with each other to be in the Potential Relay set. Potential Relay set is given by the nodes who satisfies:

$$d_{j,s} \leq d_{s,d}\ AND\ d_{j,d} \leq d_{s,d} \quad j \in N_{PotentialRelay} \quad (9)$$

If a node is in the potential relay set, it stays awake. Other nodes go back to sleep and do not wake up unless they hear another RTS message. Nodes who are in the $N_{PotentialRelay}$ set, compute the minimum transmission energy required for a cooperative data transmission from source itself which is denoted as $Energy_{s,i}$ per packet and from itself to the destination node which is denoted as $Energy_{i,d}$ per packet using equation (8). Minimum energy needed to transmit from source to destination directly is also computed which is denoted by $Energy_{s,d}$ per packet. The nodes belong to the $N_{RelaySet}$ if the following condition is met

$$Energy_{s,i} + Energy_{i,d} \leq Energy_{s,d} \quad (10)$$

Members of $N_{PotentialRelay}$ set who do not move onto the $N_{RelaySet}$ go to sleep and do not wake up until they hear the next RTS message.

Members of the $N_{RelaySet}$ now go through a fuzzy control system to determine a metric which is denoted by M_{Relay} to determine their eligibility of being elected as the final relay. The fuzzy control system is employed in all the member nodes of the $N_{RelaySet}$. The inputs of the control system are required transmission energy and residual energy of the node which are denoted by E_i^A and E_i^R correspondingly (for candidate node *i*). The output of the system is the unique metric M_{Relay} . We used the triangle membership function in this system. The inputs and outputs have three linguistic variables named *Low, Medium* and *High.* The fuzzy rules are defined in the table below.

Table 1: Fuzzy rules for sensor node selection.

		Residual Energy		
Transmission Energy		**Low**	**Medium**	**High**
	Low	Medium	High	High
	Medium	Low	Medium	High
	High	Low	Low	Medium

Each member node of the $N_{RelaySet}$ computes a unique output metric M_{Relay} to compete with each other. The node with the highest metric gets to be the chosen best relay. Nodes with their unique metric waits for a back-off time t_i and sends a beacon message with E_{Beacon} energy to the source notifying elected relay and $Energy_{s,i}$. The source node uses this energy when transmitting packets to the relay node. This back off time is related to the metric by:

$$t_i = M_{relay} \times Rand_t \quad (11)$$

Here, $Rand_t$ is a random number. The formulation of t_i with the random number ensures that no collisions occur if two or more nodes achieve the same metric. When the beacon message is heard by other member nodes of $N_{RelaySet}$, they discard the back-off time and go to sleep until they hear the next RTS message.

Data Transmission

The source node receives the beacon message from the best relay and knows the minimum transmission energy $Energy_{s,i}$ per symbol. It sends data packet with this energy to the relay node. The best relay node receives the packet and decodes the data.

The best relay decodes the data from the source node and forwards it to the destination node with minimum transmission energy $Energy_{i,d}$ per symbol. The destination node receives the data from the best relay and sends an ACK message to the source node with E_{max} . After a successful data transmission, the source, relay and destination go to sleep until they hear another RTS message.

Sleep Strategy

In order to improve network lifetime and mitigate idle listening, we employ low power listening in our cooperative communication scheme. We adopt an Ultra-low power RF wake up sensor proposed in [17]. Here, authors propose a new RF sensor which is a dedicated small RF module to check potential communication by sensing the presence of a RF signal. Effects of duty cycling, sleep delay and idle listening can be successfully mitigated by the employment of this RF wakeup sensor. While consuming only 1% energy of an idle node, this wake up sensor is able to turn off their communication module and perform carrier sensing. By employing this wakeup sensor in our scheme, nodes can detect a RTS signal which has higher signal strength than the predefined threshold while their communication module is turned off for sleeping. In this case, the RF wakeup sensor interrupts the processor to notify about the communication occurrence in the medium. This enables our scheme to achieve significant energy savings. Moreover, nodes can wake up on demand wherever there is an RTS message which enables zero sleep delay and zero idle listening.

III. PERFORMANCE EVALUATION

In this section we compare our scheme with [13] which also uses a selective single relay for cooperative communication. For fare comparison, we considered sleep delay and cost in the scheme that is described in [13]. Also, overhead for RTS/CTS signaling and relay competition were also considered for comparison.

Simulation Environment

We assume our network to be a random ad hoc network where nodes are distributed randomly in a rectangular area of 100x100 meters. Each round initiates a data flow from a source to destination which are chosen randomly. Because of the assumption of the one hop transmission range, all nodes can hear the RTS/CTS messages of each other in our network. The RTS/CTS signaling is done with maximum energy E_{max} . The channel condition remains constant for one data burst during one round of data transmission but changes as soon as the burst ends. The compared scheme [14] is denoted as EECC for presentation brevity in the performance measurement figures. The packet length is set to be 1000bits. Symbol duration, T_s is 10^{-4} s. Packet length in symbols is set to be packet length in bits (n_b)/ Data rate in symbols(R). The maximum transmission power is denoted as P_{max} . Maximum energy is given by $E_{max} = P_{max} \times T_s$. Minimum threshold energy for reliable data transfer is denoted as $E_{th} = 1 \times 10^{-8}$ J per symbol.

Overall Energy Consumption

In this case we vary the maximum transmission energy from 0 to 3.5×10^{-4} J per symbol and investigate the effect in overall energy consumption per packet in both the schemes. Number of participating nodes was 10. Symbol duration was set to be 10^{-4}s. Packet length in bits was set at 1000bits. Data rate per symbol was set to be 2. We conducted the simulation for 2000 rounds for both the schemes. Each node was given an initial energy of 5 Joules. E_{th} was set to be 1×10^{-8} J per symbol.

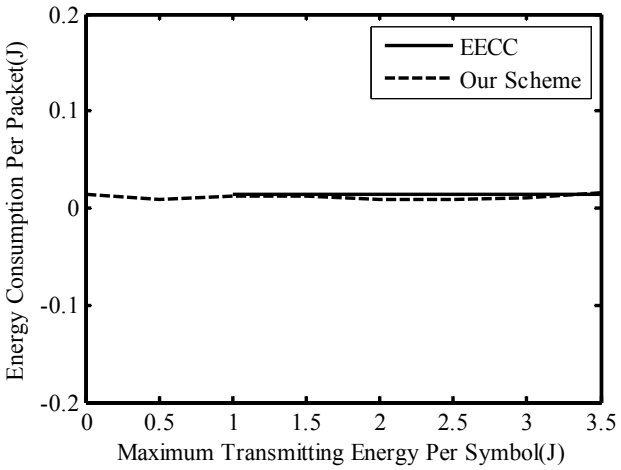


Fig. 2 Energy consumption per packet as a function of maximum transmission energy.

From Fig. 2, we can see that our scheme and the compared scheme achieve significant energy savings in overall energy consumption per packet even with the varying maximum transmission energy. As the maximum transmission energy is increased both the schemes performed well in minimizing the overall energy consumption per packet and give a steady value. It also proves that, both the schemes spend almost similar amount of energy for reliable data transmission per packet.

Network Lifetime with Varying Data Rate

In this case, we vary the data rate per symbol in both the schemes from 1 to 4 and investigate the effect in network lifetime. Maximum transmission energy was set to be 3.5×10^{-4}J per symbol. Total of 5000 rounds were conducted in this simulation.

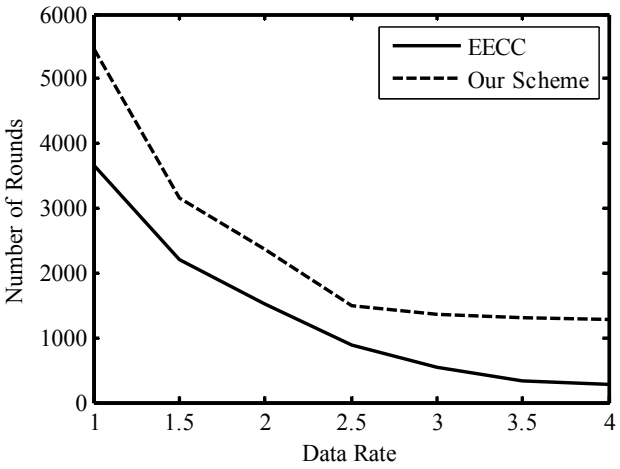


Fig. 3 Number of rounds with respect to the data rate per symbol R in a uniform traffic scenario.

We can see from Fig. 3, that as we increase the data rate, network lifetime decreases. This is because as the data rate is increased with constant packet length, more energy is spent to transmit that data. As a result of more energy consumption per packet, the nodes deplete more energy and the network lifetime decreases. However, our scheme performs significantly better than the compared scheme. This is partly due to the employment of fuzzy control system that also considers the residual energy of relay nodes. This ensures that nodes are considered of their residual energy as well as transmission energy when electing the best relay. This performance boost is also debited partly due to the employment of sleep schedule in our scheme. Nodes go to sleep when they are not receiving or transmitting any data. Moreover, nodes participating in the relay election also go to sleep when they are not selected. This ensures maximum energy savings and thus improves the network lifetime with increasing data rate.

Number of Packets Transmitted

In this case we vary the number of nodes in the network from 5 to 30 and investigate the effect on number of packets transmitted. The maximum transmission energy was set to be 1×10^{-5} J per symbol. Data rate was set to be 2 per symbol. Nodes were given an initial energy of 5 J.

From Fig. 4, we can see that with the increase of participating nodes in the network, the number of packets transmitted also increases for both the schemes. This is because as more nodes are in the network, there are more opportunities for cooperative communication. Also, as the number of nodes increases, the density of nodes increases which makes it more energy efficient for one hop cooperative communication. The transmission energy needed decreases as there are more dense nodes in the

network for cooperative communication. We can also see that our scheme achieves superior performance against the compared scheme. This is also because of jointly considering the residual energy and transmission energy needed when electing the relay nodes. With similar amount of energy consumed per packet, our scheme performs better in all the cases of varying number of nodes.

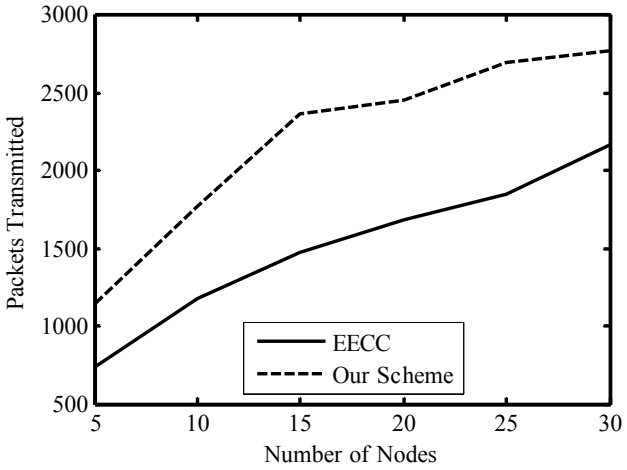


Fig. 4 Total packets transmitted as a function of number of nodes employed in the network.

IV. LIMITATIONS

Our scheme performs significantly well compared to the scheme in [13]. But it has some limitations. Our scheme uses the received signal strength to measure the minimum transmission energy needed to successfully transmit the data. In [16], authors argue that it is not always possible to calculate the optimal energy cost over a link with this principle. The link error rate has a significant effect on choosing transmission energy higher than a threshold value. The authors show that the link error rate decreases as the transmission energy is increased. Moreover, our scheme does not consider the retransmission cost for reliable packet delivery on individual links. For reliably delivering a frame, it may be needed to be delivered multiple times for successful receipt by the destination node.

V. CONCLUSION

In this paper, we propose a novel received signal strength and residual energy based single relay cooperative communication protocol for wireless sensor networks. Our protocol elects the best relay with minimum transmission energy cost and highest residual energy. Minimum transmission energy is estimated by received signal strength. Simulation results confirm that our scheme achieves significant energy savings and enhances network lifetime when compared to similar schemes.

REFERENCES

[1] J. Nicholas Laneman, David N. C. Tse, and Gregory W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," in *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062-3080, December 2004

[2] J. Nicholas Laneman, and Gregory W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," in *Proceedings of Global Telecommunications Conference,* vol. 1, pp. 77-81, November 2002

[3] Rohit U. Nabar, Helmut Bolcskei, and Felix W. Kneubuhler, "Fading relay channels: performance limits and space-time signal design," in *IEEE Journal on selected areas in Communications*, vol. 22, no. 6, pp. 1099-1109, August 2004

[4] J. Nicholas Laneman, and Gregory W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," in *Proceedings of Global Telecommunications Conference,* vol. 1, pp. 77-81, November 2002

[5] Andrej Stefanov, and Elza Erkip, "Cooperative coding for wireless networks," in *IEEE Transactions on Communications*, vol. 52, no. 9, pp. 1470-1476, September 2004

[6] Shuguang Cui, Andrea J. Goldsmith, and Ahmad Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," in *IEEE Journal on Selected Areas in Communications,* vol. 22, no. 6, pp. 1089-1098, August 2004

[7] Shuguang Cui, and Andrea J. Goldsmith, "Cross-layer optimization of sensor networks based on cooperative MIMO techniques with rate adaptation," in *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications,* pp. 960-964, June 2005

[8] Yingwei Yao, Xiaodong Cai, and Georgios B. Giannakis, "On energy efficiency and optimum resource allocation of relay transmissions in the low-power regime," in *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2917-2927, November 2005

[9] Sergio Verdú, "Spectral efficiency in the wideband regime," in *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002

[10] Shuguang Cui, Andrea J. Goldsmith, and Ahmad

Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," in *IEEE Journal on Selected Areas in Communications,* vol. 22, no. 6, pp. 1089-1098, August 2004

[11] Zhong Zhou, Shengli Zhou, Shuguang Cui, and Jun-Hong Cui, "Energy-efficient cooperative communication in clustered wireless sensor networks," in *Proceedings of Military Communications Conference*, pp. 1–7, October 2006

[12] Mazen O. Hasna, and Mohamed-Slim Alouini, "Optimal power allocation for relayed transmissions over Rayleigh fading channels," in *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1999–2004, November 2004

[13] Zhong Zhou, Shengli Zhou, Jun-Hong Cui, and Shuguang Cui, "Energy-efficient cooperative communication based on power control and selective single-Relay in wireless sensor networks," in *IEEE Transactions on Wireless Communications.* vol. 7, no. 8, August 2008

[14] Suresh Singh, and C.S. Raghavendra, "Pamas-Power aware multi-access protocol with signaling for ad hoc networks," in *ACM Communications Review,* July 1998

[15] J. Gomez-Castellanos, A. Campbell, M. Naghshineh, and C. Bisdikian, "PARO: A power-aware routing optimization scheme for mobile ad hoc networks," draft-gomez-paro-manet-00.txt, work in progress, *IETF*, March 2001

[16] Suman Banarjee, and Archan Misra, "Adapting transmission power for optimal energy reliable multihop wireless communication," in *Proceedings of International Symposium of Mobile Ad Hoc Networking and Computing,* June 2002, EPFL Lausanne, Switzerland

[17] Yong Soo Bae, Sang Hoon Lee, Lynn Choi, "The design of a ultra-low power RF wakeup sensor for wireless sensor networks," in *Proceedings of 18th Asia-Pacific Conference on Communications,* pp. 214-218, October 2012

SESSION

NOVEL COMMUNICATION SYSTEMS + ALGORITHMS + TOOLS + TRANSMISSION SYSTEMS + PROTOCOLS AND ROUTING

Chair(s)

TBA

Neural Networks for Admission Control in Wireless Mesh Networks

J. Mendoza[1]
[1]Electronic Department, Universidad Católica de Oriente, Rionegro, Antioquia, Colombia

Abstract— *This paper presents the results of a research concerning admission control (AC) on Wireless Mesh Networks (WMNs). At first, there is an introduction to WMNs and its problems. Later, topics as quality of service (QoS) and admission control en WMNs are presented, alluding to neural networks as an alternative to achieve the admission control in communication networks. Next, the results of different conduced test using a network simulator are presented to deem admission control under diverse scenarios in the same test topology, in order to quantify throughput of the network traffic. Lastly, both results are collate, those from using the traditional techniques and those with neural networks, based on the results of this work, the conclusions are drawn.*

Keywords: wireless mesh networks; admission control; neural networks; quality of service; throughput; network simulator.

1. Introduction

WMNs (WMNs) are multi-hop networks "Fig. 1", that can be defined as "an autonomous and spontaneous set of mobile routers connected by wireless links that do not require a fixed infrastructure" [1].

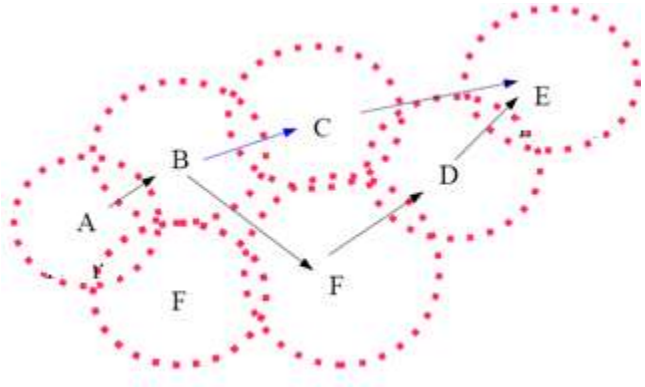


Fig. 1. Wireless mesh networks (WMNs)

On the one hand, a wireless network is made up of one or more Access points (Aps) plugged by an UTP cable to a wired network thus, be connected to more Aps in order to increase the coverage. On the other hand, WMNs allow a wireless communication among the different Aps "Fig. 2"; these are simple networks bearing in mind that all the APs share the same frequency channels which turn important when streaming the information from a point to another [2]

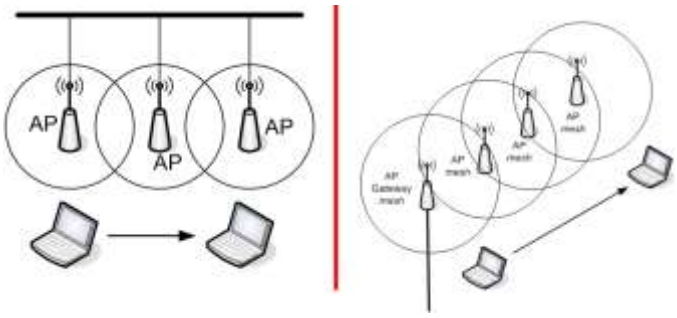


Fig. 2. A comparison between traditional networks and WMNs

Among the principal characteristics of the WMNs [3], there is the presence of redundant links and the capacity to react to changes of topology (dynamic topology).

Despite all the advantages of the WMNs, they are still networks in progress, therefore, they do exist prominent research possibilities on searching standards and protocols for the purpose of solve the existing problems among the different layers of its architecture.

There is a detailed description on [2], [4] and [5], about the WMNs performance on each layer of the Open Systems Interconnection (OSI) model, underlying those research challenges for each layer.

On the whole, for the WMNs, QoS's guarantee is a determining factor for the viability of its implementation. QoS's schemes are intended to optimize the actual capacity of the network (which is the principal limitation of the wireless networks) by: The application of admission control techniques; bearing in mind that the intention of a admission control system is to determine if the available resources in the network can allow new traffic without affecting the QoS of the existing traffic; a fair assignment of the existing resources.

2. Quality of service, admission control and neural networks

2.1 Proposal Offer to improve QoS in WMNs

Next, there is a count of those existing proposals designed to better QoS in WMNs. "Fig. 3" shows a summary of the state of the art that will help to understand the sequence in which they are going to present some findings.

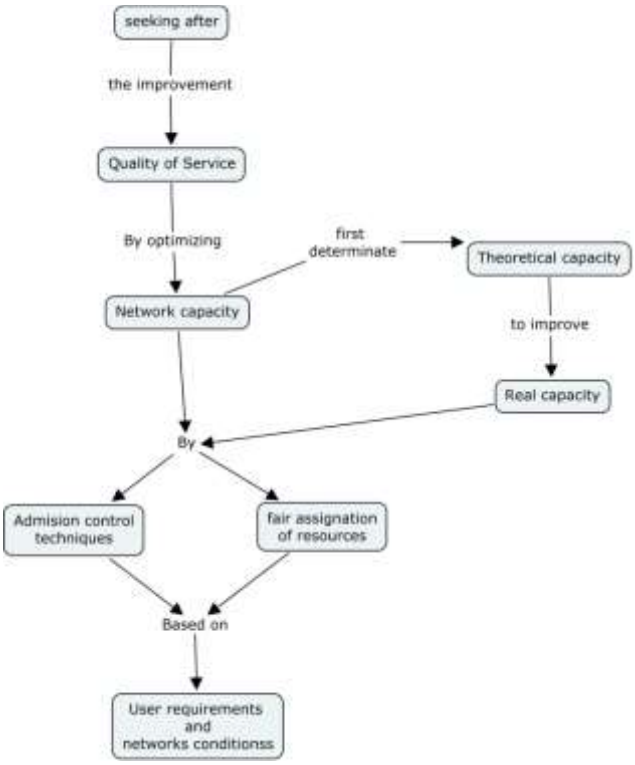


Fig. 3. Conceptual map, state of the art.

In [6], inside the formulation of the issue and before determining the WMNs' theoretical capacity, emphasis is placed on the concept of impartiality of multi-hop networks, the capacity MAC and the collision domains (physical segment where two or more nodes attempt to send a signal along the same channel at the same time) are defined as essential blocks to do the estimation. The authors demonstrate that for the performance of the WMNs decreases in the form f (1/n), where n is the total number of nodes in the network.

Due to the need of supporting diverse types of traffic on the WMNs, in [7] it is proposed a scheme to prioritize traffics, through the addition of QoS's classification methods. This becomes a requisite, since the WMNs process real-time traffic as: Voice IP and videoconferences, which present QoS's high requirements with regard to delay jitter; and non-real-time traffic as: a FTP and web browsers, in both cases, the QoS is more related with the reduction of data loss.

Admission control is used in [8], to propose a methodology to estimate the available bandwidth in the MAC layer, by using an admission control algorithm (ACA) to support QoS in WMNs and to address the traffic in accordance with the requirements. When it is real-time traffic, all the nodes in a route enable traffic based on the estimated bandwidth. In a different way, for non-real-time traffic, these procedures seek to fit the throughput so as to avoid a jammed network. Lastly, after several tests the efficiency of the algorithm is demonstrated.

To obtain the results in [8], several activities are realized under ns-2. The first one is to predict the available bandwidth. The second one is to locate the admission control. On third place, several traffics must be differentiating. Finally, to support mobility. All these lead to the conclusion that using admission control algorithms there is improvement in QoS. As a consequence to optimize admission control systems becomes a core tasks in the project.

2.2 Admission control in WMNs

The primary medium access mechanism underpinning multi-hop wireless networks is the IEEE 802.11 protocol:

- ✓ Low cost
- ✓ Easy setup
- ✓ high physical data rates (>54Mbps)

The most important feature of such IEEE 802.11-based mesh networks is that the radio links share the radio resources using a carrier sense multiple access (CSMA) based on random access protocol In CSMA for WMNs, the carrier sensing operation must now cope with the following two forms of asymmetry:

- ➢ Contention asymmetry: This introduces asymmetry in the level of contention each link/node experiences.
- ➢ Traffic asymmetry: The rate at which a link-i contends for the radio channel is a direct function of the traffic, i it needs to carry

CSMA in WMNs could present two types of problems, "Fig. 4":

- ➢ Hidden terminal problem
- ➢ Expose terminal problem

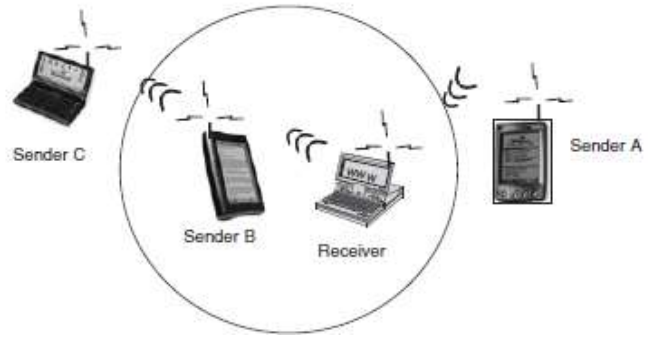


Fig. 4. Hidden terminal and exposed terminal.

To avoid the hidden and exposed terminal problem, the IEEE 802.11 basic CSMA access mechanism is extended with a virtual carrier sensing mechanism "Fig. 5".

request-to-send (**RTS**) / *clear-to-send* (**CTS**)

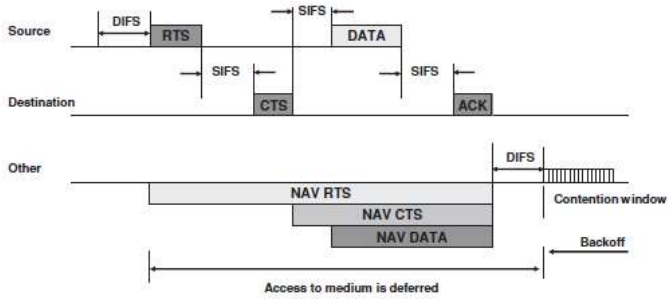


Fig. 5. RTS and CTS

2.3 Computational intelligence for admision control

In traditional admission control algorithms, there are used mathematical models; these models need to suppose traffic processes for a simple solution. This is a complex task on big networks because analytical and theoretical solutions only are valid by models. Then, QoS estimations are uncertain and admission control decisions are ineffective. In [9]-[11], three papers present the advantages of the computational intelligence methods to admission control applied in telecommunication networks.

In [10], the paper shows call admission control (CAC) problems and how the neural networks (NNs) would be a key component for CAC algorithm. NNs forecast QoS conditions and the CAC algorithm uses the NNs results to make decisions.

NNs are suitable solutions to CAC problems [10] because they reach appropriate non-lineal functions estimations. Moreover, there are certain advantages to be named: They do not need neither a precise mathematical model for traffic, nor making suppositions because NNs already know how to react based on real data behaviors; when NNs have abstracted valuable information, they are able to predict rough estimations from unknown data.

The CAC algorithm makes decisions for new traffic in the network by using the QoS estimations that the NNs do. The results show a better throughput when NNs are used for admission control in traditional networks; certainly, using NNs for admission controls in WMNs is now a relevant topic.

3. Testing and results

WMNs tests were carried out with ns-3 simulator [12]-[14] as *scilab* [15]-[17] was use for neural networks implementation.

3.1 WMNs in ns3

Ns-3 code is written in C++, for this reason, the next objects are necessary for any simulation:

- Nodes: class *node.*
- Networks devices: class *NetDivice*
- Channels: class *channel*
- Packets: *Packets* object
- Applications:
    - class *Application*
    - Defined by users

Loss propagation model:

- *Log distance loss propagation model:* this model calculates the reception power.
- *Random loss propagation model*: this model is used to introduce variance to the signal strength.
- *Delay propagation model:* this model is added due to interferences of the air.

Objects for WMNs en ns-3:

- *MeshHelper:* 802.11s implementation.
- Nodes:
    - *NodeContainer:* 8 nodes.
    - *NetDiviceContainer:* Wifi card.
    - *MobilityHelper:* node position.
- Devices:
    - *WifiMacHelper*: for routing
    - *WifiChanelHelper:* propagation model
    - *WifiPhyHelper:* physics layer dates.
    - Interfaces:
        - *InternetStackHelper:* Internet
        - *Ipv4AdressHelper:* IP address

- *Ipv4StaticRoutingHelper:* static routing

- Applications:
  - *PacketSinkHelper*: to send TCP traffic.

The WMN for these tests have 8 nodes, those nodes are distributed in a row, "Fig. 6"; the distances between two consecutive nodes are variable, "TABLE I".

Fig. 6. WMN simulated.

TABLE. I. DISTANCE AMONG NODES.

Nodes	Distances
Node 1 – node 2	37,0 m
Node 2 - node 3	36,5 m
Node 3 – node 4	36,5 m
Node 4 – node 5	33,8 m
Node 5 – node 6	39,2 m
Node 6 – node 7	25,5 m
Node 7 – node 8	44,0 m

For simulation is important to define some parameters, "TABLE II".

TABLE. II. CARACTERISTICS.

Parameter	Value
For routing	Static
Data transfer	11 Mb/s
Signal power	18 dBm
Channel frequency	2,4 x 10^6 Hz

The experimental design proposes to do the tests in 6 scenarios, under the same initial conditions; the only difference in each one of the scenarios is the number of repetitions, "TABLE III".

TABLE. III. SCENARIOS.

Scenario	Number of repetitions
1	42
2	36
3	30
4	24
5	18
6	12

The first tests were carried out for transmission between two neighbor nodes (one hop between nodes). "Fig. 7"

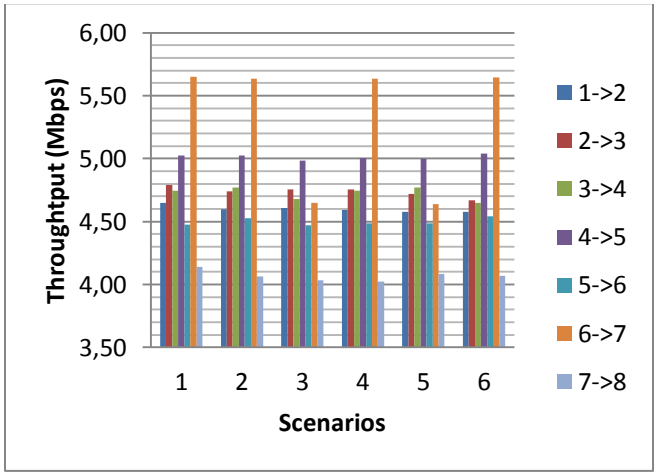


Fig. 7. Result test: One hop between nodes.

The second tests were carried out for transmission between two no neighbor nodes (two hops between nodes). "Fig. 8"

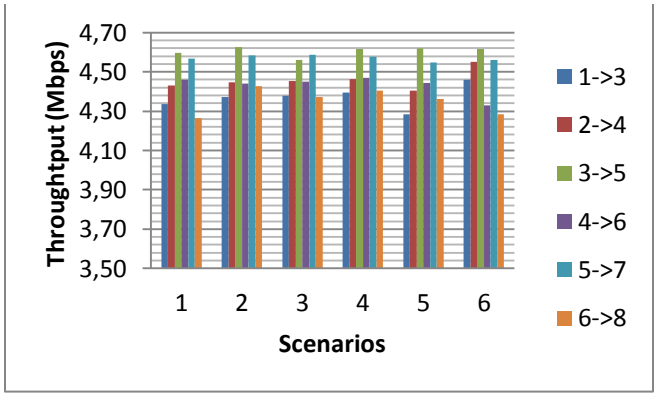


Fig. 8. Result test: Two hops between nodes.

The third tests were carried out for transmission between two no neighbor nodes (three hops between nodes). "Fig. 9"

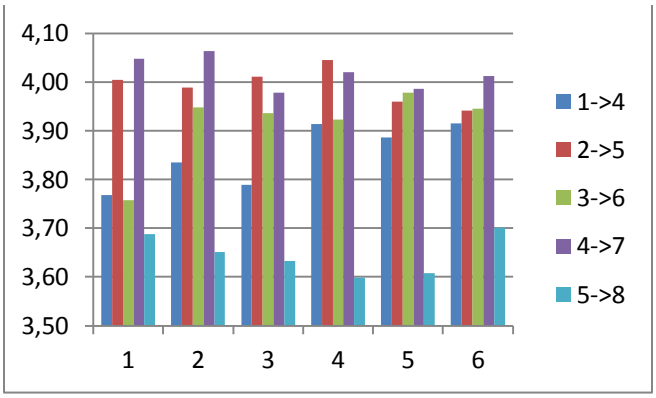


Fig. 9. Result test: Three hops between nodes.

For the last three test, if the number of hops between nodes increase, the throughput decreases. This is a logic result and presents similar outcomes in all scenarios.

For the same scenario, the throughput depends on the distance between nodes.

3.2 Admission control with neural networks

For admission control using neural networks, is desirable for the neural network inputs to have the following properties:

- Capture key elements of traffic behavior that influence the queue.
- Support a large number of traffic classes.
- Keep the number of inputs reasonably small.

As a consequence the NNs inputs are:

1. Number of Calls per Traffic Class
2. Counts of Arrivals

The values for the neural networks inputs are sent from *ns-3* to *scilab*

The neural network outputs is a decision, this decision is accept or reject traffic. The NN has learnt the boundary between the feasible and infeasible performance regions for a given input space.

The neural network output (to accept or to reject traffic) is sent from *scilab* to ns-*3*.

This test was carried out in number-3 scenario, "Fig. 10",

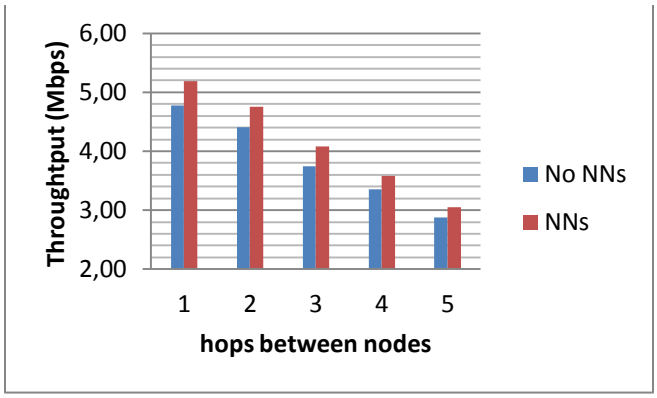


Fig. 10. Admission control: Neural network vs. no neural network

The throughput increases when neural networks are using for admission control.

3.3 Future work

Similar test measuring:

- delay
- jitter

3.4 Acknowledgment

- To Universidad de Antioquia for hosting the research.
- To Luis D. Méndez for his valuable comments and sharing his knowledge in the edition of this paper.

4. Conclusions

WMNs are an outstanding wireless technology. For this reason, this topic is full of scientific and technological interest, with many possibilities in the investigation field.

Admission control is important to improve QoS in WMNs, for this reason, if admission control shows betterment, then the QoS would be better than before.

When neural networks are used for admission control in WMNs, the throughput increases; but a QoS improvement can be warranted only when similar experiments are being carried out measuring other performance characteristics, for example, delay and jitter.

5. References

[1] D. Acuña, R. R Roncallo, (2007) "Redes Inalámbricas Enmalladas Metropolitanas". Monografía (Ingeniería Electrónica). Universidad Tecnológica de Bolívar.

[2] George Aggélou, "Wireless Mesh Networking", McGraw-Hill, 2008.

[3] Jangeun Jun and Mihail L. Sichitiu, "The Nominal Capacity of Wireless Mesh Networks". IEEE Wireless Communications • October 2003

[4] Ian F. Akyildiz, Xudong Wang, Weilin Wang, "Wireless mesh networks: a survey", Computer Networks 47 (2005) 445–487.

[5] V.C. Gungor, E. Natalizio, P. Pace and S. Avallone, "Challenges and Issues in Designing Architectures and Protocols for Wireless Mesh Networks", Part 1: Wireless Mesh Networks Architectures and Protocols, Springer, 2007.

[6] Jangeun Jun and Mihail L. Sichitiu, "The Nominal Capacity of Wireless Mesh Networks", IEEE Wireless Communications • October 2003

[7] Honglin Hu, Yan Zhang, Hsiao-Hwa Chen. "An Effective QoS Differentiation Scheme for Wireless Mesh Networks", IEEE Network • January/February 2008.

[8] Qiang Shen, Pan Li. "Admission Control Based on Available Bandwidth Estimation for Wireless Mesh Networks". IEEE transactions on vehicular technology, vol.58, June 2009.

[9] Witold Pedrycz and Athanasios V. Vasilakos, "Computational Intelligence: A Development Environment for Telecommunications Networks", Part 1 of: Computational Intelligence in Telecommunications Networks, CRC Press, 2001.

[10] Richard G. Ogier and Nina Taft-Plotkin, "Neural Network Methods for Call Admission Control", Part 2 of: Computational Intelligence in Telecommunications Networks, CRC Press, 2001.

[11] Ray-Guang Cheng and Chung-Ju Chang, "CAC and Computational Intelligence", Part 3 of: Computational Intelligence in Telecommunications Networks, CRC Press, 2001.

[12] NS-3 PROJECT. ns-3 Tutorial [online]. http://www.nsnam.org/docs/release/3.16/tutorial/ns-3-tutorial.pdf [cited on 17 march 2013]

[13] NS-3 PROJECT. ns-3 Manual [online]. http://www.nsnam.org/docs/release/3.16/manual/ns-3-manual.pdf [cited on 17 march 2013]

[14] NS-3 PROJECT. ns-3 Model Library [online]. http://www.nsnam.org/docs/release/3.16/models/ns-3-model-library.pdf [cited on 17 march 2013]

[15] SCILAB PROJECT. Introduction to scilab [online]. http://www.scilab.org/resources/documentation/tutorials [cited on 17 march 2013]

[16] SCILAB PROJECT. Scilab for very beginners [online]. http://www.scilab.org/resources/documentation/tutorials [cited on 17 march 2013]

[17] R.M Hristev, "The ANN Book", GPL license. http://www.gsd.ece.buap.mx/DocumentosSistemasDigitales/NeuralNetworks/Hritsev_The_ANN_Book.pdf [cited on 17 march 2013]

[18] Qiang Shen, Pan Li. "Admission Control Based on Available Bandwidth Estimation for Wireless Mesh Networks". IEEE transactions on vehicular technology, vol.58, June 2009.

An Unambiguous Correlation Function With a Sharp Main-Peak for BOC Signal Tracking

Youngseok Lee, Jeongyoon Shim, Youngpo Lee, Jaewoo Lee, and Seokho Yoon†
College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Gyeonggi-do, Korea
†Corresponding author

Abstract— *This paper proposes an unambiguous binary offset carrier (BOC) correlation function based on a combination of partial correlations composing the BOC autocorrelation, which has a main-peak sharper than those of the conventional unambiguous correlation functions, thus providing a better tracking performance. From numerical results, it is confirmed that the proposed unambiguous correlation function provides a significant tracking performance improvement over the conventional unambiguous correlation functions in terms of the tracking error standard deviation.*

Keywords: BOC; correlation function; ambiguity problem; tracking; TESD

1. Introduction

New global navigation satellite systems (GNSSs) including European Galileo and modernized global positioning system (GPS) employ the binary offset carrier (BOC) modulation providing a higher positioning accuracy than the conventional phase shift keying (PSK) modulation used in the conventional GPS [1], [2]. In the BOC modulation, the signal is generated by multiplying a pseudorandom noise (PRN) code with a sub-carrier of sine- or cosine-phased square wave, denoted as $\text{BOC}_{\text{sin}}(kn, n)$ or $\text{BOC}_{\text{cos}}(kn, n)$, respectively: Here, k represents the ratio of the chip period T_c of the PRN code to the period of the sub-carrier, and n denotes the ratio of T_c^{-1} to 1.023 MHz [3], [4]. In general, the BOC signal is known to provide a better tracking performance than the GPS signal. The performance improvement on the signal tracking, and consequently, the positioning accuracy comes from the fact that the main-peak of the BOC autocorrelation is much narrower than that of the PSK autocorrelation [5]. However, the main drawback of the BOC modulated signal is that its autocorrelation has multiple side-peaks, making the signal tracked at one of the side-peaks, and thus, resulting in the biased tracking measurements, which is referred to as the ambiguity problem [2].

To deal with the problem, several unambiguous correlation functions [6]-[8] have been proposed removing side-peaks directly. In [6], subtraction of the crosscorrelation between the BOC and PRN signals from the BOC autocorrelation is employed, on the other hand, in [7], an unambiguous correlation function was proposed based on a combination of the crosscorrelations between the received BOC and specially designed local signals; however, the correlation functions in [7] and [8] are applicable only to $\text{BOC}_{\text{sin}}(n, n)$ and $\text{BOC}_{\text{sin}}(kn, n)$ signals, respectively. Recently, in [8], an interesting correlation function applicable to both $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$ was proposed by combining the sub-correlations of the BOC autocorrelation. However, the correlation function in [8] is focused only on removing side-peaks of the autocorrelation without an effort to improve its tracking performance.

In this paper, we propose an unambiguous correlation function with a improved tracking performance by splitting a sub-carrier pulse into two sub-pulses and combining the partial correlations obtained using the sub-pulses. The proposed correlation function has a narrower main-peak than those of the conventional correlation functions and is applicable to both $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$. Moreover, it is demonstrated that the proposed correlation function offers a performance improvement over the conventional correlation functions in terms of the tracking error standard deviation (TESD).

The rest of this paper is organized as follows. In Section 2, we describe the BOC signal model and its partial correlations. Section 3 proposes an unambiguous correlation function with a sharper main-peak. In Section 4, tracking performances of several unambiguous correlation functions are compared in terms of the TESD. Finally conclusion is drawn in Section 5.

2. Signal Model

Interpreting a sub-carrier pulse as the sum of two rectangular sub-pulses, we can express the baseband equivalent of a BOC signal as

$$b(t) \;=\; \sqrt{P} \sum_{i=-\infty}^{\infty} c_i p_{T_c}(t - iT_c) d(t) c_{sc}(t), \qquad (1)$$

where P is the signal power, $c_i \in \{-1, 1\}$ is the ith chip of a PRN code with period T, $p_{T_c}(t)$ is the PRN code waveform defined as a unit rectangular pulse over $[0, T_c)$, and $d(t)$ is the navigation data. In addition, $c_{sc}(t) = \sum_{l=0}^{\infty} h_l p_{T_s}(t - iT_c - lT_s)$ is the square wave sub-carrier, where N is

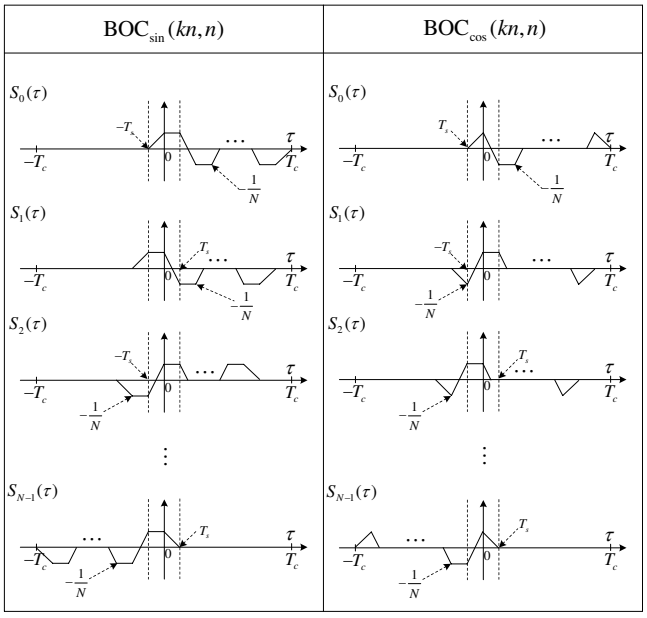


Fig. 1: Partial correlations for $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$.

number of the sub-pulses in T_c, $h_l \in \{-1, 1\}$ is the sign of the lth sub-pulse, and $T_s = T_c/N$ is the duration of a sub-pulse: For $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$, the set (N, h_l, T_s) is specified as $\left(4k, (-1)^{2ki+\lfloor \frac{l}{2} \rfloor}, \frac{1}{4kn \times 1.023 \text{ MHz}}\right)$ and $\left(4k, (-1)^{2ki+\lceil \frac{l}{2} \rceil}, \frac{1}{4kn \times 1.023 \text{ MHz}}\right)$, respectively, where $\lceil x \rceil$ is the smallest integer not smaller than x, and $\lfloor x \rfloor$ is the largest integer not larger than x. We assume that every chip of the PRN code is an independent random variable taking on +1 and -1 with equal probability and the code period T is sufficiently large compared with the chip period T_c. It is also assumed that a pilot channel (i.e., $d(t) = 1$) for tracking is provided [9].

The normalized BOC autocorrelation can be expressed as

$$\begin{aligned} R(\tau) &= \frac{1}{PT} \int_0^T b(t)b(t+\tau)dt \\ &= \sum_{l=0}^{N-1} \{ \frac{1}{N} \sum_{m=0}^{N-1} h_l h_m \Lambda_{T_s}(\tau + (l-m)T_s) \} \\ &= \sum_{l=0}^{N-1} S_l(\tau), \end{aligned} \tag{2}$$

where

$$\Lambda_\epsilon(\tau) = \begin{cases} 1 - \frac{|\tau|}{\epsilon}, & |\tau| \le \epsilon, \\ 0, & |\tau| > \epsilon \end{cases} \tag{3}$$

is a triangular function and

$$S_l(\tau) = \frac{1}{N} \sum_{m=0}^{N-1} h_l h_m \Lambda_{T_s}(\tau + (l-m)T_s) \tag{4}$$

is the lth partial correlation. Fig. 1 shows partial correlations for $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$.

3. Proposed Correlation Function

Figs. 2 and 3 show the first and second steps to generate the proposed correlation function for $\text{BOC}_{\text{sin}}(kn, n)$ by combining the partial correlations, respectively. From Fig. 2, we can see that a correlation function $R_0(\tau)$ with no side-peak can be generated by combining $S_0(\tau)$ and $S_{N-1}(\tau)$ as

$$\begin{aligned} R_0(\tau) &= S_0(\tau) \oplus S_{N-1}(\tau) \\ &\triangleq |S_0(\tau)| + |S_{N-1}(\tau)| - |S_0(\tau) - S_{N-1}(\tau)|. \end{aligned} \tag{5}$$

Moreover, we can observe that the main-peak width of $R_0(\tau)$ is determined by the location of a zero-crossing point (denoted by α) nearest to the point that $\tau = 0$.

Thus, in this paper, we construct correlation functions $T_1(\tau)$ and $T_2(\tau)$ with α closer to the point that $\tau = 0$ by combining $R_0(\tau)$, $S_1(\tau) - S_{N-2}(\tau)$, and $S_{N-2}(\tau) - S_1(\tau)$ as

$$\begin{cases} T_1(\tau) & = (S_1(\tau) - S_{N-2}(\tau)) \boxplus R_0(\tau) \\ & \triangleq |S_1(\tau) - S_{N-2}(\tau) + R_0(\tau)| \\ & \quad -|S_1(\tau) - S_{N-2}(\tau)| \\ T_2(\tau) & = (S_{N-2}(\tau) - S_1(\tau)) \boxplus R_0(\tau) \end{cases} \tag{6}$$

as shown in Fig. 2. Then, we combine $T_1(\tau)$ and $T_2(\tau)$ as

$$R_1(\tau) = T_1(\tau) \oplus T_2(\tau) \tag{7}$$

yielding an unambiguous correlation function with a narrow main-peak; however, the correlation function $R_1(\tau)$ provides a poor tracking performance since it is the combination of only four partial correlations out of $N = 4k$ partial correlations.

Thus, as shown in Fig. 3, we combine the unambiguous correlation function $R_1(\tau)$ with partial correlations $\{S_l(\tau)\}_{l=1}^{N-2}$ as

$$R_{\text{proposed}}(\tau) = R_1(\tau) + \sum_{l=1}^{N-2} S_l(\tau) \oplus R_1(\tau). \tag{8}$$

Although the figures are shown for $\text{BOC}_{\text{sin}}(kn, n)$ only, the proposed scheme is directly applicable to $\text{BOC}_{\text{cos}}(kn, n)$ using (5)-(8). The height of the main-peak of proposed correlation function is two regardless of the value of k. The main-peak widths are $\frac{17}{25}T_s$ and $\frac{1}{2}T_s$ for $\text{BOC}_{\text{sin}}(kn, n)$ and $\text{BOC}_{\text{cos}}(kn, n)$, respectively.

In the tracking process, the discriminator output

$$D(\tau) = R^2_{\text{proposed}}\left(\tau + \frac{\Delta}{2}\right) - R^2_{\text{proposed}}\left(\tau - \frac{\Delta}{2}\right) \tag{9}$$

is applied to the loop filter to drive the numerically controlled oscillator (NCO), which advances or delays the clock of the local signal generator until τ become zero, where Δ is the early-late spacing.

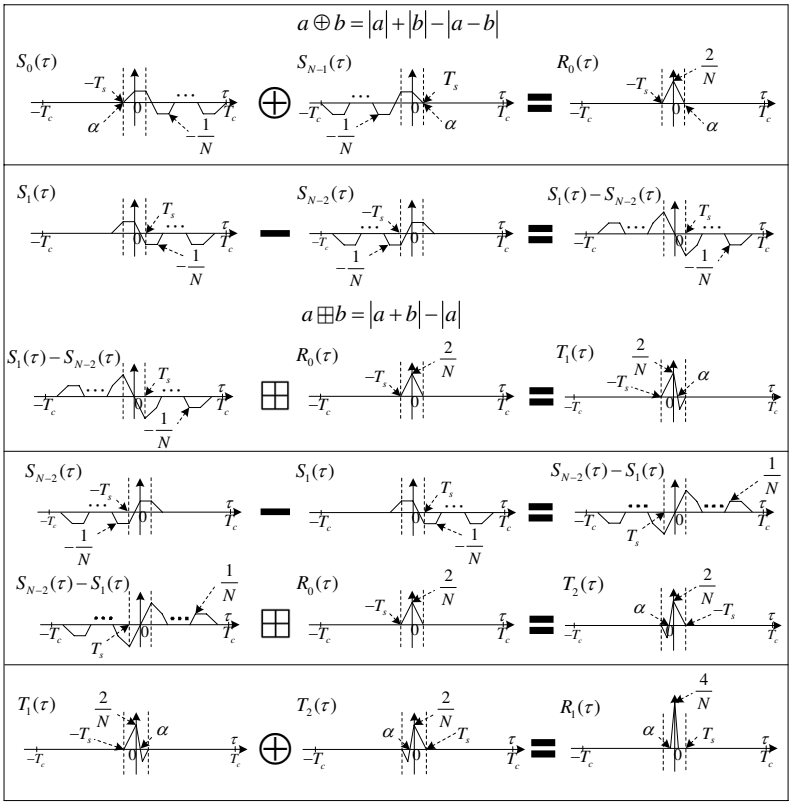


Fig. 2: The first step in generating the proposed correlation function for $\text{BOC}_{\text{sin}}(kn, n)$.

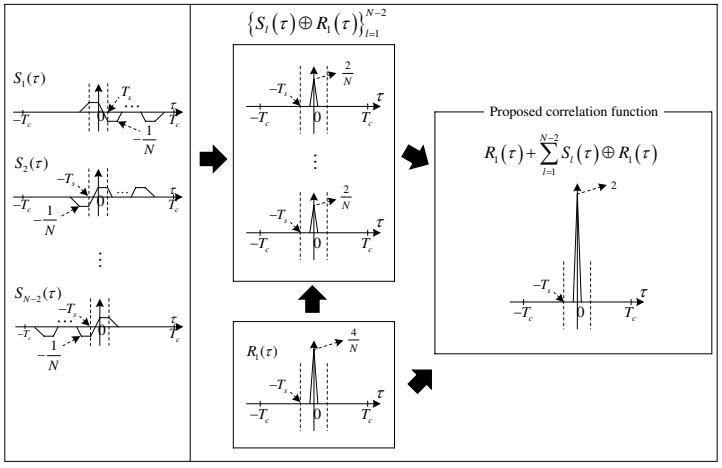


Fig. 3: The second step in generating the proposed correlation function for $\text{BOC}_{\text{sin}}(kn, n)$.

4. Numerical Results

The tracking performances of the several unambiguous correlation functions are compared in terms of the TESD: Here, the TESD is defined as $\frac{\sigma}{G}\sqrt{2B_L T_I}$, where σ is the standard deviation of $D(\tau)|_{\tau=0}$, B_L is the bandwidth of the loop filter, T_I is integration time, and $G = \frac{dD(\tau)}{d\tau}\big|_{\tau=0}$ is the discriminator gain [10]. We assume the following parameters: Galileo E1-B PRN code with period $T = 4$ ms, $T_I = T$, $\triangle = T_s/4$, $B_L = 1$ Hz, and $T_c^{-1} = 1.023$ MHz.

Fig. 4 shows the TESD performances of the proposed and conventional correlation functions as a function of the carrier-to-noise ratio (CNR) when $k = 2$: Here, the CNR is defined as P/N_0 with N_0 the noise power spectral density. The scheme in [7] is shown for sine-phased BOC signals only, since the correlation function is only dedicated to $\text{BOC}_{\text{sin}}(kn, n)$. From the figure, It is clearly observed that

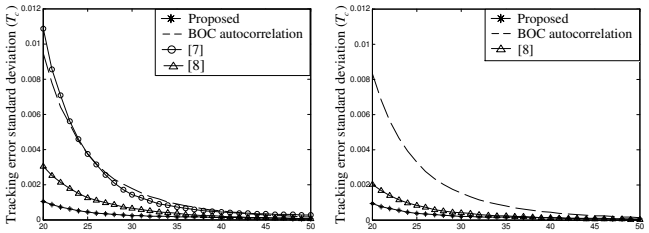


Fig. 4: Tracking error standard deviation of the proposed and conventional correlation functions as a function of CNR when $k = 2$.

the proposed correlation function provides a better TESD than the conventional correlation functions in the CNR range $20 \sim 50$ dB-Hz of practical interest.

5. Conclusions

In this paper, we have proposed an unambiguous BOC correlation function with a sharp main-peak based on a combination of partial correlations. At first, we have generated an unambiguous correlation function with a narrow main-peak using four partial correlations. Then, we have obtained an unambiguous correlation function with a sharp main-peak by adding multiple unambiguous correlation functions with same main-peak width, which come from combinations of the $4k - 2$ partial correlations and the narrow unambiguous correlation function. Finally, it has been observed that the proposed correlation function offers a performance improvement over the conventional correlation functions in terms of the TESD.

6. Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2012R1A2A2A01045887 with funding from the Ministry of Science, ICT&Future Planning (MSIP), Korea, by the Information Technology Research Center (ITRC) program of the National IT Industry Promotion Agency under Grant NIPA-2013-H0301-13-1005 with funding from the MSIP, Korea, and by National GNSS Research Center program of Defense Acquisition Program Administration and Agency for Defense Development.

References

[1] J. W. Betz. "Binary offset carrier modulations for radionavigation," *J. Inst. Navig.*, vol 48, no. 4, pp. 227-246, Dec. 2001.

[2] E. Kaplan and C. Hegarty, Understanding GPS: *Principles and Applications*, 2nd ED., Norwood: Artech House, 2006.

[3] J. Wu and A. G. Dempster, "Applying a BOC-PRN discriminator to cosine phased BOC(f_s, f_c) modulation," *Electron. Lett.*, vol. 45, no. 13, pp. 689-690, June 2009.

[4] J. A. Avila-Rodriguez, "On generalized signal waveforms for satellite navigation," Ph.D. dissertation, Dept. Aer. Engineer., University of Munich, Munich, Germany, 2008.

[5] A. Burian, E. S. Lohan, and M. K. Renfors, "Efficient delay tracking methods with sidelobes cancellation for BOC-modulated signals," *EURASIP J. Wireless Commun. Network*, vol. 2007, article ID. 72626, 2007.

[6] O. Julien, C. Macabiau, M. E. Cannon, and G. Lachapelle, "ASPeCT: unambiguous sine-BOC(n, n) acquisition/tracking technique for navigation applications," *IEEE Trans. Aer., Electron. Syst.*, vol. 43, no. 1, pp. 150-162, Jan. 2007.

[7] Z. Yao, X. Cui, M. Lu, Z. Feng, and J. Yang, "Pseudo-correlation-function-based unambiguous tracking technique for sine-BOC signals," *IEEE Trans. Aer., Electron. Syst.,* vol. 46, no. 4, pp. 1782-1796, Oct. 2010.

[8] Y. Lee, D. Chong, I. Song, S. Y. Kim, G-I Jee, and S. Yoon, "Cancellation of correlation side-peaks for unambiguous BOC signal tracking," *IEEE Commun. Lett.,* vol. 16, no. 5, pp. 569-572, May 2012.

[9] F. D. Nunes, M, G. Sousa, and J. M. N. Leitao, "Gating functions of for multipath mitigation in GNSS BOC signals," *IEEE Trans. Aer., Electron. Syst.,* vol. 43, no. 3, pp. 951-964, July 2007.

[10] A. J. Van Dierendonck, P. Fenton, and T. Ford, "Theory and performance of narrow correlator spacing in a GPS receiver," *J. Inst. Navig.,* vol. 39, no. 3, pp. 265-283, Fall 1992.

A General Review of different methods for Wireless Power Transmission

Shahram Javadi[1], Aliasghar Mohamedi[2]
Islamic Azad University, Central Tehran Branch
[1] Assistant Professor of Electrical Engineering Dept.
[2] Electrical Engineering Dept., IAU, Central Tehran Branch

Email[1]: sh.javadi@iauctb.ac.ir
Email[2]: aa.mohamedi@gmail.com

Abstract

The concept of wireless power transmission is as old as the electrical power itself. The scientists were interested in this subject since the generation of the electricity and using it. In fact, whenever the electrical power is generated, the matter of transmitting that power is bounded with it too. So far, wired transmission is still the common method because the wireless power transmitting methods have some drawbacks that make them not ready to be widely used.

This paper is a general review on wireless power transmission methods that are invented so far. In this paper, it is focused on 4 methods rather than the others because of their practicality. They are magnetic resonance, capacitive coupling, laser and microwave. First of all, The systems and general performances of each method is explained, then some examples of works done so far around the world for the introduced methods are presented and finally, the methods are compared with each other in several aspects.

Keywords: Wireless Power Transmission, Magnetic Resonance, Capacitive Coupling, Microwave, Laser

1. Introduction

The first attempt for wireless power transmission is registered for the experiments of Heinrich Hertz that he transferred high frequency power by parabolic reflectors in 1888 [1]. In the late 18th century and the beginning of the 19th century, Nicola Tesla proposed the idea of wireless power more widely. He transferred the electrical power using magnetic resonance systems. In 1899 Tesla did a major breakthrough at Colorado Springs by transmitting 100 million volts of high-frequency electric power wirelessly over a distance of 26 miles at which he lit up a bank of 200 light bulbs and ran one electric motor. With this souped up version of his Tesla coil, Tesla claimed that only 5% of the transmitted energy was lost in the process, but broke of funds again, he looked for investors to back his project of broadcasting electric power in almost unlimited amounts to any point on the globe. The method he would use to produce this wireless power was to employ the earth's own resonance with its specific vibrational frequency to conduct AC electricity via a large electric oscillator [2].Tesla also designed a demonstrations of wireless energy transfer that is called the "Electrostatic Method" or Capacitive Coupling. An electric field of alternating current, high potential, high frequency is generated across a ground plane. The sending conductor (primary) must have a matched frequency receiving conductor (secondary) in order to maintain high efficiency in energy transfer [3]. In 1961 William C. Brown published a paper on the feasibility study of transmitting the electrical power using microwave. Microwave power transmission of tens of kilowatts has been well proven by existing tests at Goldstone in California (1975) and Grand Bassin on Reunion Island (1997) [1]. In the case of electromagnetic radiation closer to visible region of spectrum (10s of microns (um) to 10s of nm), power can be transmitted by converting electricity into a laser beam that is then pointed at a solar cell receiver This mechanism is generally known as "power beaming" because the power is beamed at a receiver that can convert it to usable electrical energy [1].

Wireless power transmission methods are divided into 2 categories of “near field” and “far field” according to

the distance of the transmitted power. In near field methods, the distance between the transmitter and the receiver is less than the "Fresnel" parameter:

$$S = \frac{D^2}{4\lambda}$$

Where "D" is the antenna's length or diameter and λ is the wavelength. Magnetic and electrostatic coupling are counted as near field and microwave and laser methods are counted as far field method [1].

2. Brief explanation of the methods

a) Magnetic Resonance

In wireless power transmission by resonance method, we have two coils with a determined selfish capacity, each coupling to a capacitor with a determined capacity and meantime they have mutual inducting effects with each other. They play the role of our transmitter and receiver. One of the coils is connected to the AC power supply and the other is connected to the load. The performance of this system is that when the AC source is connected to the transmitter coil, an AC current is established. The AC current in the transmitter coil establishes an AC flux with the same frequency. The frequency of the source must be equated to resonant frequency of the transmitter and receiver circuit. Alternating flux produced by the transmitter coil, would be received by the receiver coil by mutual induction and since its frequency is equal to the resonant frequency of the receiver circuit, the maximum energy transfer amount occurs. Overview of such system is shown in figure (1).

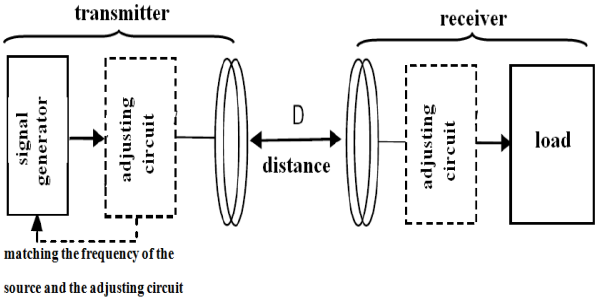


Figure (1) - The overview of magnetic resonance system

Consider a circuit comprising a capacitor and an inductor. Each of the inductor and capacitor has impedance which can be calculated as follows:

$$Z_L = j\omega L \quad (1)$$

$$Z_C = \frac{1}{jC\omega} \;\; ; \;\; \omega = 2\pi f \quad (2)$$

Now if the working frequency of the circuit is such that the absolute value of Z_L and Z_C is the same, we can say that the circuit is at resonant mode and the working frequency of the circuit in this mode is called the resonant frequency. If we equate the absolute values of Z_L and Z_C we will have:

$$Z_L = Z_C \Longrightarrow j\omega L = \frac{1}{j\omega C} \Longrightarrow f = \frac{1}{2\pi\sqrt{LC}} \quad (3)$$

The last equation is used to calculate the resonant frequency of an LC circuit.

b) Capacitive coupling

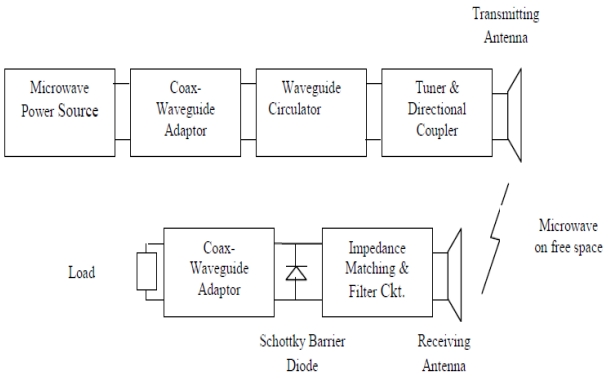
Electrostatic induction or capacitive coupling is the passage of electrical energy through a dielectric. In practice it is an electric field gradient or differential capacitance between two or more insulated terminals, plates, electrodes, or nodes that are elevated over a conducting ground plane. The electric field is created by charging the plates with a high potential, high frequency alternating current power supply. The capacitance between two elevated terminals and a powered device form a voltage divider [1].

The electric energy transmitted by means of electrostatic induction can be utilized by a receiving device, such as a wireless lamp. Nikola Tesla demonstrated the illumination of wireless lamps by energy that was coupled to them through an alternating electric field [1].

"Instead of depending on *electrodynamics induction* at a distance to light the tube . . . [the] ideal way of lighting a hall or room would . . . be to produce such a condition in it that an illuminating device could be moved and put anywhere, and that it is lighted, no matter where it is put and without being electrically connected to anything. I have been able to produce such a condition by creating in the room a powerful, *rapidly alternating electrostatic field*. For this purpose I suspend a sheet of metal a distance from the ceiling on insulating cords and connect it to one terminal of the induction coil, the other terminal being preferably connected to the ground. Or else I suspend two sheets . . . each sheet being connected with one of the terminals of the coil, and their size being carefully determined. An exhausted tube may then be carried in the hand anywhere between the sheets or placed anywhere, even a certain distance beyond them; it remains always luminous." [1]

c) Microwave

William C. Brown, the pioneer in wireless power transmission technology, has designed, developed a unit and demonstrated to show how power can be transferred through free space by microwaves. The concept of Wireless Power Transmission System is explained with functional block diagram shown in Figure 2. In the transmission side, the microwave power source generates microwave power and the output power is controlled by electronic control circuits. The wave guide ferrite circulator which protects the microwave source from reflected power is connected with the microwave power source through the Coax – Waveguide Adaptor. The tuner matches the impedance between the transmitting antenna and the microwave source. The attenuated signals will be then separated based on the direction of signal propagation by Directional Coupler. The transmitting antenna radiates the power uniformly through free space to the rectenna. In the receiving side, a rectenna receives the transmitted power and converts the microwave power into DC power. The impedance matching circuit and filter is provided to setting the output impedance of a signal source equal to the rectifying circuit. The rectifying circuit consists of Schottky barrier diodes converts the received microwave power into DC power [4]. The



schematic view of the microwave method is shown in figure (2).

Figure (2) - the schematic view of the microwave method

d) Laser

Laser power beaming is the wireless transfer of energy (heat or electricity) from one location to another, using laser light. The basic concept is the same as solar power, where the sun shines on a photovoltaic cell that converts the sunlight to energy. Here, a photovoltaic cell converts the laser light to energy. The key differences are that laser light is much more intense than sunlight, it can be aimed at any desired location, and it can deliver power 24 hours per day. Power can be transmitted through air or space, or through optical fibers, as communications signals are sent today, and it can be sent potentially as far as the Moon [5]. The schematic view of the laser method is shown in figure (3).

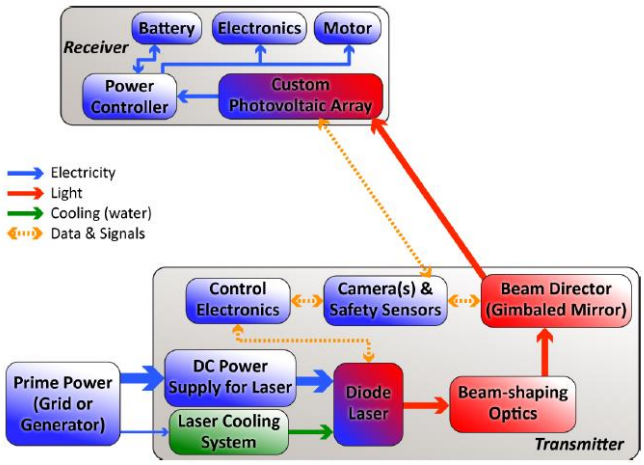


Figure (3) - the schematic view of the laser method

3. Comparison of the methods

In this section, the proposed methods are comparing with each other in 5 aspects of: transmitting power range, transmitting distance range, commercial aspects, efficiency and biologic effects. The summary of the results is shown in the table (1).

4. Challenges and future view

Initial costs of the WPT systems, especially in microwave and laser methods are almost high. This is one of the major reasons that WPT is not being widely used [17, 18].

To sustain the constant power level, there are few challenges for WPT. This is due to the electromagnetic wave scatters freely in space as it propagates, which causes the efficiency to be much lower. Leaving some energy left unused or transferred unused. But using multiple antenna arrays will be able to solve the problem [18].

Since the world is lit by wires and every electrical device is fed with wires, it will be biggest challenge to implement the WPT technology. There should be complete revolution in the electrical world for manufacturing and designing. Still the safety of the microwaves remains a question for the public. Some of the countries which depend on electrical energy for the economy like our country Bhutan: it will be a heavy blow. This was the main reason that Tesla failed to convince then the sponsors and entrepreneurs of his time to carry out his project [18].

Table (1) – Comparison of the presented methods

Comparison parameter / Method's name	Power range	Distance range	Commercial	Efficiency	Biological
Magnetic resonance	Up to 60 watts has been reported [6]	Up to 8 times of the transmitter and the receiver [6]	Not expensive [7]	Up to 45% have been reported [6]	Not reported any harmful effect [1,8]
Capacitive coupling	It is almost like the magnetic coupling				
Microwave	Up to 100kw feasibility study [9]	Scientists are studying on transmitting from the Moon to the Earth [10]	Expensive	Up to 54% has been reported [12]	If the wave parameters don't increase so mauch more than the usual doesn't have harmful effects [4,13]
Laser	Up to several hundred kw has been reported [14]	Up to 1 km has been reported [14]	Expensive	Up to 30% has been reported [15]	If the radiation steadies, damages the living tissues [16]

Many thinks that WPT is not safe and fear its impact in human health and environment, but as per the IEEE standard, but the safety studies has been taken that its radiation level would be never higher than the dose received while opening the microwave oven door, meaning it is slightly higher than the emissions created by cellular telephones[1]. Thus the public exposure of WPT fields would be below existing safety guidelines (ANSI/IEEE exposure standards) [18].

More efficient energy distribution systems and sources are needed by both developed and under developed nations. In regards to the new systems, the market for wireless power transmission is enormous. It has the potential to become a multi-billion dollar per year market [19].

The increasing demand for electrical energy in industrial nations is well documented. If we include the demand of third world nations, pushed by their increasing rate of growth, we could expect an even faster rise in the demand for electrical power in the future. These systems can only meet this 90–94 % efficient transmission [19].

5. An example of works done so far for each method

The MIT researchers successfully demonstrated the ability to power a 60 watt light bulb wirelessly, using two 5-turn copper coils of 60 cm (24 in) diameter, that were 2 m (7 ft) away, at roughly 45% efficiency. The coils were designed to resonate together at 9.9 MHz (≈ wavelength 30 m) and were oriented along the same axis. One was connected inductively to a power source, and the other one to a bulb. The setup powered the bulb on, even when the direct line of sight was blocked using a wooden panel [6].

In [20], authors have proposed wireless power distribution with capacitive coupling to overcome disadvantages in magnetic coupling like power decrease for inexact connection. In this paper a new wireless power distribution using capacitive coupling is proposed excited by multi-stage switched mode active capacitor to increase provided power. The proposed system improves power transfer efficiency

without LC resonance so that it is robust against parameter change. In addition the proposed active negative capacitor works stable without any feedback loop.

In [21], A microwave power beaming system was developed to realize wireless power supply to a Micro Aerial Vehicle. This system consists of transmitting, tracking, and receiving systems. In the transmitting system, a 5.8GHz microwave beam was irradiated from an active phased array antenna. Transmitting power was 4W and the beam divergence angle was 9deg. In the tracking system, a 2.45GHz pilot signal was detected by a two-dimensional tracking antenna and the position was deduced though the software retro-directive functions. The maximum tracking error was 1.97deg in the azimuth direction and 1.79deg in the traverse direction. Then, FM/AM wireless camera signal was used the pilot signal for tracking. The maximum error was 3.22deg in the traverse direction and was 4.97deg in azimuth direction that were slightly larger than those without modulation.

A method to use laser to supply power remotely for multimode wireless sensor networks is proposed in [22]. In the working space of wireless sensor networks, laser is transformed to a spatial distributed light field with certain uniformity and its wavelength is converted to the most sensitive wavelength of solar cell by the phosphor element. Thus, more wireless sensor networks nodes can be powered at the same time. A demonstration experiment is carried out, an yttrium aluminum garnet phosphor element is stimulated by the 3 W 457 nm laser and a uniform diffused light field with 3 π/4 space angle is obtained. The average optical energy density is 85 μ W/cm^2 at the distance 1.5 m away from the phosphor surface. An ultra low-power consumption energy harvesting system is designed.

6. Conclusion

In this paper the concept of wireless power transmission was being discussed. After a brief history, 4 methods of WPT were introduced. Then each of the methods was briefly explained. In the next section, the methods were compared to each other in several aspects. It is shown that the far field methods are more expensive than the near field methods. Then challenges and the future views for wireless power transmitting were discussed. As we saw, WPT is not widely used because of some reasons such as small amount of efficiency, commercially expensive etc. And in the last section, an example of works done so far for each method was presented.

References

1- www.wikipedia.org
2- www.mind-course.com
3- www.element14.com
4- S. Sheik Mohammed, K. Ramasamy, T. Shanmuganantham, "wireless power transmission- a next generation power transmission system", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 13, 2010
5- Laser Motive company's "laser power beaming fact sheet", 2012
6- Andre kurs, Aristeidis karalis, Robert Moffatt, J.D. Joannopoulos, Peter Fisher, Marin Soljacic, "Wireless Power Transfer via Strongly Coupled Magnetic Resonances", International Journal of Advanced Engineering & Applications, Jan. 2010, pp. 177-181
7- Derek Runge, Carl Westerby, "Design Review: Wireless Power Transmission Using Resonant Coils" , ECE 445, Fall 2008
8- Peter Fisher, Robert Moffatt, Marin Soljacic, Andre Kurs, John Joannopoulos, Aristeidis Karalis, "Goodbye wires... MIT experimentally demonstrates wireless power transfer", MIT experimentally demonstrates wireless power transfer." PHYSorg.com. 7 Jun 2007
9- Matsumoto, Hiroshi. "Research on Solar Power Satellites and Microwave Power Transmission in Japan", IEEE Microwave Magazine. December 2002. pp. 36 – 45
10- Dr. David R. Criswell, Dr Robert D. Waldron, "LUNAR SYSTEM TO SUPPLY SOLAR ELECTRIC POWER TO EARTH", 25th Intersociety Energy Conversion Engineering Conference, Reno, Nevada, August 12-17, 1990
11- Group Members: Eric Lo, Hau Truong , Louis Elnatan, Alvin Mar, Ha Nguyen, Adviser: Dr. Ray Kwok, " WIRELESS BATTERY CHARGER", Dec 02, 2005
12- Andrew Bomber, Professor La Rosa, "Wireless Power Transmission: An Obscure History, Possibly a Bright Future", Physics 464: Applied Optics, March 4, 2006
13- ALLAHYAR KANGARLU, PIERRE-MARIE L. ROBITAILLE, "Biological Effects and Health Implications in Magnetic Resonance Imaging", *MRI Facility, 1630 Upham Drive, Columbus , OH 43210, Center for Advanced*

Biomedical Imaging, Department of Radiology, The Ohio State University,Columbus, OH, 43210 Received 24 June 1999; revised 27 March 2000; accepted March 28, 2000

14- http://www.nasa.gov/home/hqnews/2009/nov/HQ 09-261 power beam.html

15- www.powerbeaminc.com

16- Shreyas Srinath, Sahana S Bhandari, "Optic based wireless power transmission for wireless sensor networks", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 10, December-2012

17- Sagolsem Kripachariya Singh, T. S. Hasarmani, and R. M. Holmukhe, "Wireless transmission of electrical power overview of recent research & development", *International Journal of Computer and Electrical Engineering, Vol.4, No.2, April 2012*

18- Rajen Biswa , "Feasibility of Wireless Power Transmission", Electronics and Communication Engineering College of Science and Technology Rinchending PhuentsholingMay, 2012

19- Sourabh Pawade, Tushar Nimje, Dipti Diwase, "Goodbye Wires: Approach to Wireless Power transmission", International journal of emerging technology and advanced engineering, ISSN 2250-2459, volume 2, Issue 4, April 2012

20- Fnato, Hirohito Dept. Electr. & Electron. Eng., Utsunomiya Univ., Utsunomiya, Japan Chiku, Yuki; Harakawa, Kenichi, "Wireless power distribution with capacitive coupling excited by switched mode active negative capacitor", Electrical Machines and Systems (ICEMS), 2010 International Conference

21- Ishiba, Mai, Graduated Sch. of Frontier Sci., Univ. of Tokyo, Chiba, Japan Ishida, Jun; Komurasaki, Kimiya; Arakawa, Yoshihiro, "Wireless power transmission using modulated microwave", crowave Workshop Series on Innovative Wireless Power Transmission: Technologies, Systems, and Applications (IMWS), 2011 IEEE MTT-S International

22- Wang, Ning Key Lab. of Opto-Electron. Technol. & Syst., Educ. Minist. of China, Chongqing, China Zhu, Yong; Wei, Wei; Chen, Jianjun; Liu, Shenshen; Li, Ping; Wen, Yumei M, "One-to-Multipoint Laser Remote Power Supply System for Wireless Sensor Networks", Sensors Journal, IEEE, Feb 2012

Design and Evaluation of Wireless Tele-Control System Considering Channel Equalizer

[1]Faramarz Alsharif,[2]ShiroTamaki,[3]Tustomu Nagado,[4]Mohammad Reza Alsharif, [5] Heung Gyoon Ryu
[1,2, 3,4] Graduate School of Engineering and Science
University of the Ryukyus
Nishihara, Japan
[5] Chungbuk National University
Cheongju, South Korea
Department of Electronic Engineering
[1]faramarz_asharif@yahoo.com
[2]shiro@ie.u-ryukyu.ac.jp
[3]nagado@eee.u-ryukyu.ac.jp
[4]asharif@ie.u-ryukyu.ac.jp
[5]ecomm@chungbuk.ac.kr

***Abstract*— In this paper, we aim to design a Tele-Control system for the closed-loop control system. Tele-Control system consists of control plant, controller, feedforward channel and feedback channel in the closed-loop control system. Feedforward channel is located between controller and control plant and feedback channel, respectively. Basically, channels would be multipath channel due to the reflections. Therefore, in order to design a suitable controller for the closed-loop system, we have to consider the feedforward channel and feedback channel. In other words we design the controller corresponding to the open-loop system. In this research we have considered the PID controller. Moreover, the effects of channels are reduced after equalization by FIR filter. The stability and performance of the closed-loop system can be evaluated by step response. The control plant is set to be an Osprey in helicopter mode when it takes off. Moreover, we have compared the conventional control system which is without channel and Tele-Control system. Eventually, in conclusion we discussed about the performance and stability of Tele-Control system that even though the closed-loop system can be stabilize but performance will be degraded due to existence of multi delay in channel. Thus, PID parameters are tuned in order to overcome the degradation. Eventually, we could obtain a desirable response. However, there was a reverse overshoot due to existence of multipath channel.**

Keywords: Tele-Control System, Multipath Channel, Equalization, PID Controller, Time Lag

1. INTRODUCTION

The utilization of Tele-Control system is one of the significant issues in the servo systems. Especially, when system requires control in distant. The advantage of Tele-Control System is that we can realize servo systems to be managed and observed it's behaviors from distant and for maintenance of controller since controller is located in observation center. Let us clarify the Tele-Control system. Basically, in Tele-Control system they are always two channel. One is the feedforward channel to send the optimal or compensated input to the control plant and the other one is the feedback channel since output signal should be sent to the controller side in order to calculate the error and to minimize it. So, these channels are disadvantages of utilization of Tele-Control system. First of all due to the usage of communication system in the closed-loop system we would have some impairment such as phase noise, Doppler effects, frequency offset, delays and attenuations. The mentioned impairment can be solved by implanting the system that has high function capabilities. Therefore, phase noise, Doppler effects, frequency offset can be repaired by installing the advanced function capability. However, the received signal should be equalized to get the original information from sender. Therefore, in order to get the exact data from sender we have to equalize the received signal. The received signal may be distracted by the multipath channel. Multipath channel effect occurs concerning the circumstances of the environment of control plant. In other words, multipath channel is inclusion of accumulated delayed and attenuated direct path signal. Even though sender has sent the original signal but in receiver side we will have distracted signal by the multipath channel. Thus equalization of signal is required in receiver side. For equalization, first we have to compose the replica of the unknown channel. The composition of the replica channel of the unknown channel can be done by FIR adaptive filter. However, the composition of the replica channel is not sufficient. We have to realize the inverse system of replica Channel. Therefore, the inverse channel is realized after the receiving the distracted signal. This has role of equalizing the received signal. These kinds of process should be implemented in two different stages. One is in the feedforward side of the receiver and the other one in the feedbackpart of the receiver since we have round trip multipath channel in the closed-loop system or Tele-Control System. After realizing the equalizer, we implement it in the closed-loop system. The control Plant is considered to be Osprey [1] in helicopter mode when it has to maintain the commanded desired altitude. In the Next chapters more details about the design of controller and the closed-loop system considering equalizer are mentioned.

2. FUNDUMENTAL THEORY OF CONVENTIONAL FEEDBACK CONTROL

As we have discussed in introduction previously, the principal purpose of utilization of Tele-Control System is that observe the control plant behaviors and attitudes from a distant and send an optimal input to control plant. Basically, when Tele-Control System is required that control plant should be located in distant. For instance unmanned vehicle, missiles, Telescope in space and so on. So one of the advantages of usage of Tele-Control System is that plant can be controlled from a distant and controller maintenance can be done rapidly in case of impairment since controller is located in our side. Anyhow, let us define our proposed Tele-Control system. Before moving to Tele-Control System let's have a review of the conventional feedback control system. Assume that we have a control plant as *P*, Controller *K* and *r, u, y, e* stand for reference, input, output and error signals, respectively. In Fig. 1, it shows how Conventional Feedback Control System [2-5] work.

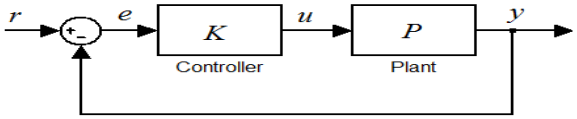


Fig. 1 Conventional Feedback Control System

Through Fig.1 we could obtain the relation between reference and output signal as follows.

$$y = KPe \quad (1)$$

$$e = r - y \quad (2)$$

By substituting equation (2) to equation (1) we have obtained the transfer function from *r*to*y*which has shown in equation (3).

$$y = \Delta K \Pr \quad (3)$$

That $\Delta = \frac{1}{1+KP}$ is sensitivity function of the closed-loop system and $\Delta^{-1} = 0$ is characteristicequation of the closed-loop system.They are several methods to design the controller. For instance, the most known controller is PID Controller. PID parameters can be set by obtaining the crossover frequencyω_{ph}. Crossover frequency is the frequency that when theopen-loop's phase becomes -180°. Subsequently, we can obtain the limited stable proportional gain *Kc* and oscillation period *Tc*of the plant. The derivation of *Kc*and*Tc* shown as follows.
Suppose that the open-loop of Fig. 1 is

$$y = G_O r \quad (4)$$

Where, $G_O = PK$. Here we assume that *K=1*.

By solving the characteristicequation $\Delta^{-1} = 0$. Solution of characteristicequation can be expressed as follows.

$$\Delta^{-1} = 0 \rightarrow 1 + P = 0$$

$$\rightarrow P = -1$$

The polar expression of the given plant is $P(j\omega) = |P(j\omega)| e^{j\angle P(j\omega)}$ .Where ω is angular frequency and *j* is square root of *-1*.
Then we will have $|P(j\omega)| e^{j\angle P(j\omega)} = e^{-j180}$.
Eventually, $\angle P(\omega_{ph}) = -180$.

Then gain margin of the given plant is as follows

$$gm = \frac{1}{\mathrm{Re}[G(j\omega_{ph})]} \quad (5).$$

As well *Kc* and *Tc* could be obtained as follows.

$$20\log_{10} Kc = gm \rightarrow Kc = 10^{\frac{gm}{20}} \quad (6)$$

and

$$Tc = \frac{2\pi}{\omega_{ph}} \quad (7).$$

Subsequently P, PI, and PID parameters could be summarized in table1.

Table1. PID Parameters

Controller	*Kp*	*Ti*	*Td*
P	*0.5Kc*	∞	*0*
PI	*0.45Kc*	*0.83Tc*	*0*
PID	*0.6Kc*	*0.5Tc*	*0.125Tc*

Which *Kp*, T_i and T_d are proportional gain, Integrator gain and derivative gain, respectively.
Thus the PID controller can set as follows in frequency domain.

$$K(s) = Kp + \frac{1}{T_i s} + T_d s \quad (8)$$

Where, *s*is Laplace operator.
Furthermore, the stability can be guarantee by the small gain theorem is satisfied. That is the gain of the open-loop transfer function with controller, should be less than *1*,

$$|P(j\omega)K(j\omega)| < 1 \quad (9)$$

that is

$$|P(j\omega)||K(j\omega)| < 1$$

$$\rightarrow \sqrt{K_p^2 + \left(T_d\omega - \frac{1}{T_i\omega}\right)^2} < \frac{1}{|P(j\omega)|} \quad for \ \forall\omega .$$

3. TELE-CONTROL SYSTEM

So far we have discussed about the process of design the controller. Next let us define Tele-Control System as follows. Following figure shows the structure of Tele-Control system.

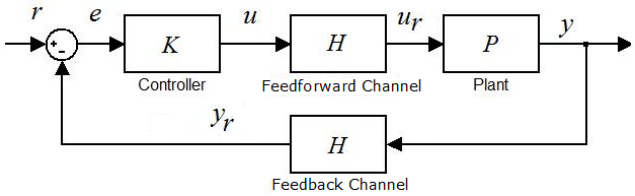


Fig. 3 Tele-Control System

Here H is Multipath channelanu_r,y_r are received input and received output signal, respectively. Through Fig.2, we can get the closed-loop system's transfer function according to following equations.

$$y = PHKe \quad (10)$$

$$e = r - Hy \quad (11)$$

Afterward we get the transfer function between r and y which iscomplementary sensitivitytransfer function as follows.

$$y = \Delta_H PHKr \quad (12)$$

Where, $\Delta_H = \dfrac{1}{1+PKH^2}$ stands for sensitivity transfer function which is from r to e.

As we can see in sensitivity function of the closed-loop system, it has been involved with Channel's square. Our proposed method is to reduce the effect of the channel in the sensitivity function. The proposed method has shown in following Fig. 4.

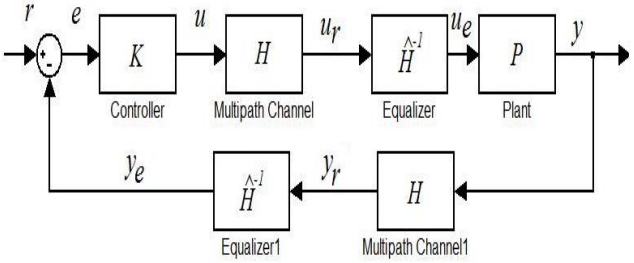


Fig. 4 Configuration of the proposed method

Here, y_e, $u_e$and $\hat{H}^{-1}$stand for the equalized input signal, equalized output signal and Equalizer, respectively. $\hat{H}$ itself is the replica channel of H that is estimated with adaptive filter. However, before getting starting the proposed method let us see how we can design a controller for Tele-control system without considering channel equalizer.

For fig.3, we could design a PID controller according to the explained procedure in previous chapter.

Therefore, the gain margin of system with multipath channel can be calculated as follows.

$$1 + PH^2 = 0 \quad (13).$$

Here Multipath channel's model can be express as follows. $H(s) = \sum_i \alpha_i e^{-L_i s}$. Where $\alpha$is the attenuation factor and L is Time-Delay and $i$is the number of taps. In next chapter Multipath channel will be introduced in details.

The polar Expression of Multipath channel is:

$$H(j\omega) = \sum_i \alpha_i e^{-jL_i\omega} \Rightarrow H(j\omega) = |H(j\omega)| e^{j\angle H(j\omega)}.$$

Where, $|H(j\omega)| = \sqrt{\left(\sum_i \alpha_i \cos(L_i\omega)\right)^2 + \left(\sum_i \alpha_i \sin(L_i\omega)\right)^2}$

and $\angle H(j\omega) = -\tan^{-1}\left[\dfrac{\sum_i \alpha_i \sin(L_i\omega)}{\sum_i \alpha_i \cos(L_i\omega)}\right]$.

Then we have:

$$P(j\omega)H(j\omega)H(j\omega) = -1$$

$$|P(j\omega)||H(j\omega)|^2 e^{j\left(\angle P(j\omega) + \angle H^2(j\omega)\right)}$$

$$= |P(j\omega)|\left(\left(\sum_i \alpha_i \cos(L_i\omega)\right)^2 + \left(\sum_i \alpha_i \sin(L_i\omega)\right)^2\right)$$

$$\times e^{j\left(\angle P(j\omega) - 2\tan^{-1}\left[\frac{\sum_i \alpha_i \sin(L_i\omega)}{\sum_i \alpha_i \cos(L_i\omega)}\right]\right)} = 1e^{j(-180)}.$$

The crossover frequency can be obtained by solving the following equation.

$$\angle P(j\omega_{ph}) - 2\tan^{-1}\left[\frac{\sum_i \alpha_i \sin(L_i\omega_{ph})}{\sum_i \alpha_i \cos(L_i\omega_{ph})}\right] = -180.$$

Afterward, gain margin will be obtained and eventually controller can be designed.However, the stability point of view, as it is clear, it is very hard to satisfy the small gain theorem.

$$\left|P(j\omega)H^2(j\omega)K(j\omega)\right|<1 \qquad (14)$$

$$\rightarrow \left|K(j\omega)\right|<\frac{1}{\left|P(j\omega)\right|\left|H^2(j\omega)\right|}$$

$$=\sqrt{K_p^2+\left(T_d\omega-\frac{1}{T_i\omega}\right)^2}<\frac{1}{\left|P(j\omega)\right|\left(\left(\sum_i \alpha_i \cos(L_i\omega)\right)^2+\left(\sum_i \alpha_i \sin(L_i\omega)\right)^2\right)}$$

for $\forall\omega$.

As it can be seen in equation (14)'s conditions, it is very hard to maintain the small gain theorem due to the existence of multipath channel. Thus, we have to reduce the effect of multipath channel in sensitivity transfer function. In the next chapter proposed method, characteristic of multipath channel and Multipath channel canceller are introduced.

4. REDUCTION OF MULTIPATH CHANNEL EFFECTS IN THE CLOSED-LOOP SYSTEM

As we have discussed previously, existence of the multipath channel make the systemunstable and it is very hard to determine the PID parameter that satisfies the small gain theorem which has been mentioned in Equation (14). Therefore, somehow the multipath channel should be eliminated in order to get rid of the instability. Thus, equalizer is required in the receiver side of the plant for the feedforwardmultipath channel and another equalizer is required in the controller side for feedback multipath channel. By implementation of the equalizer we can reduce the effect of the multipath channel. However, before equalizing the received signal estimation of multipath channel is required. Estimation of multipath channel can be done by adaptive filter. After reconstructing the replica of multipath channel, inversion of the replica channel should be implemented in cascade to vanish the multipath channel. In following figure the process of the proposed system is indicated in detail.

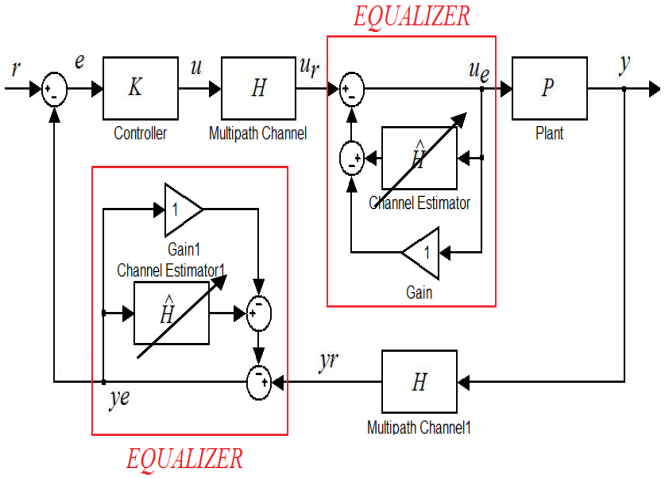


Fig. 5. Configuration of the proposed method in detail

$$y=KH\hat{H}^{-1}Pe \qquad (15)$$

$$e=r-H\hat{H}^{-1}y \qquad (16)$$

Afterward, we have as follows.

$$y=\frac{PH\hat{H}^{-1}K}{1+PK\left(\frac{H}{\hat{H}}\right)^2}r \qquad (17)$$

That $\frac{1}{1+PK\left(\frac{H}{\hat{H}}\right)^2}$ is the sensitivity function of the proposed method.

Here, if and only if when $\hat{H}=H$, then we would obtain equation (12) that is identical to the conventional feedback control system. However, this does not happen since replica channel cannot realize the precise characteristic of multipath channel. Nevertheless, we can reduce the effects of the multipath channel. So this makes the closed-loop system stable. However, performance is going to be degraded anyhow.

5. MULTIPATH CHANNEL

Basically, multipath channel is consequences of the reflected desired signal or in other words accumulation of several attenuated and delayed reference signal. Especially, this phenomenon would be occurred easily and frequently in metropolitan ambit which comprised of high density of building and so. Also, it would occur in mountainous area as well. Following shows the signal composition in time domain of reflected signal which comprise ofmultipath channel.

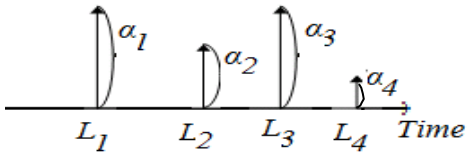


Fig. 6. Signal compositions in time domain

According to Fig. 6, it can be expressed in mathematical model for multipath channel$H$ such as follows.

$$h(n)=\sum_i \alpha_i \delta(n-\tau_i) \qquad (18)$$

Where, α and τ stands forattenuation andtime delay factor of multipath channel, respectively.As it is clear in equation (18), we need to estimate the multipath channel in order to get rid of instability in the closed-loop system. Therefore, for estimation of multipath channel FIR (finite impulse response) adaptive filter is utilized. The tap number of FIR adaptive filter concerns the length of multipath channel. Hence, the length of filter should exceed the length of multipath channel. Otherwise reconstruction of replica multipath channel becomes hard. In next chapter several adaptive filter algorithms are introduced.

6. ADAPTIVE FILTERS

An adaptive algorithm [6-7]is a set of recursive equations used to adjust the weight vector of replica multipath channel*H*automatically to minimize the effect of multipath channel in sensitivity function.Such that the weight vector converges iteratively to the optimum solution that corresponds to the bottom of the performance surface, i.e. the minimum of MSE (Mean Square Error). The Least- Mean- Square (LMS) algorithm is the most widley used among various adaptive algorithm because of its using the negative gradient of the instantaneous squared error. In general expression for*H*that intend to adapt itself to *H*. The derivation of updated weight vector of LMS algorithm can be shown as follows. Here the adaptation done in time domain so we consider the *h(n)* as the imverse Laplace trasfer of *H(s).* As well for $\hat{h}(n)$ is the inverse Laplace transfer of $\hat{H}$ (s). Before getting start the calculation, let us define the error signal in the adaptive filter e_f(n).

$$e_f(n) = h(n) * u(n) - \hat{h}(n) * u_e(n) \qquad (19)$$

According to stochastic gradient algorithm, we would have as follows. Here *n* and *i* are iteration and filter's tap number, respectively.

$$\hat{h}_i(n+1) = \hat{h}_i(n) - \mu \nabla e_f^2(n)$$

$$= \hat{h}_i(n+1) = \hat{h}_i(n) - \mu \frac{\partial e_f^2(n)}{\partial \hat{h}}$$

$$\nabla e_f^2(n) = \frac{\partial}{\partial \hat{h}(n)} (h(n) - \hat{h}(n))(h(n) - \hat{h}(n))^T * (u(n) \times u^T(n))$$

$$= \frac{\partial}{\partial \hat{h}(n)} (h(n) - \hat{h}(n)) \times (h^T(n) - \hat{h}^T(n)) * (u(n) \times u_e^T(n))$$

$$= \frac{\partial}{\partial \hat{h}(n)} h(n) \times \hat{h}^T(n) - 2\frac{\partial}{\partial \hat{h}(n)} h(n) \times \hat{h}^T(n) + \frac{\partial}{\partial \hat{h}(n)} \hat{h}(n) \times \hat{h}^T(n)$$

$$= (0 - 2h(n) + 2\hat{h}(n)) * (u_e(n) \times u_e^T(n))$$

$$= -2(h(n) - \hat{h}(n)) * u(n) \times u_e^T(n)$$

$$= -2e_f(n) u_e^T(n)$$

Eventually, we obtain the following equation.

$$\hat{h}_i(n+1) = \hat{h}_i(n) - 2\mu e_f(n) u(n-i) \qquad (20)$$

where μ is the step size or convergence factor that determines the stability and the convergence rate ofthe algorithm.

In the case of Normalized LMS, the LMS algorithm normalizes the step size with respect to the input signal power.

$$\hat{h}_i(n+1) = \hat{h}_i(n) - \frac{2\mu e_f(n) u(n-i)}{N\sigma_x^2} \qquad (21)$$

Where, $\sigma_x = \frac{1}{N}\sum_{i=0}^{N-1} u^2(n-i) \rightarrow$ and N is tap number of adaptive filter.

Step size is now bounded in the range of 0 to 2. It makes the convergence rate independent of signal powerd by normalizing the input vector with the energy of the input signal in the adaptive filter.

7. SIMULSTION AND RESULTS

In order to evaluate the perforamnce and stability of the proposed method, we have simulated for a system that it might requried Tele-Control system in some particular situation. The plant is chosen to be an Osprey. Osprey needs to take off in helicopter mode. After take it will change to airplain mode and it flies away. The first objective is to take off Ospery safely and with good performace by using Tele-Control System. As we disscused, in the case of Tele-Control System for sure we have multipath channel. In this situation we can consider that in the closed-loop system we would have two iedndtical multipath channel, one for feedforward and the other for the feedback. For evaluation, we will simulate the step response of the closed-loop system without equlizer, with equzlier and conventional feedback control method (without channel) to compare the performances and stability among them. The other evaluation is for Channel estimator that we have disscused in chapter 6. Following shows the Conditions of Simulation.

- The referecne altitude 50[m]. ($R(s) = \frac{50}{s}$)
- The altitude dynamics of an Osprey :

$$P(s) = \frac{1}{100\,s^3 + 215\,s^2 + 30.5\,s + 1}$$

- Controller of altitude:

$$K(s) = \frac{0.2s^2 + 0.1s + 0.05}{s}$$

- Channel Specification (with 10 taps M=10):

$$H(s) = \sum_{i=1}^{M} \alpha_i e^{-sL_i}$$

.

Where, attenuated and time delay factor are indicated as follows.

$\alpha_i = rand(i) e^{\frac{-0.1i}{M}}$, $\tau_i = 0.5i \times sort(|randn(i)|)$ (i=1 to 10).

For adaptive filter's performance, we have evaluated for equalized input signal which corresponds to feedforward multipath channel and equalized output multipath channel which corresponds to feedback multipath channel. Evaluation of each equalizer has been done by learning curve which is NMSE (Normalized Mean Square Error) showing as follows.

$$NMSE_{feedforwrd} = \frac{\sum_{n=1}^{N} \left| u_r(n) - \hat{h}(n) * u_e(n) \right|^2}{\sum_{n=1}^{N} \left| u_r(n) \right|^2}$$

$$NMSE_{feedback} = \frac{\sum_{n=1}^{N} \left| u_r(n) - \hat{h}(n) * u_e(n) \right|^2}{\sum_{n=1}^{N} \left| u_r(n) \right|^2}$$

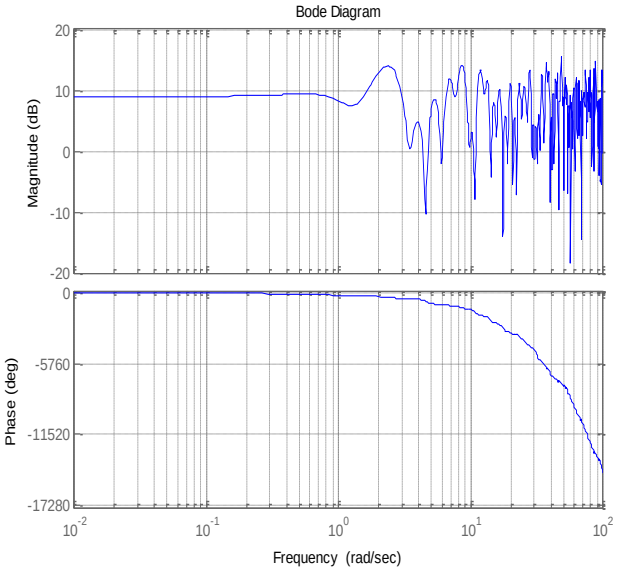


Fig. 7. The frequncy Response of Multipath Chnanel

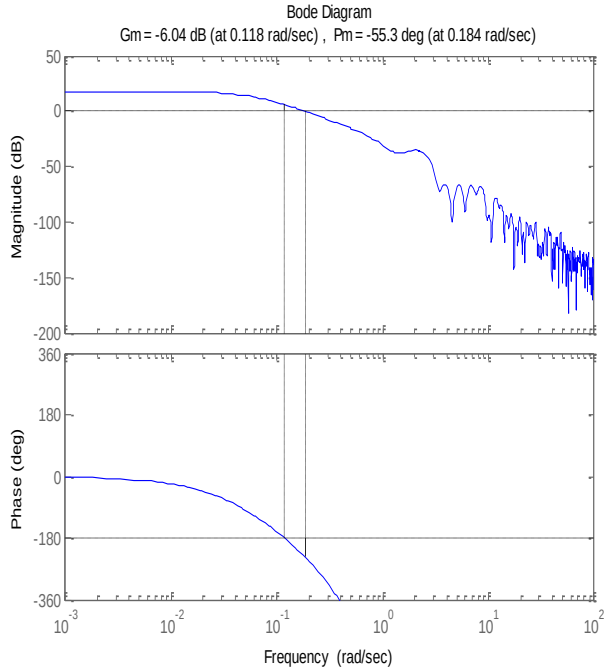


Fig.8. Bode diagram of the the open-loop without controlller and equalzier

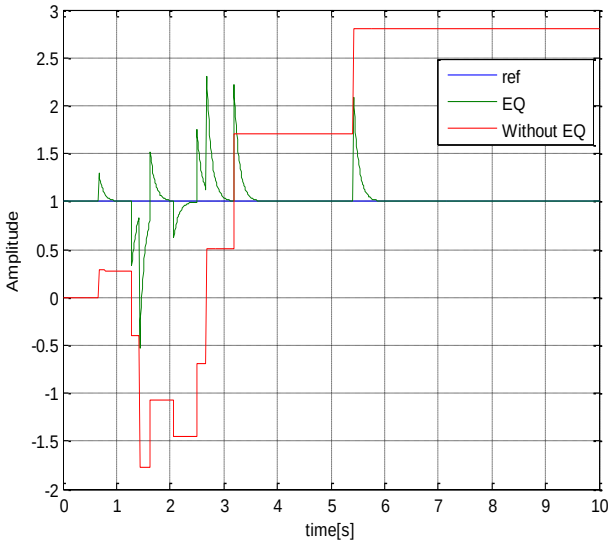


Fig. 9. Time response of mutipath Channel with and without equalzier

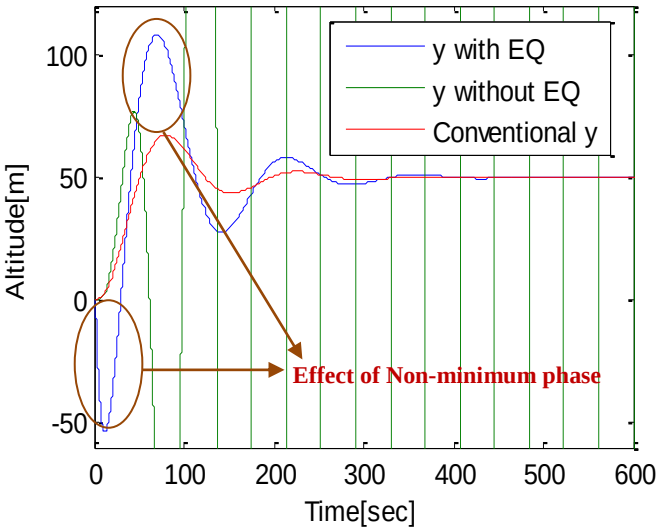


Fig. 10. The desired altitude response of the closed-loop system with and without equalizer and for conventional feedback control

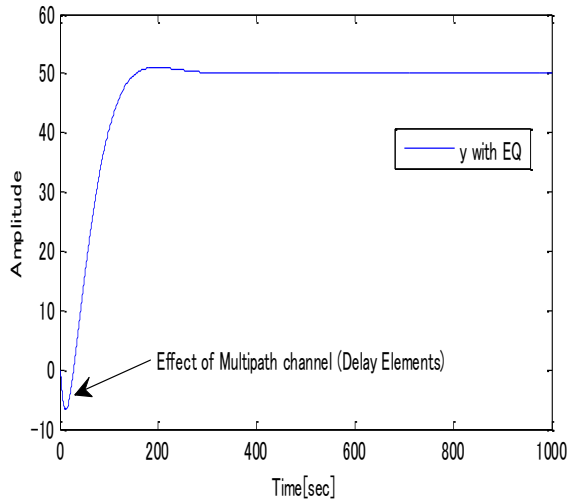


Fig. 11. The compensated step response of wireless closed-loop system with equalizer

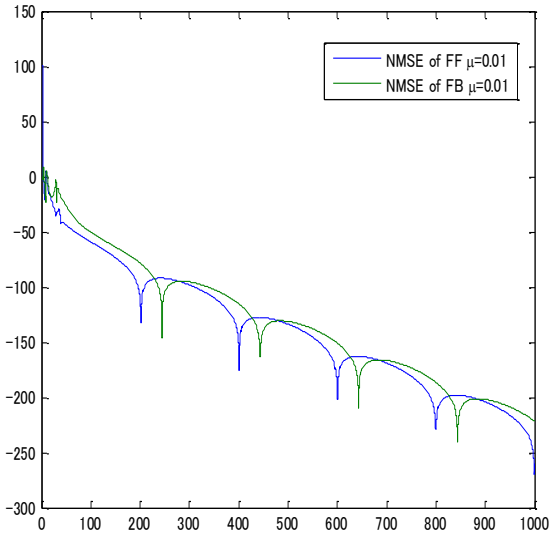


Fig. 12. Learning cutve of adaptive filters

Fig. 7 shows the bode diagram of the multipath channel. As we can see the boundary of channel exceeds 10 [dB] in several frequencies and it's phase drops very rapidly. Thus we can estimate that the closed-loop of system may become unstable that can be confrimed in Fig. 10. Fig. 8 indicates the phase margin and gain margin of thethe open-loop system without controoller. As it is claer the gain margin is -6.04[dB] and phase margin is -55.2 [deg] . Both margins are negative so the closed-loop becomes unstable. In order to get rid of multipath channel affects equalizer has been imlemented. Fig .9 shows the equalizer's performances. Eventually the step response of the closed-loop system with and with equlizer and with conventional method are indicated in Fig. 10. As it is obvious that for conventional curve it converges to reference signal with settele time of 300[sec] and roughly 10% overshoot. However, considering Tele-control system, it becomes unstable without equalizer. By utilization of equalizer we can confirm that it maintain the stability and the settle time is almost same as the convetntional curve. However, Overshoots of the system with equalzier is 100% that is not acceptable performance in the actual situation. Even the reverse oversoot that is indicated in a brwon circle is not suitable for the servo system. This caused by reverse overshoot by non-minimum phase affection. For these unsuitable matters, PID controller are tuned untill the overshoot is minimized. Thus, in Fig.11 we obtained better response compare to Fig.10. Eventually, the learning curve of the adaptive filters for feedb-forward and feedback channel are shown in Fig.12 base on two different algorithm LMS and NLMS. The learning curve for LMS becomes NaN which is unstable due to the regardlessing the normalization of the input signal. For NLMS algprithm eventhogh the learning curve is flactuatig frequently butit is decreasing by each iteration andhad a acceptable performances and can realize the replica of multipath channel in the both side of multipath feedforward and feedback channel.

8. CONCLUSION

In this paper we have proposed to implement equalizer in Tele-control system in order to get rid of instability and performance degdration of the closed-loop system which cuased by multipath channel. As a result the closed-loop system is asysmptoticaly stable. However, the performance of the closed-loop does not satisfy the condition practically. Thus, as a future work the enhacment of perforamnce should be considered in Tele-control system. Moreover, in plant there are unecrtanity factor practically. That means plant's paarmeters may changes. So far, we did not have considered the uncertanity in pant. So, if there is any uncertanity in plant the closed-loop system may be come unstable. In order to avoid instability due to uncertainty robust control should be appliyed. Therefore, to realize the joint system of roubst control and equalizer system is one of the important issue and future works.

ACKNOWLEDGMENT

This work was supported by Japan Society for Promotion of Science and Marubun Research Promotion Foundation. We wish to deeply express our gratitude to their supports.

REFERENCES

[1] R.C.Dorf, R. H .Bishop "Modern d ControlSystem",Prentice Hall2002
[2] WitoldPedrycz, "Robust Control Design an Optimal control Approach" ,Wiley 2007
[3] R. Oboe, K. Natori, K. Ohnishi, "A Novel Structure of Time Delay Control System with Communication Disturbance Observe" International Workshop on Advanced Motion Control, AMC '08. 10th
[4] F.Asharif, S. Tamaki, T.Nagado, T. Nagata and M. R.Alsharif,"Design of Adaptive Friction Control of Small-Scaled Wind Turbine System Considering the Distant Observation" ICCA, LNCS Springer pp213-221, Nov. 2012.
[5]Guillermo J., SilvaAniruddhaDatta, S.R Bhattacharyya,"PID Controller for Time-Delay System" Birkhauser, 2004
[6] Kong-Aik Lee, Woon-SengGan and SenM.Kuo,"Subband Adaptive Filtering Theory and Implementation," 2009, John Wiley & Sons, Ltd
[7]F. Asharif, S. Tamaki, M. R. Alsharif, H. G. Ryu"Performance Improvement of Constant Modulus Algorithm Blind Equalizer for 16 QAM Modulation"InternationalJournal on Innovative Computing, Information and Control, Vol. 7, No. 4, pp.1377-1384, April 2013.

Dynamic Packet Size Adaptation for Efficient Bluetooth Communication

Daniel Driscall and Kyoung-Don Kang

Department of Computer Science

State University of New York at Binghamton
{ddrisca1, kang}@binghamton.edu

Abstract

Bluetooth is a wireless protocol designed to support short-range communications with low-power consumption for ubiquitous computing. It is important to maximize the transfer rate of Bluetooth without requiring additional hardware that may increase the cost and power consumption. However, Bluetooth goodput is highly variable in space and time. To improve the goodput as much as possible without requiring a better antenna or more communication bandwidth, we develop a new lightweight method that dynamically adapts the packet length by applying hill climbing methods. To avoid potential local optima in hill climbing, we also support controlled random exploration of different hills, i.e., different packet sizes selected randomly. Our approach is different from most existing work on adaptive rate control in that it measures and enhances the goodput in the application-layer feedback loop neither relying on physical layer metrics (e.g., the received signal strength or signal to noise ratio) nor requiring any changes to the underlying hardware or firmware. Our approach is implemented as an application-layer library by extending a Bluetooth host stack in Linux. Our experiments show that the proposed method for dynamic packet size adaptation efficiently selects an appropriate Bluetooth packet size for goodput enhancement with minimal overheads.

***Keywords*—Bluetooth, Goodput, Dynamic Packet Size Adaptation via Feedback**

I. INTRODUCTION

Bluetooth is designed for low cost, low power, and short range (up to 100 meters) wireless communications in ubiquitous computing environments with normal speeds of up to 1 Mbps [2]. It consumes an order of magnitude less power compared to 802.11. It is originally designed as a single-hop cable replacement, for example, to eliminate the wire for data transfer between a smart phone and laptop. In addition, a large number and variety of Bluetooth devices have recently been developed. Due to the popularity of Bluetooth devices for ubiquitous/pervasive computing, it is desirable to significantly enhance the *Bluetooth communication goodput*, which is the application layer throughput indicating the number of useful data bits successfully transmitted to the receiver per unit time. If a higher goodput is supported, the overall user experience and energy efficiency can be enhanced due to the reduced time for data transfer and reduced retransmissions for many applications ranging from data transfer between a pair of data Bluetooth devices and multimedia applications. For example, the quality of service and energy efficiency for peer-to-peer multimedia streaming via 802.11 accompanied by Bluetooth [12] can enhanced considerably, if the Bluetootch goodput is maximized.

Generally speaking, it is challenging to support high goodput for a Bluetooth communication. Bluetooth uses the 2.4 GHz ISM band. The same radio band is used by IEEE 802.11 (WiFi), which is widely deployed to build wireless local area networks, consumes an order of magnitude higher power to support a higher bit rate and longer communication range than Bluetooth does. Other Bluetooth and non-Bluetooth devices, such as Bluetooth baby monitor and microwave ovens, also use the same band. Therefore, Bluetooth communications may suffer from interferences from other communication sources in addition to being subjected to the various vagaries of wireless links [5].

To address these challenges, we investigate how to increase Bluetooth goodput as much as possible without requiring more expensive hardware or any changes to firmware subject to the cost increase and deployment issues. We avoid extensive mathematical modeling of wireless channels based on the physical layer metrics such as the received signal strength or signal to noise ratio [1,3,5,8,10], since a mathematical model of a wireless channel derived in a specific environment may become largely inaccurate due to, for example, fluctuating channel conditions and mobility. In uncertain environments, physical layer metrics often fail to capture the application layer goodput. Therefore, in this paper, we design an adaptive and lightweight approach that hunts for an improved goodput via dynamic packet size adaptation in the application layer without modifying the underlying hardware, firmware, or protocol stack.

Adaptive rate control techniques have been studied for efficient wireless communications [1-3,5-10]. However, most existing work assumes that short packets are less susceptible to losses or use a fixed single packet size. In this paper, we empirically show that *a larger Bluetooth packet is not necessarily more subject to losses* and a lower goodput as a result (Fig. 2 in Section III). We also show that *a single packet size may support largely fluctuating goodputs*, even if the other communication settings, such as the distance between two communicating Bluetooth devices and WiFi interference, do not change significantly. Based on this observation, we

design an opportunistic approach to dynamic packet size adaptation to enhance the Bluetooth goodput. By applying the hill climbing technique, we continue to go up or down a hill (i.e., increase or decrease the packet size) as long as the data transfer rate does not decrease. On the other hand, we begin to explore the opposite direction of the hill, if the effective transfer date decreases; that is, we begin to decrease the packet size, if the transfer rate decreased as the packet size was increased in the previous trial or vice versa. In addition, to avoid local maxima, we introduce a probabilistic yet controlled random exploration technique to explore other randomly selected hills for potential enhancement of the goodput in a systematic manner. In this way, we decrease the probability of settling for a packet size selected by hill climbing even though there could be a different packet size that provides a considerably higher goodput.

We have implemented our adaptive algorithm by extending the Bluetooth host stack in Linux. Especially, we extend L2CAP (Logical Link Control and Adaptation Protocol) [11,4] without changing the underlying Bluetooth radio stack. Although the default packet size is 672 bytes in L2CAP, it can be increased up to 65536 bytes. In connection-oriented Bluetooth communications considered in this paper, an L2CAP sender retransmits a packet until it receives an acknowledgement from the receiver or the connection fails.

We have performed experiments in a residential building considering good and poor wireless connections. To create a good connection, we place a pair of Bluetooth devices one foot apart from each other. On the other hand, to create a poor connection, a sender is placed two floors away from the receiver in a residential house. In the experiments with a poor connection, our approach considerably outperforms a baseline approach, i.e., the original unmodified L2CAP, which uses a fixed default packet size. Comparing to the baseline, our approach improves the average effective transmission rate by approximately 48%, while significantly reducing the variance. For good wireless connections, it provides similar performance to the baseline.

The rest of the paper is organized as follows. Our approach to Bluetooth goodput enhancement via a feedback in the application layer is described in Section II. Performance evaluation results are discussed in Section III. Related work is discussed in Section IV. Finally, Section V concludes the paper and discusses future work issues.

II. ALGORITHM DESIGN FOR DYNAMIC PACKET SIZE ADAPTATION

In this section, background information of the Bluetooth protocol stack is given. In addition, our approaches to hill climbing and random exploration for Bluetooth goodput enhancement are discussed.

A. Bluetooth Protocol Stack

The Bluetooth protocol stack consists of two parts: (1) a controller stack implemented in hardware to deal with the timing critical radio interface and (2) a host stack implemented in the operating system to handle high level data.

In this paper, we focus on extending the host stack to support dynamic packet size adaptation for goodput enhancement. Since our approach does not directly deal with low-level radio, it is not necessary to modify the control stack. Hence, our approach is easier to deploy than an alternative approach that requires control stack modifications is.

In L2CAP, the MTU (Maximum Transmission Unit) must be at least 48 bytes long but not longer than 65535 bytes, while the default MTU size is 672 bytes. Generally, a large MTU is more efficient with lower overheads, since the packet header size is fixed regardless of the MTU size. Other overheads such as the packet scheduling delay can be amortized better for a bigger MTU. However, simply using the maximum MTU size may not result in the highest goodput due to stochastic wireless channel characteristics.

We have experimentally observed that any packet size larger than 43860 bytes results in significant packet losses. Thus, an ideal MTU size should be big enough to amortize transmission overheads but small enough to avoid excessive packet losses. In our experiments, the relation between the MTU size and effective transmission rate is nonlinear but highly stochastic when 48 bytes $\leq$ MTU $\leq$ 43860 bytes. (More details are given in Section III.) Based on this observation, we design an opportunistic approach to increasing the goodput via hill climbing and random exploration within the range of 48 – 43860 bytes. The notations used for the design of an algorithm are summarized in Table 1.

Notation	Meaning
MTU	Max. Transmission Unit (48 – 43860 bytes; 672 bytes by default)
α, β ($0 < \alpha < \beta < 1$)	Magnitude of packet size adaptation
N	Number of packets used as a unit for effective transfer rate measurement and packet size adaptation
R_i	Measured goodput for the i^{th} set of N packets where $i \geq 1$
D	Current direction for hill climbing
D_{new}	New direction for hill climbing
p	Probability for a random jump
k	Number of climbs for tryout of a new hill

Table 1. Notations

B. Hill Climbing

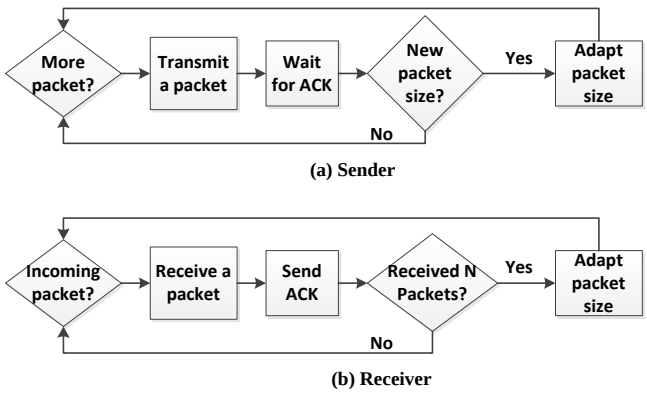


Fig. 1 Hill Climbing for Packet Size Adaptation

The overall procedure of our hill climbing for enhancing the effective transmission rate is depicted in Figure 1. Also, a more detailed description of the dynamic packet size adaptation follows:

1. Initially, use the default MTU size and set the preferred direction *D* = *increase*.

2. For the ith set of N successfully received packets where i ≥ 1, the receiver computes the goodput is:
   $R_i = \frac{1}{T_N} \sum_{j=1..N} s_j$ bps (bits per second)
   where TN is the total time for transmitting the N packets and s_j is the size of the j^{th} packet in bits.

3. **If** $R_i \le R_{i+1}$, the receiver sets the new preferred direction $D_{new} = D$. If D_{new} = *increase*, MTU_{new} = MTU(1+α) where 0 < α < 1. Else, MTU_{new} = MTU(1-α).

4. **Else** set D_{new} = opposite of *D*. If D_{new} == *decrease*, MTU_{new} = MTU(1-β). Otherwise, MTU_{new} = MTU(1 + β) where 0 < α < β < 1.

5. Set D = D_{new}.

6. Start random exploration if appropriate.

7. Piggyback MTU_{new} to an acknowledgement packet (or any other packet) transmitted to the sender.

8. Repeat Steps 2 – 7 until all data is completely transmitted to the receiver.

Algorithm 1. Dynamic packet size adaptation

In Step 1, we assume that an increase in packet size is preferred, since it is initially unknown whether we need to increase or decrease the MTU length to increase the goodput. Hence, we set the current hill climbing direction D = increase.

In Step 2, we compute the goodput R_i for the i^{th} set of the *N* consecutively received packets. Similarly, R_{i+1} is computed for the $(i+1)^{th}$ set of *N* packets.

In Step 3, if $R_i \le R_{i+1}$, the effective transmission rate has increased. If the current direction for packet size adaptation was the increasing (decreasing) direction, we further increase (decrease) MTU by αMTU to continue going up (down) the hill, i.e., to continue increasing the effective transmission rate.

On the other hand, in Step 4, we observe that the goodput has decreased if $R_i > R_{i+1}$. In this case, we start going to the opposite direction, because the current hill climbing strategy failed to increase the goodput. If the current direction is increasing the packet size, we decrease it by βMTU or vice versa. Notably, by setting α < β, we take a conservative approach to avoid going too far in a wrong direction. By doing this, our algorithm aims to avoid indefinitely oscillating between a packet size increase and decrease, which may happen in the worst case if α = β.

In Step 5, we set $D = D_{new.}$ In Step 6, random exploration is performed, if necessary, to avoid local optima in a probabilistic manner. In Step 7, the receiver sends the new MTU value to the transmitter to enhance the goodput. Since the MTU size can be expressed using only two bytes, it is piggybacked to an acknowledgement (or any other packet) transmitted from the receiver to the sender. Steps 1 – 7 are repeated until all data is successfully transmitted to the receiver.

C. Random Exploration

When the effective transmission rate is computed in Step 2 of Algorithm 1, the receiver picks a random number between 0 and 1. If the random number is not bigger than a pre-set probability *p* for random exploration, the hill climbing procedure picks a random packet size between 48 and 43860 bytes. The hill climbing routine is restarted at the newly picked random packet size.

The new hill will be climbed for another pre-set parameter *k* number of climbs. At the end of *k* climbs, if the new goodput is higher than the value measured before the random jump event, this new hill will be climbed continuously. If the goodput decreases, the hill climbing algorithm will abandon this new hill and return to the original hill climbed before the random search.

If the algorithm is in the new hill tryout phase (i.e., *k* number of climbs have not yet been completed), a random shift to a new hill is suspended to give each random search a chance to prove or disprove the new randomly picked hill is a viable one.

Note that our hill climbing algorithm strives to converge to an appropriate packet size that can increase the goodput. Our algorithm is lightweight. Also, it does not require extensive modeling of wireless communication channel characteristics in a specific environment. Thus, it consumes little resources and readily usable for Bluetooth communications in any mobile/ubiquitous environment.

III. Performance Evaluation

For performance evaluation, we have a sender transfer a file in a PC to a receiver running on a laptop via Bluetooth. This could easily be applied to other mobile devices or game consoles to transfer data with each other. Although file transfer is used for experimental purposes, our approach can be used to transfer any data.

First, we capture data to analyze the goodput as a function of the packet size. In this experiment, we intend to observe whether or not short packet lengths generally provide higher transfer rates, while observing that a single packet length provides different goodputs over time due to varying conditions for Bluetooth communications.

Second, we evaluate the goodput supported by our hill climbing method, while trying different parameters for random exploration. Our objective is to observe whether our approach can improve the effective transmission rate compared to the baseline method that uses a fixed default MTU size of 672 bytes in L2CAP. In addition, we aim to investigate issues related to random exploration. Although random exploration may help hill climbing avoid local maxima, excessive randomness may decrease the goodput, while increasing potential fluctuations. We intend to find appropriate parameters for random search to increase the effective transfer rate, while avoiding severe rate variances.

A. *Packet Length vs. Goodput*

In this subsection, we perform an experiment to show the need for hill climbing to enhance the goodput of a Bluetooth communication. More specifically, a 3.2MB file is transferred once between a pair of Bluetooth devices. We have used two setups for experiment: one good connection and one poor connection, respectively. In the good connection, a pair of Bluetooth devices is 1 foot apart. The poor connection uses the same two Bluetooth devices; however, the devices are two floors away from each other in a house. Since Bluetooth is mainly designed to support short-range communication, the poor connection represents a challenging situation for a Bluetooth communication. However, this is a realistic scenario in which users may want to avoid frequent trips between several places in a building for data transfer.

Figure 2 shows the results of the experiments. One experiment is done for a good connection. This experiment shows that a high goodput can be achieved at many different packet sizes. From the figure, we observe that a relatively small packet size (<1,000 bytes) often yields largely oscillating transfer rates, because short packets are subject to the large overhead associated with the Bluetooth protocol. These results contradict to the commonly accepted rule of thumb to use small packets in lossy environments to support a higher goodput [1-3,5-10]. From Figure 2, we observe that a large packet size does not necessarily provide a higher goodput. Thus, hill climbing for dynamic packet size adaptation is a viable approach to enhancing the Bluetooth goodput.

In addition, we create one poor connection at a time to transfer a 3MB file and repeat this experiment three times. Figure 2 shows that the effective transfer rate widely varies among the three experiments using the poor connection. From this, we observe that the goodput of a poor connection is affected by not only the packet size but also other unpredictable stochastic wireless channel characteristics that may largely vary. However, the results show that it is feasible to increase the goodput rate by dynamically adapting the packet size. Also, in Figure 2, there are many individual peaks and valleys. Therefore, random exploration can further improve the goodput by avoiding local maxima.

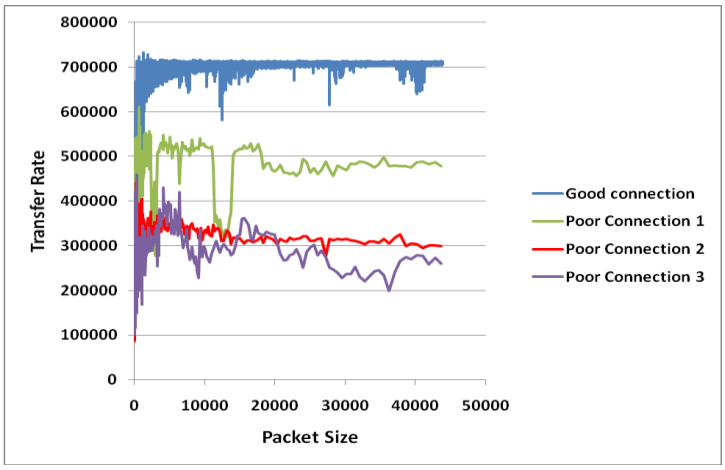


Fig. 2 Packet Size (bytes) vs. Goodput (bits per second)

B. *Tuning Random Exploration Parameters*

In the experiments presented in this subsection, the 3.2MB file is repeatedly transferred from the sender to the receiver for 15 minutes. At the end of a 15 minute experiment, we measure the total amount of the data successfully transferred to the receiver to compute the average goodput. In a good (poor) connection, the Bluetooth device pair is 1 foot (two floors) away, similar to the previous subsection.

Parameter	Value
α	0.1
β	0.2
N	20 packets
Initial value of MTU	672 bytes
Initial value of *D*	Increase
p	0.01, 0.02, ..., 0.1
k	1, 2, ..., 10
Length of an experimental run	15 minutes

Table 2. Experimental Parameter Settings

Table 2 summarizes the parameter settings used for the experiments presented in this subsection. The receiver runs our adaptive method upon receiving 20 new packets (i.e., N = 20) to enhance the effective transfer rate compared to the basic approach that uses the fixed default packet size (672 bytes in L2CAP). We use different values of p (the probability for a random jump) and k (the number of climbs for tryout of a new hill) for random jumps. Specifically, we consider p = 1%, 2%, …, 10% and k = 1, 2, …, 10. A 15 minute file transfer experiment is performed for each pair of p and k parameters to measure the average goodput. Therefore, we have executed 100 experiment runs in total where each run is 15 minutes long. In addition, we compute the standard deviation for the measured transfer rates.

Figure 3 shows the average goodput for the good connection as a function of the two parameters p and k. Also, Figure 4 shows the standard deviation of the goodput for the good connection as a function of the two parameters p and k. From these figures, we observe that the p and k parameters have *little influence* on a good Bluetooth connection, since the goodput is relatively stable for a reliable connection as shown in Figure 2. As shown in Figure 4, the standard deviation is small with *moderate values of p and small values of k*. For a good connection, a good hill can be found quickly and wasting a large amount of time by deviating from a good hill is unnecessary.

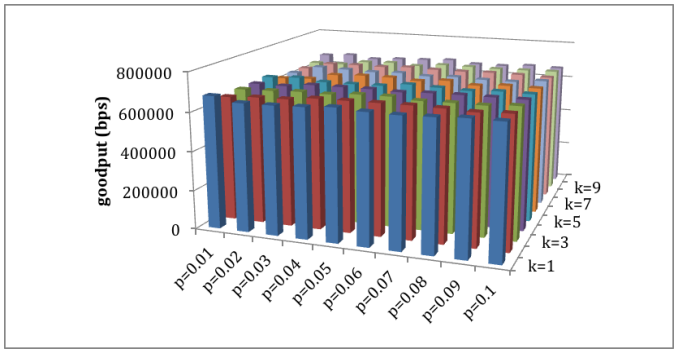


Fig. 3 Average goodput of a good connection for different p and k values

Figure 5 shows the average goodput for the poor connection as a function of the two parameters p and k. Figures 5 and 6 show the standard deviation of the goodput for the poor connection. The experimental results show that a pair of moderate p and a small k supports a high transfer rate with small standard deviation, similar to the result for the good connection.

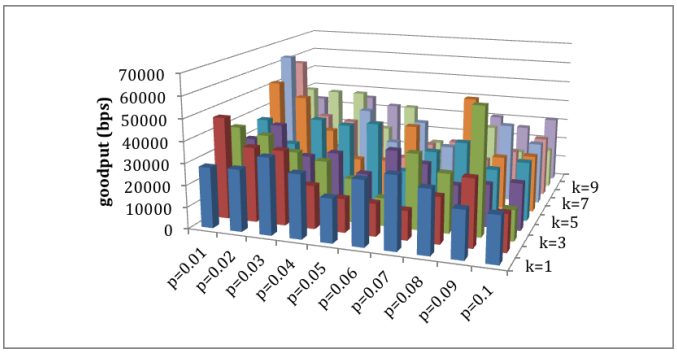


Fig. 4 Standard deviation of the goodput for a good connection

C. *Evaluation of the Goodput*

Overall, the experimental results show that the *hill climbing with controlled random explorations significantly outperforms the basic approach for a poor connection and does not severely decrease the goodput for a good connection*. For example, in our approach, selecting p = 4% and k = 1 yields an average goodput of approximately 490kbps with a poor connection. For more than 90% of the time, the transfer rate stays within 11% of the average. For a good connection, the same p and k valus yield a goodput of approximately 660kbps. For approximately 90% of the time, the transfer rate stays within 9% of the average.

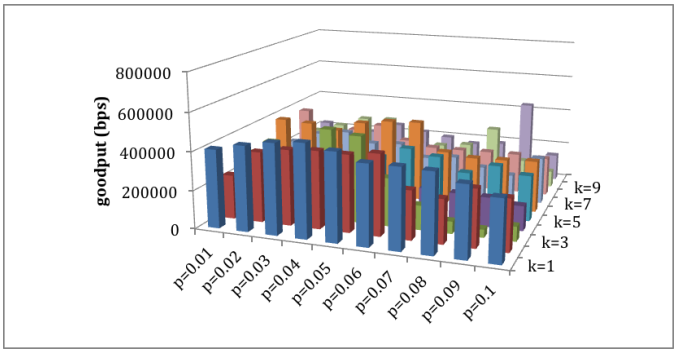


Fig. 5 Average goodput of a poor connection for different p and k values

On the other hand, the baseline supports only 330kbps with the poor connection. Also, the transfer rate fluctuates within 48% of the average. It supports roughly 690kbps with the good connection. For 90% of the time, the transfer rate stays within 10% of the average.

Hence, in these experiments, our adaptive approach enhances the goodput by more than 48% for the poor

connection, while significantly reducing the variations. For the good connection, our adaptive approach with p = 4% and k = 1 decreases the goodput by less than 5% compared to non-adaptive approach. Therefore, the overhead of our adaptive approach is acceptable.

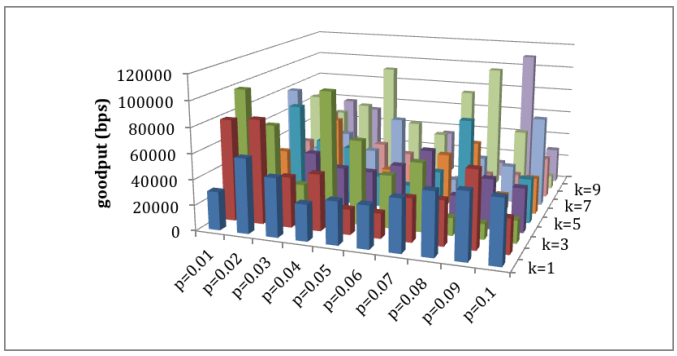


Fig. 6 Standard deviation of the goodput for a poor connection

A possible extension is to turn off our hill climbing and random exploration methods, for example, when the goodput exceeds a pre-specified threshold that indicates a very good connection. A challenge is how to predict when to turn it on again to avoid goodput losses in varying wireless communication conditions. A thorough investigation is reserved for future work.

IV. Related Work

Research on dynamic packet length adaptation has been done to maximize the rate of wireless data transfer. However, a common assumption of the previous work is that a large packet is subject to lower overheads but more losses [3, 5, 7, 10]. Based on this assumption, the packet size is reduced further as more packets are lost. In this paper, we empirically show that the assumption does not necessarily hold in practice. Thus, we hunt for a higher goodput by dynamically adapting the packet size within the whole range of the minimum and maximum packet size via efficient hill climbing based on the observed packet length vs. goodput relations.

In [8], an optimal packet size is selected based on pre-known wireless network conditions. However, this approach may not perform well in practice, because wireless network conditions often largely fluctuate due to wireless uncertainties and mobility. Packet loss models [3,10] are developed assuming a constant bit error rate. They assume that most packet errors are due to random bit errors in the payload. However, Vyas et al. [9] experimentally disprove this assumption in an 802.11a network. In [2, 6], wireless sensor nodes adapt the packet length to increase the transfer rate based on physical channel conditions. Dong et al. [2] observe that commonly used physical layer metrics such as the RSSI (received signal strength indicator) and LQI (link quality indicator) are not effective to measure the throughput in a wireless sensor network. They propose to support dynamic packet size adaptation based on the observation that most work on packet length optimization uses a fixed optimal length scheme. However, their work implemented in TinyOS 2.1 is closely tied to the underlying protocol stack and does not consider local maxima problems unlike the work presented in this paper. Instead, it focuses on sensor reading aggregation and multi-hop transmissions important in the context of wireless sensor networks.

Bluetooth uses the frequency hopping technique. Different packets are sent using different channels. Leveraging the frequency hopping nature of Bluetooth, Sarkar et al. [5] use different packet sizes for different channels based on their conditions. Chen et al. [1] propose an analytic link-layer model to find an optimal packet type or size to maximize the TCP transfer rate over Bluetooth. Instead of developing yet another analytic model based on physical layer metrics such as the RSSI or LQI, which could be largely inaccurate, we take a practical approach to improving the goodput of data transfer in Bluetooth, which dynamically adapts the packet length based on the application layer goodput feedback. As wireless channel characteristics are highly stochastic, our empirical approach enhances the goodput with little computational complexity and communicational overhead in practice. Neither is it tied to a specific mathematical model. Furthermore, unlike a majority of work based on analytic modeling and simulation studies, our approach is actually implemented in Linux and thoroughly evaluated in a real residential environment.

V. Conclusions

Bluetooth is a wireless protocol designed to support short-range data exchanges for ubiquitous computing. In this paper, we aim to considerably increase the Bluetooth goodput. To achieve this objective, we have designed a new method based on hill climbing with controlled random exploration. As our approach is an application layer solution, no modification of the underlying Bluetooth firmware or hardware is required. Thus, it can be easily deployed. The experimental results show that the proposed method can find an appropriate packet length to substantially enhance the goodput in highly dynamic wireless environments. Further, our approach is lightweight in that it does not require extensive modeling of wireless channel characteristics, complex computation, or additional communications between Bluetooth devices.

Acknowledgement

This work was supported, in part, by the NSF grant CNS-117352.

References

[1] L.-J. Chen, R. Kapoor, M. Y. Sanadidi, M. Gerla, "Enhancing Bluetooth TCP Throughput via Link Layer Packet Adaptation," IEEE ICC, 2004.

[2] W. Dong, X. Liu, C. Chen, Y. He, G. Chen, Y. Liu, J. Bu, "DPLC: Dynamic Packet Length Control in Wireless Sensor Networks," IEEE INFOCOM, 2010.

[3] X. He, F. Y. Li, and J. Lin, “Link Adaptation with Combined Optimal Frame Size and Rate Selection in Error-Prone 802.11n Networks”, IEEE ISWCS, 2008.
[4] A. Huang, “An Introduction to Bluetooth Programming,” [Online] Available at http://people.csail.mit.edu/albert/bluez-intro/
[5] S. Sarkar, F. Anjum, R. Guha, ``Optimal Communication in Bluetooth Piconets," IEEE Transactions on Vehicular Technology, Vol. 54, No. 2, pp. 709-721, March, 2005.
[6] W. Song, M. N. Krishnan, and A. Zakhor, “Adaptive Packetization for Error-Prone Transmission over 802.11 WLANs with Hidden Terminals”, IEEE MMSP, 2009.
[7] T. Vu, D. Reschke, W. Horn and Tu Ilmenau “Dynamic Packet Size Mechanism (DPSM) for Multimedia in Wireless Networks,” Multimediale Informations- und Kommunikations systeme (MIK), 2002.
[8] M. C. Vuran and I. F. Akyildiz, “Cross-layer Packet Size Optimization for Wireless Terrestrial, Underwater, and Underground Sensor Networks,” IEEE INFOCOM, 2008.
[9] A. K. Vyas, F. A. Tobagi, and R. Narayanan, “Characterization of an IEEE 802.11a Receiver using Measurements in an Indoor Environmant”, IEEE GLOBECOM, 2006.
[10] J. Yin, X. Wang, and D. P. Agrawal, “Optimal Packet Size in Error-prone Channel for IEEE 802.11 Distributed Coordination Function”, IEEE WCNC, 2004.
[11] “Bluetooth Protocols,” [Online] Available at http://en.wikipedia.org/wiki/Bluetooth_protocols#Logical_link_control_and_adaptation_protocol_.28L2CAP.29 .
[12] Yao Liu, Lei Guo, Fei Li, Songqing Chen, “An Empirical Evaluation of Battery Power Consumption for Streaming Data Transmission to Mobile Devices,” Proceedings of the 19th ACM International Conference on Multimedia, 2011

Implementation of Automotive Platform for Real-Time Diagnosis

Minwoo Jung and Jeonghun Cho
School of Electronic Engineering, Kyungpook National University, Daegu, Republic of Korea

Abstract - *As vehicle and computer technique are advanced, a vehicle has many electronic control unit (ECU). The ECUs control a series of actuators on an internal combustion engine to ensure the optimum running through reading values from a multitude of sensors within the engine bay, interpreting the data using multidimensional performance maps, and adjusting the engine actuators accordingly. It has utilized on vehicular diagnostic system such as on-board diagnostic version II (OBD-II). The OBD-II collects state of health information for various vehicular sub-systems. We propose automotive platform that can manage vehicle in real-time. It enables an early diagnosis based on storing information in server. It ensures high safety, more comfort.*

Keywords: Automotive platform; Real-Time diagnosis; Bluetooth; Electronic control unit;

1 Introduction

As computer and automotive technology are advanced, many ECUs are included in vehicle. It has utilized on vehicular diagnostic system such as OBD-II. OBD is protocol that was developed in the 1980's to help technicians diagnose and service the computerized engine systems of modern vehicles [1]. A new generation of these systems called OBD-II is present on 1996 and newer vehicles. OBD-II is an automotive term referring to a vehicle's self-diagnostic and reporting capability. They enable the vehicle owner or a repair technician to access to state of health information for various vehicle sub-systems. They are implemented via OBD-II interface such as control area network (CAN) that needs for information sharing between ECUs [2]. The CAN was developed for the automotive market to reduce the weight and cost of wiring harnesses and add additional capabilities. It is also used in factory automation, medical, marine, military and anywhere a simple yet robust network is needed. It is not necessarily a complete network system. The vehicle is diagnosed by OBD-II that based on collected vehicular information via OBD-II interface. Most of vehicle requires extra device in order to monitor state of vehicle [3]. Recently, mobile device applications allow smart devices to display and manipulate the OBD-II data accessed via USB adaptor cables or Bluetooth adapters plugged in to the car's OBD-II connector. It enables a real-time monitoring through smart devices [4].

We propose a converter that OBD protocol is converted into Bluetooth protocol and a graphic user interface (GUI) for smart devices. The smart devices have limit of storage. A server needs in order to store vehicular information that is transported in real-time. We propose cloud computing for construction server. A repair technician would analyze failure cause through stored data in server. It enables to detect potential error in vehicle. The second section describes background, third section presents proposed scenarios. Finally, last section presents conclusion and future work.

2 BACKGROUND

2.1 In-Vehicle Network

IVN with point-to-point connections is the simplest topology and can make functionality easily, shown as Fig. 1 b). But this topology has complex wiring which is hard to modify and extend. And as ECUs are increased, the wiring and connection points are increased exponentially. This is not good for production and maintenance of vehicle.

Networked IVN uses a bus connection to which other ECUs are connected this provides flexibility and scalability, as shown Fig. 1 c). ECUs, which have their own sensors and actuators, are connected with others through Bus and share information. This is better than point-to-point connection but there are some limitation to provide flexibility and scalability. Sensors and actuators do not have their own network interface, so they have to be a part of some kind of ECUs [5].

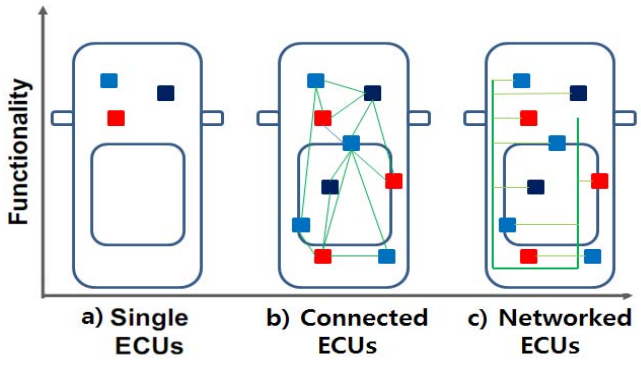


Figure 1 Vehicle E/E architecture

Fully Networked IVN, in addition to ECUs, consists of Sensors and Actuators which have their own network

interfaces thus they can connect to IVN independently and assure high flexibility and scalability. But, with the increase in number of ECUs, sensors and actuators the design, verification and validation of Automotive E/E systems become complicated. Moreover, increasing the IVN traffic causes congestion, delay and omission of IVN thus safety and dependability is at stake. This paper proposes distributed hierarchical service architecture to solve problems which are described before. Distributed Service Management Point (SMP) maintains services hierarchically and provides redundant services to assure dependability. According to the requirements, SMP can consist of various services and can modify services easily and support on-line diagnosis to validate services [6].

2.2 Bluetooth

The Medical Working Group of the Bluetooth SIG began defining a specification addressing the needs of the medical community. Under Bluetooth, a profile defines the characteristics and features including function of a Bluetooth system. The end result of this work was the HDP specification that included the MCAP (Multi-Channel Adaptation Protocol) and made use of the Device ID Profile (DI). Fig. 2 describes the interaction between a Bluetooth Protocol.

Medical Application describes the actual device application, including its user interface, application behavior, and integration layer to IEEE 11073-20601 stack implementation. IEEE 11073-20601 stack performs building, transmission, reception, and parsing of IEEE PDU packets for the associated PHD being developed. This component will directly link to the HDP. Device ID (DI) Profile is a Bluetooth profile designed to provide device specific information through use of the Service Discovery Protocol (SDP). If vendor specific information is required as part of a particular Medical Device, this profile provides specific behavior to acquire this information. A good HDP implementation offers API's to register and query for such vendor specific information.

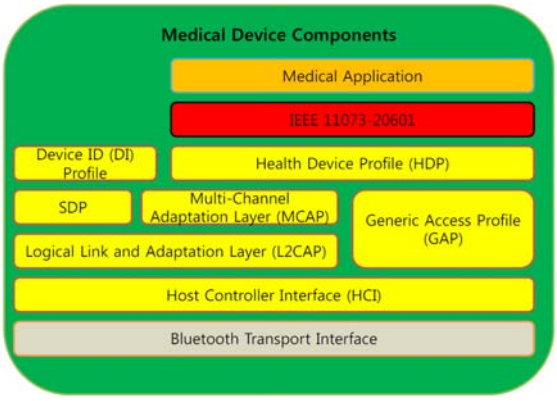


Figure 2 Bluetooth Stack

These API's can then be integrated directly into the Medical Application. Health Device Profile (HDP) is the core Bluetooth profile designed to facilitate transmission and reception of Medical Device data. The API's of this layer interact with the lower level MCAP layer, but also perform SDP behavior to connect to remote HDP devices. SDP is the Service Discovery Protocol used by all Bluetooth profiles to register and discover available services on remote devices so that connections over L2CAP can be established. Multi-Channel Adaptation Layer (MCAP) is used by HDP and facilitates the creation of a Communications Link (MCL) for exchanging generic commands, and also one or more Data Links (MDL) to transfer actual Medical Device data. MCAP is specific for the HDP and guarantees reliable transmission of data. Generic Access Profile (GAP) describes the required features of all core Bluetooth profiles including inquiry, connection, and authentication procedures. Logical Link and Adaptation Layer (L2CAP) supports protocol multiplexing, packet segmentation and reassembly, quality of service, retransmission, and flow control for the Bluetooth packets transmitted through MCAP. Host Controller interface (HCI) describes the commands and events that all Bluetooth hardware implementations can understand. Bluetooth Transport Interface describes the UART, USB, SDIO, 3-wire, ABCSP, etc. transport interface to the actual Bluetooth hardware components being used. Typically, UART and USB are the most widely used transports [7, 8].

3 IMPLEMENTATION

Experimental environment consists of 10 ECUs, shown as Fig. 4. ECUs consist of high speed module and low speed module. The high speed module requires continuously monitoring, the low speed module requires intermittently. Each ECU manages service state with state messages and replies the service requests. The ECUs can use various backbone network and sub-network in implementation level. The backbone network can use CAN2.0 or FlexRay, the sub-network can use CAN 1.0 and LIN.

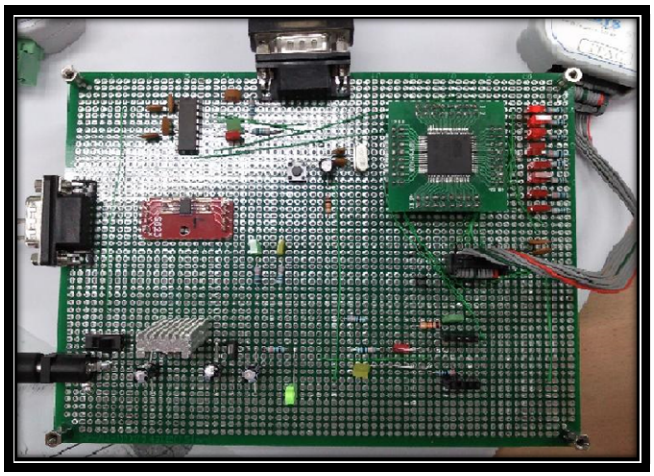


Figure 3 Bluetooth Converter

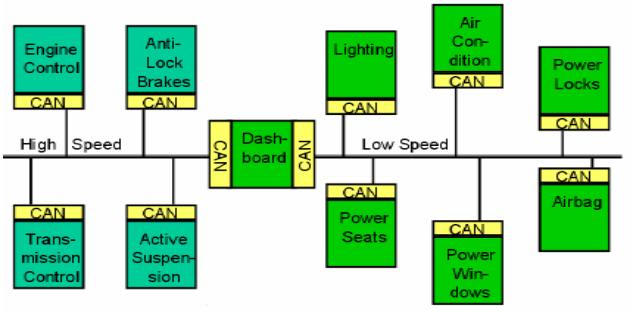


Figure 4 Experimental block diagram

A dashboard that collects information of ECUs through IVN includes Bluetooth module, micro control unit (MCU). The MCU converts OBD-II to Bluetooth in order to communicate with smart device, shown as Fig. 3. We use AT90CAN128 which is 8-bit microprocessor. The data that is converted to Bluetooth protocol transmits to Bluetooth module. The Bluetooth module performs pairing for Bluetooth communication. The Bluetooth stack implement as Bluetooth Health Device Profile (HDP) for future research, we don't consider it in this paper.

A smart device performs pairing in order to communicate with dashboard. Fig. 5 presents screen of paring processing. If not available, the smart device presents message about failure. After pairing success, data of ECUs are presented and transmit server via WiFi. The server store the data that receive from smart device. Fig. 6 presents storing data in server. The data consists of protocol type, arbiter, length, HEX type, ASCII type. A repair technicaian can request the data to server, can monitor vehicular state current and confirm vehicular history.

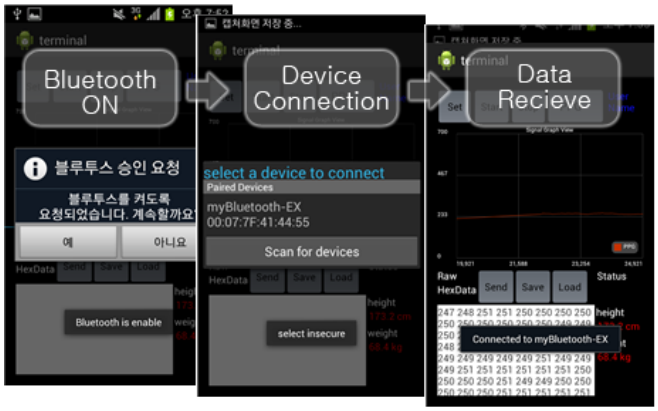


Figure 5 Bluetooth Paring

4 Conclusion

We have proposed automotive platform for vehicular real-time diagnosis. It consists of ECUs, dashboard, smart device, server. ECUs detects vehicular state and transmits data to dashboard via OBD-II interface. There are CAN, LIN, FlexRay in OBD-II interface. We establish network using CAN. The dashboard converts OBD-II to Bluetooth protocol and transmits to smart device via Bluetooth. The smart device presents information of vehicular state, transmits information to server. The server stores vehicular information and provides services to clients such as vehicular owner, repair technician. Storing vehiclular information in server will ensure exact vehicular diagnosis and convenience on repairing. We use technology about in-vehicle network, Bluetooth, cloud computing. In future, we will apply to u-Health, intelligent transportation system (ITS) and research Bluetooth pairing using NFC.

Protocol	Arbiter	Length	DATA(HEX)	DATA(ASCII)
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	EA E0	7	53 76 63 33 52 65 71	Svc3Req
Std(2.0A)	E9 E0	7	53 76 63 32 52 65 71	Svc2Req
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	D9 E0	8	53 4D 50 32 53 54 41 ...	SMP2STAT
Std(2.0A)	D8 E0	8	53 4D 50 31 53 54 41 ...	SMP1STAT
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	EB E0	7	53 76 63 34 52 65 71	Svc4Req
Std(2.0A)	E8 E0	7	53 76 63 31 52 65 71	Svc1Req
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	EA E0	7	53 76 63 33 52 65 71	Svc3Req
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	E9 E0	7	53 76 63 32 52 65 71	Svc2Req
Std(2.0A)	D8 E0	8	53 4D 50 31 53 54 41 ...	SMP1STAT
Std(2.0A)	D9 E0	8	53 4D 50 32 53 54 41 ...	SMP2STAT
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	E8 E0	7	53 76 63 31 52 65 71	Svc1Req
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	EB E0	7	53 76 63 34 52 65 71	Svc4Req
Std(2.0A)	EA E0	7	53 76 63 33 52 65 71	Svc3Req
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	E9 E0	7	53 76 63 32 52 65 71	Svc2Req
Std(2.0A)	D8 E0	8	53 4D 50 31 53 54 41 ...	SMP1STAT
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	D9 E0	8	53 4D 50 32 53 54 41 ...	SMP2STAT
Std(2.0A)	E8 E0	7	53 76 63 31 52 65 71	Svc1Req
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	EB E0	7	53 76 63 34 52 65 71	Svc4Req
Std(2.0A)	EA E0	7	53 76 63 33 52 65 71	Svc3Req
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	D8 E0	8	53 4D 50 31 53 54 41 ...	SMP1STAT
Std(2.0A)	DA E0	8	53 4D 50 33 53 54 41 ...	SMP3STAT
Std(2.0A)	E9 E0	7	53 76 63 32 52 65 71	Svc2Req
Std(2.0A)	ED E0	7	53 76 63 35 52 65 71	Svc5Req
Std(2.0A)	E8 E0	7	53 76 63 31 52 65 71	Svc1Req

Figure 6 Stroting data in server

5 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.

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the CITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1006) supervised by the NIPA(National IT Industry Promotion Agency.)

6 References

[1] H. Kopez, R. Obermaisser, C. El Salloum, B. Huber, "Automotive Software Development for a multi-core System-on-a-chip," in ICSE workshops SEAS'07, Minneapolis, 2007.

[2] C. Pinello, L. P. Carloni, A. L. sangiovanni-Vincentelli, "Fault-Tolerant Deployment of Embedded Software for Cost-Sensitive Real-Time Feedback-Control Applications," IEEE Design, Automation and Test in Europe Conference and exhibition, vol. 2, pp. 1164-1169, Febrary 2004.

[3] J.R. Pimentel, M. Salazar, "Dependability of Distributed Control system Fault Tolerant Units," in IECON 02, vol. 4, pp. 3164-3169, November 2002.

[4] Yeon Kyu Bong, Jeong Ki Yun, "The Dependability Analysis of LIN Network for Adaptive Front-Lighting System," in ISOCC'08, vol. 1, pp.425-428, November 2008.

[5] Kabsu Han, Yongseop Kwon, Wooyeon Kim, Jeonghun Cho, "Distributed Hierarchical Service Network for Automotive Embedded System", Proceedings of the International Conference on Information Network 2012, pp. 188-192, 2012.

[6] H. Schweppe, A. Zimmermann, D. Grilly, "Flexible In-Vehicle Stream Processing with Distributed Automotive Control Unis for Engineering and Diagnosis," in SIES' 08, vol. 1, pp. 74-81, June 2008.

[7] Juan M. Corchado, Javier Bajo, Dante I. Tapia, Ajith Abraham, "Using Heterogeneous Wireless Sensor Networks in a Telemonitoring System for Healthcare," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, No. 2, March 2010.

[8] Minwoo Jung, Jeonghun Cho, "Platform development for medical system in u-Health care environment", International Conference on Wireless Networks 2012, pp. 521-525, 2012.

Cooperative Communication in Free Space Optical Systems

Vineeta Dubey , D. Chadha , and Vinod Chandra
Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi-110016, India
vineetakhare@gmail.com,dchadha@ee.iitd.ernet.in,vchandra@ee.iitd.ernet.in

Abstract *- We have investigated the performance improvement of cooperative free space optical (FSO) communication over single input single output (SISO) system in this paper. Bit error rate (BER) analysis for gamma-gamma channel model with additive white Gaussian noise (AWGN) has been demonstrated for SISO and cooperative system as well where the performance improvement with different combining techniques in cooperative system has been depicted in this paper. Significant performance improvement is achieved in case of asymmetrical channel model where direct link is facing higher turbulence than the indirect link as compared to symmetrical channel model where all the links are considered to be at same turbulence level. We have shown very good BER performance by using amplify and forward (AF) scheme over decode and forward (DF) scheme for cooperative FSO systems.*

Keywords: *FSO, SISO, Cooperative communication, Decode and forward, Amplify and forward.*

1 Introduction

Free Space Optical (FSO) communication finds applications in last-mile access, back-haul for wireless cellular networks, fiber backup and disaster recovery. Unlimited bandwidth, low cost of installation and excellent security are attractive features of an FSO communication[1]. But there are also some problems with the usage of FSO systems such as rain, fog, dust, snow, turbulence and misalignment as well. To combat the effects of turbulence, the technologies such as MIMO (Multiple input multiple output) and relay networks have been used [2].

In radio frequency (RF) systems, the nature of broadcasting of signal is taken into account and the signal is sent through different relays to enhance system performance. These relays can be ordered in a serial or parallel fashion according to the need of user. Series combination (also known as multihop communication) is generally used to increase the range of communication whereas parallel combination (also known as cooperative communication) is meant for increasing the system performance [3]. Relay nodes can be chosen in the network as per the requirement of cooperation and hence it does not include any extra hardware requirement like multiple input multiple output system. FSO provides the facility of point to point communication only, so different modules of laser diodes (LD) and photo detectors (PD) are used to get the advantage of cooperative diversity.

In this paper, we have analyzed single relay assisted cooperative system using amplify and forward and decode and forward strategies with symmetrical and asymmetrical channel environments. We have used binary pulse position modulation (BPPM) for data modulation, Gamma-Gamma with AWGN as the atmospheric turbulent channel model and direct detection at the receiver end. Different combining techniques such as maximal ratio combining (MRC), equal gain combining (EGC) and selection combining (SC) are used at the receiver for comparison.

The paper is organised as follows. Section 2 discusses the system model and channel model of the FSO. Gamma-Gamma probability density function channel model is used based on the Kolmogrov theory. Section 3 deals with the block diagram of cooperative system . Further, the subsection give the details of PPM. Section 4 has the simulation results conclusion for SISO and cooperative system for symmetrical and asymmetrical channel models.

2 Atmospheric channel model

Optical channel is affected by parameters such as scattering and turbulence. Gamma-Gamma PDF closely models experimental results over low to high turbulence strengths and is most suitable for studying link performance parameters for slow fading conditions. Therefore, Gamma-Gamma model is used as channel model for both the direct and indirect paths.

The irradiance of optical field in Gamma-Gamma channel is defined as the product of two random processes, i.e. $I = IxIy$, where Ix arises from large scale turbulent eddies and Iy from small-scale eddies leading to the so-called Gamma-Gamma PDF, i.e.

$$f(I) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}} I^{\left(\frac{\alpha+\beta}{2}\right)-1} K_{(\alpha-\beta)}\left(2\sqrt{\alpha\beta I}\right)}{\Gamma(\alpha)\Gamma(\beta)} \qquad (1)$$

where $K_{\alpha-\beta}(.)$ is the modified Bessel function of the second kind of order α-β. Here, α and β are the effective number of small-scale and large scale eddies of the scattering environment given below, Γ is the gamma function.

These parameters can be directly related to atmospheric conditions according to

$$\alpha = \left(exp\left[\frac{0.49\sigma_R^2}{(1+1.11\sigma_R^{12/5})^{7/6}} \right] - 1 \right)^{-1} \quad (2)$$

$$\beta = \left(exp\left[\frac{0.51\sigma_R^2}{(1+0.69\sigma_R^{12/5})^{5/6}} \right] - 1 \right)^{-1} \quad (3)$$

where σ_R^2 is the Rytov variance given by

$$\sigma_R^2 = 1.23 C_n^2 k^{7/6} L^{11/6} \quad (4)$$

Here k is the optical wave number given by $k=2\pi/\lambda$; λ is the wavelength and C_n^2 is the atmospheric structure parameter [4].

3. Block Diagram

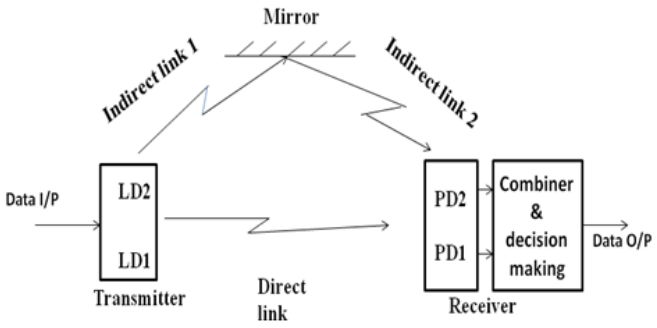


Fig.1: Block diagram of single-relay cooperative communication system for decode and forward scheme.

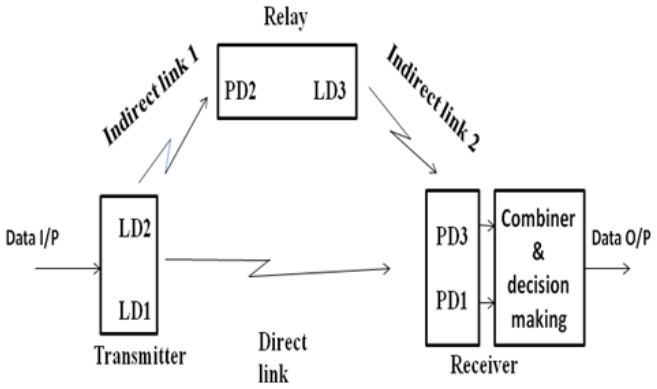


Fig.2: Block diagram of single-relay cooperative communication system for amplify and forward scheme.

Two practical cases of relay can be considered in FSO cooperative systems. In Fig.1, we consider a single relay cooperative DF relaying network with one transmitter, one receiver and one mirror which works as a reflector to provide the diverse path. The transmitter is equipped with two LDs; LD1 and LD2, pointing out in the direction of a destination node and a corresponding relay node, respectively.The same signal is sent through both the diodes and one direct and the other diverse beams are collected at destination end by using two PDs; PD1 and PD2 by using different combining techniques.

In the second case, one opto-electronic-opto (O-E-O) converter can be mounted at relay node which gives the advantage of amplify and forward.

In Fig.2, we consider a single relay cooperative DF relaying network with one transmitter, one receiver and one relay node which works as a transceiver. The transmitter is equipped with two LDs; LD1 and LD2, pointing out in the direction of a destination node and a corresponding relay node, respectively. The source node transmits the same signal to the relay and destination node. Relay node decodes, amplifies and retransmits the signal to the destination. The transmitted signal from the relay has same power as at LD2. The signals from source and relay are collected at PDs; PD1 and PD3, respectively and then processed by using different combining techniques.

As shown, we have one direct link and two indirect links in these systems. We consider two cases of the set- up in our study. In case-1, we have assumed symmetrical channel environment and in case-2, we consider the asymmetrical channel environment for direct and indirect links. Intensity modulated BPPM has been used as the modulation technique. At the receiver with direct detection we compare three types of combining techniques of the received signals. In MRC, the received signals are weighted with respect to their signal to noise ratios and then summed. In the case of EGC, the received signals are summed coherently with equal weights, and in SC the strongest signal is selected out of the two received signals.

3.1 Pulse Position Modulation

Proper choice of digital modulation techniques is very important with respect to the link design keeping in view the various parameters such as fog scintillation and scattering that govern its performance, limiting the FSO link. Restriction on laser power output due to safety concerns to human eye and skin [1] require power efficient modulation schemes for better range. Earlier studies for optical space communication have shown that Q-PPM is an energy-efficient and readily implemented modulation choice for optical communication. In Q-PPM, a signaling interval of length T_s is subdivided into Q slots, each of length $T = T_s / Q$, and a bit group comprising of $\log_2 Q$ bits is represented by a laser pulse which has duration equal to one slot. If a constant average power, P_r watts is received at the receiver which in this case is same as the peak power, then the received optical energy per symbol is $E_s = P_r T = P_r T_s / Q$ joules and is related to energy of the information bit as $E_s = E_b \log_2 Q$. It is advantageous to employ PPM over OOK (On-Off Keying) since no threshold detection at the receiver will then be

required. However, synchronization requirements at the receiver make it necessary to have an accurate clock. In terms of energy per bit ($E_s = E_b \log_2 Q$), larger Q decreases energy per bit for a fixed error probability. This efficiency comes at the expense of peak power. However, the peak power is limited and large spectrum occupancy and additional synchronization difficulties with large-alphabet PPM are the issues to be dealt with [5].

4. Results and Conclusion

Performance evaluation in terms of BER with respect to E_bN_o was carried out using ML detection in MATLAB by transmitting BPPM modulated data streams in blocks of 10,00,000 bits, using Gamma-gamma channel model, AWGN and different combining techniques.

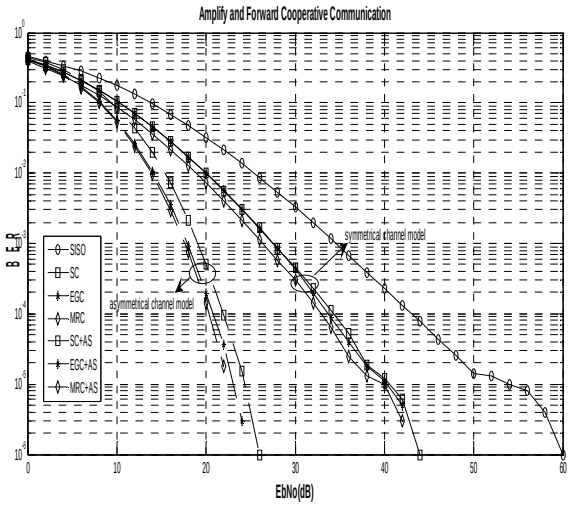


Fig.3 Comparative results of SISO and cooperative diversity with MRC, EGC and SC for symmetrical and asymmetrical channel models with decode and forward scheme.

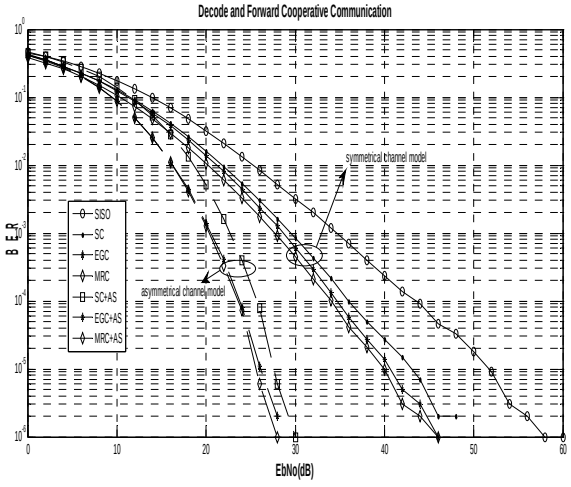


Fig.4 Comparative results of SISO and cooperative diversity with MRC, EGC and SC for symmetrical and asymmetrical channel models with amplify and forward scheme.

In Fig.3, we have shown the results for symmetrical channel model with DF relaying scheme, where all the three links (direct link and both the indirect links) are facing the same turbulence value (standard deviation=0.8)and asymmetrical channel model, where the direct link and both the indirect links are facing different turbulences. In case of asymmetrical channel model, standard deviations for the direct link and both the indirect links are 0.8 and 0.2, respectively. Link length of direct link is 1.414 Km and indirect links is 1 Km. From Fig.3, we conclude that the BER performance can be sufficiently improved by using cooperative diversity as compared to SISO. A diversity gain of 8 dB is obtained with SC over SISO system,while EGC and MRC provide the gain of 9 dB and 10 dB over SISO for BER of 10^{-5} in case of symmetrical channel model. In case of asymmetrical channel model, a diversity gain of 23 dB is obtained with SC,while EGC and MRC provide the gain of 25 dB and 26 dB over SISO for BER of 10^{-5} .

From Fig.4, we conclude that the BER performance can be sufficiently improved by using cooperative diversity as compared to SISO (Single input single output) by using AF relaying scheme. A diversity gain of 11 dB is obtained with SC over SISO system,while EGC and MRC provide the gain of 12 dB and 13 dB over SISO for BER of 10^{-5} in case of symmetrical channel model. In case of asymmetrical channel model, a diversity gain of 26 dB is obtained with SC,while EGC and MRC provide the gain of 27 dB and 28 dB over SISO for BER of 10^{-5} .

Comparative results show that cooperative diversity for symmetrical channel shows better performance as compared to SISO whereas cooperative diversity for asymmetrical channel shows better performance as compared to cooperative diversity with symmetrical channel model for AF and DF relay schemes. We can conclude, therefore, that if we use asymmetrical channel model, which indeed is a practical case with cooperative communication in FSO facing different atmospheric conditions, then we can obtain better performance by using any of the above mentioned diversity combining techniques. AF relaying scheme provides much better performance as compared to DF relaying for the cases of symmetical and asymmetrical channel model as well.

5 References

[1] David 1. T. Heatley, David R. Wisely, Ian Neild, and Peter Cochrane: "Optical Wireless: The Story so Far", IEEE Communications Magazine, December 1998, pp. 72-78.

[2] M. Safari and M. Uysal, "Relay-assisted Free-space optical communication", IEEE Trans. on wireless Commun., vol. 7, no. 12, pp. 5441-5449,Dec.2008

[3] J. N. Laneman and G. W.Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, Chicago, IL, Sept. 2000.

.
[4] Zhu and Joseph M. Kahn, "Free-space optical communication through atmospheric turbulence channels", *IEEE Trans. on Commun,* vol. 50.no.8,pp1293-1300,Aug 2002.

[5] Harinder Singh Sandhu, Thesis of Master Of Technology: "Power Efficient Free Space Optical Multiple Input Multiple Output Wireless Communication", *Department Of Electrical Engineering, Indian Institute Of Technology Delhi*, May 2008.

Train safety improvements using microwave networks

Pavel Rozsival, Jan Pidanic, Petr Dolezel, Pavel Bezousek
Department of Electrical engineering, University of Pardubice, Pardubice, Czech Republic
pavel.rozsival@upce.cz

Abstract - *Article describes use of radio network using 2.45GHz (ISM) frequencies for safety management and logistic in train transportation. Possible uses and expected properties of low cost network based on radio nod equipped train wagons are described. Design of such radio network based on NRF24L01 radios is outlined. This paper also describes testing scenarios for estimated parameter proofing. At the end of the paper are presented results from testing dependency of movement on reliability of communication and first results from real scenario tests. Data shows that concept of placing low cost radios on wagons can be easily used in various situations like radio identification even in high speeds between train and ID reader.*

Keywords: train; safety; RFID; NRF24l01

1 Introduction

Railway transportation is historically one of the most safe and reliable way of transport people and goods [1] [2].With increased traffic on railroads a lot of safety issues were solved and are still in solve. These days safety management in train transportation is one of the most developed, but still dealing with issues. Some errors are unpredictable as human being and mostly caused by human [3]. Most of the mistakes are easy to prevent or fix, like entering wrong track or direction. But some of them are still uncovered by any safety system.

One of the unsolved problems is train integrity detection for example. Usually most of automatic safety features is connected to locomotive, but if the train loses some wagons, it's hard to detect them on tracks, or especially in freight trains to even notice missing car. Another problem is identification of trains or train wagons, most of tracks and safety electronics are able to detect train on track, but only few are equipped for reading ID of train, and again ID is placed on locomotive only.

If all train vehicles including both locomotives and wagons are equipped with programmable radio NOD it can help to solve, or be a part of solution of most of these issues, and secondary help with other non-safety related demands.

2 Possible Solution

Our scenario uses radio nod on every train car used on selected tracks, around these tracks are radio base stations placed on important parts of track. Nods should be universally reconfigurable to act in different ways.

This system can work in more ways, to increase safety or comfort of train use. At first, if every nod contains specific ID code, it can be used for identification. In this case, it can help with train composing. Now the worker have to go to vehicle read the number, check and type in, other one is selecting manually and composing by the list. RFID based train cars can make this partially autonomous. It can help to detect presence of trains in parking places, train stations, maintains stations if equipped with ID readers. Also tracks equipped with readers can detect presence of every car. System equipped with readers among tracks can detect differences in cars passed through check points and detect train integrity loss. System can be also used for positioning of trains with dangerous or valuable freight or even freight itself.

NODs don't have to act only as ID radio tag, but can also contain sensors to measure conditions for fault prevention, or damaged goods complains and much more. Radio tags can also create networks and detect changes in neighbors. Networks can be configured by higher system, or can be made as self-configurable, but because of possible presence of nods that are not part of the train or low possibility to detect faulty NOD, it's better to use higher system to configure train car network. This network can easily detect neighbor loss and create integrity warning. [4]

Because of huge amount of train cars, radio nods should be as cheap as possible. Nods have to offer long live span. Because of train speed varying from 0-160 km/h (in the Czech Republic) system might be able to work up to 200 km/h. Also tags will be exposed to weather conditions, so they should withstand temperatures from -20 to 60°C, high humidity or running water contact.

3 System design

3.1 Radio network

Radio network is composed from nods and base stations. In our testing scenario, every carriage is equipped with 2 nods, one nod on every side of train car preventing radio shadow creating and increasing reliability by this doubling. Radios must be placed on the outer side of wagon due metallic (short wave radio proof) walls of trains. Side of under frame was chosen. Place is covered from the top, hidden from eyes and doesn't affect the profile of the train.

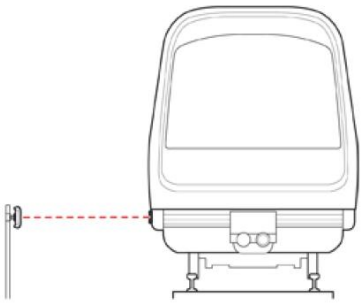


Figure 1: Tag/Reader placement

Rails in testing area are equipped with base stations collecting data on passing trains. Most of stations are equipped with GSM modem for data retransmit to central database for later processing.

3.2 Tag

Tag is radio NOD basically programmed to transmit unique ID for wagon identification.

Nod as itself is consisted of radio chip, low power microprocessor, battery and few discrete components.

Selection of radio components highly affects properties of whole system. Trying to keep hardware part as simple as possible, we have to use radio chip with high level of integration of all major radio parts and low support components count. As a compromise between simplicity and efficiency NRF24L01 from Nordic semiconductor was chosen.

The Nordic NRF24L01+ is a highly integrated, ultra-low power (ULP) 2Mbps RF transceiver IC for the 2.4GHz ISM (Industrial, Scientific and Medical) band. With peak RX/TX currents lower than 14mA, a sub µA power down mode, advanced power management, and a 1.9 to 3.6V supply range, the nRF24L01+ provides a true ULP solution enabling months to years of battery life from coin cell or AA/AAA batteries. The Enhanced ShockBurst™ hardware protocol accelerator offloads time critical protocol functions from the application microcontroller enabling the implementation of advanced and robust wireless connectivity with low cost 3rd-party microcontrollers. The Nordic nRF24L01+ integrates a complete 2.4GHz RF transceiver, RF synthesizer, and baseband logic including the Enhanced ShockBurst™ hardware protocol accelerator supporting a high-speed SPI interface for the application controller. No external loop filter, resonators, or VCO varactor diodes are required, only a low cost ±60ppm crystal, matching circuitry, and antenna. The nRF24L01+ comes in a compact 20-pin 4 x 4mm QFN package [5].

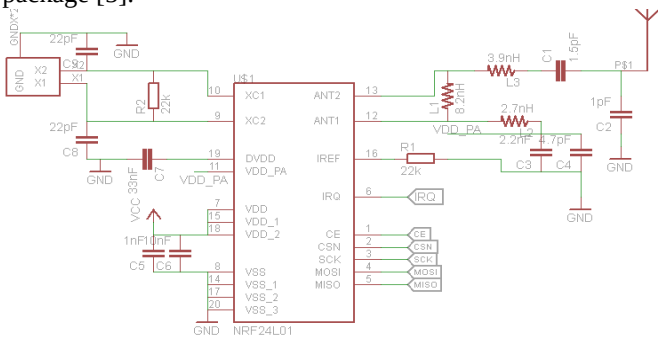


Figure 2: Radio chip connection diagram

Microprocessor selection is not as important as radio chip. Low power, low voltage and low cost are the only demands. Processor from Atmel AVR family was chosen, namely ATmega48PV. ATmega48 is the cheapest available processor from Atmel, it is important to choose PV suffixed parts. P is suffix for processors with picoPower technology with very low power consumption sleep modes. V stays for low voltage. Processors with V in name are able to run from 1.8-5.5 V compared to others that works in range from 2.7 to 5.5 volts. This processor offers plenty of peripheries that can be used for extending functionality of application. In our case only SPI port is used to connect radio chip and UART for debugging or connecting when the nod is used as transceiver.

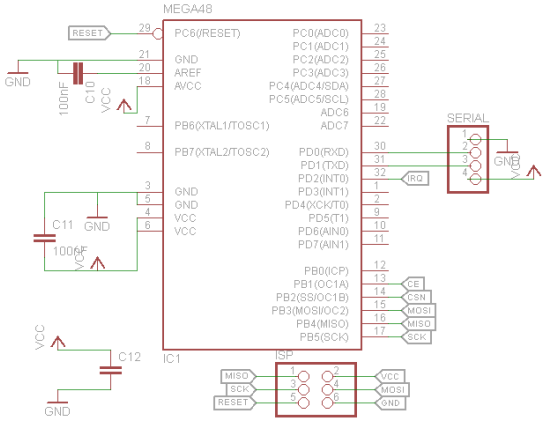


Figure 3: Microprocessor connection

Chip antenna solution was used for tags as space saving solution. For special purposes was developed nod with SMA connector for external antenna connection.

Battery selection must be compromise between size and capacity. Lifespan of tag should be at least 1 year. First iteration used CR2450 battery offering theoretical lifespan over 2 years. Testing in climatic box showed that battery based on Li-MnO_2 suffers from extreme capacity los under low temperatures. In other iterations CR14250 battery with Li-$SOCl_2$ mechanism were used. This battery is only slightly affected by temperature and offers 3 times higher capacity compared to CR2450, with only small size adding (tag with this battery is 5 mm higher).

For outside condition deployment, tag is fitted in Hammond1551 flanged or non-flanged 20x35x60 mm box.

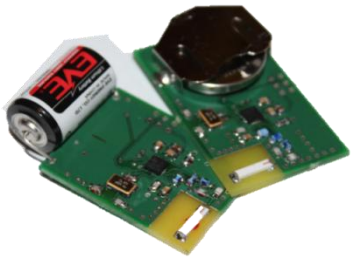


Figure 4: Nods

Figure 5: Boxed Nods

Due to high parasitic properties of component used in this solution, range of radio showed high variation in distance. Tag design was improved to 0402 sized components, increasing stability of design and decreasing size of board.

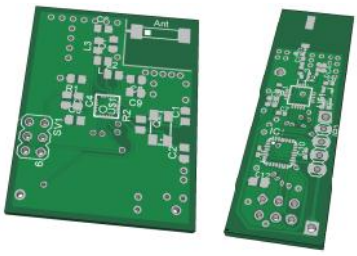


Figure 6: Board size comparsion

Size of the board was decreased to 15x45 mm with all components on one side and more peripherals taken out. Design upgrading to 0402 passive components showed huge improvement in stability and radio performance. Smaller PCB with components on one side decreased costs and production time.

Figure 7: Final tag

3.3 Base station

Base stations are special kind of NOD, usually consisted from tag electronics acting as transceiver and higher computer that offers better connectivity (serial, Ethernet…) or functionality (database, data logging, data analysis…). We are using 3 kinds of base stations. First one is mobile base station consisted from tag with external antenna connector and GSM modem for data harvesting and retransmitting in situation where other data network connection is not possible.

Figure 8: Reader with GSM modem

Second one is tag connected to 3.5" industrial PC. PC based computer PEB2737 offers all possible connections and programing versatility. Low cost demand for base stations is no longer needed. Amount of base stations needed compared to tags is negligible.

Figure 9: Base station with embeded PC

3.4 Principle of operation

NOD programmed as radio ID tag is acting like active RFID, which means it transmits unique ID. This can be done by two ways, transmitting ID every fixed moment, or on demand from reader. First option was selected for testing purposes. In this case whole tag can stay in sleep state with very low power consumption for most of the time. As a compromise cycle of 1 sec was chosen. It means that communication usually takes 1ms with 20-25 mA consumption and rest of second stays in sleep mode with around 6μA consumption. This enables tag lifespan for years. Communication is addressed and packet oriented with CRC enabled, using 4 bytes of data as ID.

Base stations acts as RFID readers, they are trying to collect ID from all passing trains and put the data into database. At the very beginning of experiment, base stations were made as self-powering, and hanged on power electricity wires holding posts. Base stations are always in active mode; this makes them more current hungry. Short time between battery change and attack of vandals made us move them in backup boxes that are along the rails, safely closed and providing power supply. All stations were equipped with GSM modem, because of lack of other networking option.

4 Pre deployment testing

Before the system was deployed on railway, basic parameter testing was done to determine behavior and some parameter tuning.

Measured parameters are power consumption of tags, communication distance (tag to base station), and effect of relative speed.

Power consumption was measured with scope meter connected to resistor placed between battery and tag. Measured consumption was varying around predicted values, around 6-7 μA for 998 ms and 20-30 mA for 1 ms, this means average power consumption is max 37 μA.

Maximum communication distance is also important. During radio identification on tracks, train can be passing at full speed, Czech limit is 160 km/h but with some reserve we wanted to guarantee up to 200 km/h. While the car is passing in speed of 200 km/h it will move 56 m every second. If we have reader with radial pattern antenna and tag communicating every 1s. We need reading distance at least 50 m to guarantee at least 1-2 readings by covering area of 100 m. For this test 5 Tags were picked and placed in open field with no obstacles in 600 m radius. Base station with ¼ wave antenna was moved until communication was lost. 1 way traffic from tags, 1 transmit per second was used, table 1 shows received vs. expected amount of data.

Table 1:Hits vs. expected ratio

reading/distance	5m	10m	30m	50m	60m
tag 4	100%	100%	100%	100%	96%
tag 5	13%	0%	0%	0%	0%
tag 7	100%	100%	97%	93%	67%
tag 11	100%	98%	100%	96%	74%
tag 14	100%	93%	91%	91%	87%

Table shows that most of NODs are able to communicate on distance greater than 50 m. Tag 5 vas found faulty and discarded.

Last test was relative speed test. This test was performed at the Hradec Kralove airport, where the nod was placed in the middle of tarmac and base station in the car. Car was passing in different speeds up to 180 km/h. Due the weather condition of that day, it was impossible safely reach 200 km/h. Positive ID count readings was logged.

- 140 km/h 4 hits
- 90 km/h 3 hits (returning car)
- 140 km/h 4 hits
- 120 km/h 2 hits (returning car)
- 140 km/h 3 hits
- 150 km/h 2 hits (returning car)
- 160 km/h 3 hits
- 180 km/h 2 hits (returning car)
- 160 km/h 2 hits
- 180 km/h 2 hits (returning car)

Car was moving in both directions, and the position of equipment was mirrored on return, this can change radio wave propagation. Setup of this test was reversed compared to real scenario, tag was placed outside and base station was placed in car. This setup was chosen because of heavy rain falls. Weather condition also made impossible to safely drive car in higher speeds and keep this speed for 100 m around the tag.

This test proved presumption of tag requirements and setting. 50 m+ range, 1 ID per second is compromise setting for still reliable setup. If higher ratio of possible reading is needed to increase probability of safety ID, repeating period can be shortened but for the price of battery live drop, almost all battery consumption is made when on air, this means doubling communication will cut battery live to half. This experiment also proved minimal impact of Doppler frequency shift calculated to 133-570 Hz [6].

5 Preliminary testing results

For preliminary testing one train car was marked with tags and two readers were deployed. Two tags were used, one tag per side of the train car.

Figure 10: Tag placement

Base stations were placed on both sides of tracks in different locations. Reader 1 was mounted 2 m from tracks and battery operated; antenna was approximately in same high as tag on wagon. Reader 2 was mounted in rail house with electronics close to railroad crossing and this reader was live power operated for the price of 11 m distance through plastic composite wall, also antenna was 1-2 m higher than ideal position.

Figure 11: Position of reader 1

Figure 12: Reader 1 placed

Figure 13: Position of reader 2

Figure 14: Reader situation

Figure 15: Reader 2 placed

Table 2: Hits per passes of the train

Tag no.	reader 1	reader 2	passes	Near side ratio	Far side ratio
13	17	10	17	100,00%	58,80%
10	13	15	17	88,20%	76,50%

You can see impact (table 2) of reader placement on system properties, from later signal propagation measurement in place, reliability can be increased by using antenna with shaped pattern. You can also see that it's hard to communicate with module covered by body of the train (far side reading). Special pattern antenna based on collected data is under development.

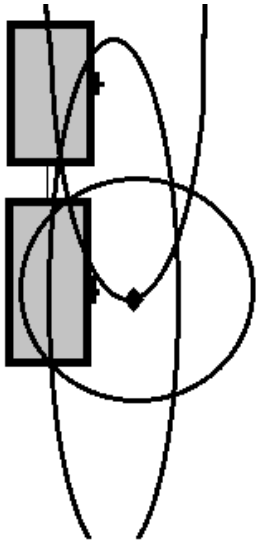


Figure 16: Train with tags (marks on sides); coverage by radial pattern antenna (circle) and modified pattern antenna (eliptic) antenna and directional antena (U shape); antenna position is in the middle of the circle.

6 Full scale test

Previous testing showed proof of concept of our design. Some minor bugs were fixed and tag design was upgraded. Before full test placement speed test was repeated. This time with multiple tags placed and condition good enough to reach speed of 205 km per hour. At least 2 positive ID logged were taken as positive identification.

- 190 km/h 4/4 logged
- 185 km/h 4/4 logged
- 188 km/h 3/4 logged (possible shadowing)
- 205 km/h 4/4 logged
- 205 km/h 4/4 logged

After this recheck preparation of full scale long time testing took place. First goal was to find appropriate tracks and positions for reader placement with heavier train traffic. After discussion main corridor was chosen and places for readers were found. Because of long time testing expectation, places with power supply and GSM signal coverage was needed.

First reader was placed in Pardubice/Slovany in train security booth 8 meters from tracks. Trains are passing in almost full speed on two parallel tracks.

Figure 17: Reader 1 placement

Second reader was placed close to Uhersko in security booth only 3 meters from tracks.

Figure 18: Reader 2 placement

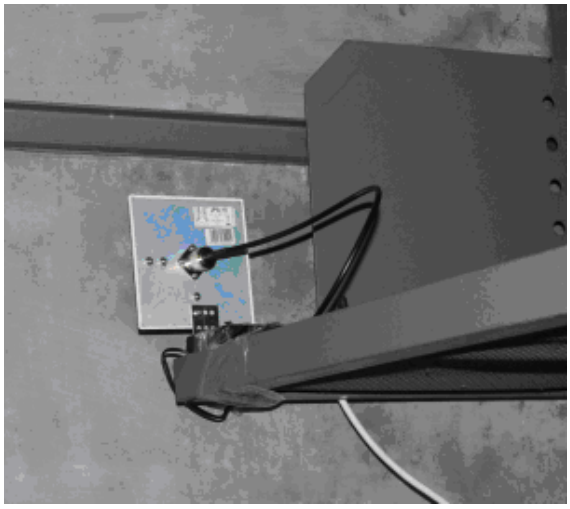


Figure 19: Reader 2 antenna placement

Second and third reader uses concept of directional antenna pointing in sharp angle with tracks. Panel antenna with 60 degrees pattern was used.

Figure 20: Reader situation

Third Reader was placed in Sedlistka close to the railway station in train security and driving house. It's the only concrete house used.

Figure 21: Reader 3 situation

One reader was left on track used in preliminary testing to cover trains that are not returning directly but circling around. Position of antenna was rearranged in correct high.

In this time tags are being placed on cars, and only partial amount of them is currently in process. But until now all of them are working perfectly. Even in accidental situation when tags were transported by super city train travelling mostly in full speed around 160 km/h (speed showed by infotainment system in train) through installed area 12, 9 and 13 from 20 were identified when passing around reader 1, 2, 3. Tags were placed in static dissipative bag in luggage inside of the train.

Collected data are stored on server and displayed on web server application. Application is accessible online on http://vlaky.eparo.cz. In this time it is too soon to create statistic report but at least we can say it looks like it works.

7 Conclusion

This article shows the possibilities of radio network used for safety and logistics management. Preliminary testing shows that covering trains by radio network can be used as a safety improvement, but mostly as a support not as the only solution. But for the price less than 10 USD for prototype per radio nod (4-5 USD in larger volumes) it's still huge improvement compared to price paid.

Long time testing should last for one year to show efficiency, reliability and security of this approach.

Questionable is radio band selection 2.4 GHz offers fast data transfer, leading to short time to ID, sub GHz offers higher range with lower power. Interesting option is 5GHz band but with lack of selection of integrated radios. Ultra wide selection of various integrated radios in 2.4 GHz kept us in this band.

Acknowledgment

The research was supported by the postdoctoral project Strengthening of Research and Development Teams at the University of Pardubice No. CZ.1.07./2.3.00/30.0021 and by the Czech Ministry of Industry and Trade project No. FR-TI1/084. At this point I want to thanks for collaboration and help to DKV (Railway Car Depot) Pardubice with tag placement and SZDC (Railway Infrastructure Administration) Pardubice with reader placement.

8 References

[1] Wikipedia, "Aviation safety," 10 april 2013. [Online]. Available: http://en.wikipedia.org/wiki/Aviation_safety. [Accessed 2013 april 2013].

[2] R. Ford, "INFORMED SOURCES," 2010. [Online]. Available: http://dspace.dial.pipex.com/town/square/ca14/ALYCIDON%20RAIL/INFORMED%20SOURCES%20ARCHIVE/INF%20SRCS%202000/Informed%20Sources%2010%202000.htm. [Accessed 10 april 2013].

[3] "Lists of rail accidents," 10 april 2013. [Online]. Available: http://en.wikipedia.org/wiki/Lists_of_rail_accidents. [Accessed 10 april 2013].

[4] H. Scholten, R. Westenberg and M. Schoemaker, "Sensing Train Integrity," in *Sensors, 2009 IEEE*, Chrischurch, New Zeland, 2009.

Parintins Smart Grid Project Backhaul

Carlos Henrique Rodrigues de Oliveira and José Carlos de Medeiros
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 / jose.medeiros@eletrobras.com

***Abstract*—Smart Grid solutions are being driven by the desire for more efficient energy usage worldwide. Nowadays Smart Grid communications network is still a heterogeneous network based on many different standards. This paper describes the Smart Grid backhaul deployed in Parintins, city of Brazil, based on network topology presented in [1].**

***Index Terms*— Backhaul, Smart Grid, Wi-Fi.**

I. Introduction

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 [2]. 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 recently deployed.

II. 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.

III. Backhaul topology

In Figure 1 is presented the backhaul topology deployed in April 2013.

Figure 1. Backhaul topology

The backhaul topology consists of four mesh equipments each one comprises two radios (2.4 GHz and 5.8 GHz) to cover the entire city.

The backhaul network is a layer to transport data from HAN (Home Area Network), FAN (Field Area Network) and NAN (Neighborhood Area Network) of different sort of services incluiding AMI (Advanced Metering Infrastructure), DA (Distribution Automation), DR (Demand Response), videomonitoring and local government access.

[1] This work was supported by the Smart Grid Parintins Project of Eletrobras - Brazil.

IV. BACKHAUL INFRASTRUCTURE

The backhaul infrastructure was structured by one tower in the substation and tree poles deployed in strategic points (Aninga Road, Macurani Road and Itacoatiara Street) of the city to facilitate line of sight, physical access and power suply.

The poles were made of fiber glass and the size of the tower and the poles is 30 m (net height). It was installed a protection structure to allow secure installation and future maintainance.

Figure 2 shows the pole photo mounted in the ground.

Figure 2. Pole photo mounted in the ground

Figure 3 shows the pole photo installed in one of the three points (Macurani Road).

Figure 3. Pole photo installed in the Macurani Road

V. RADIO LINKS

The master mesh equipment was installed in the subestation tower. It has three actived interfaces two radios (2.4 GHz and 5.8 GHz) and one gigabit ethernet port connected with the data center switch. Figure 4 shows the established radio link between Aninga pole and subestation tower.

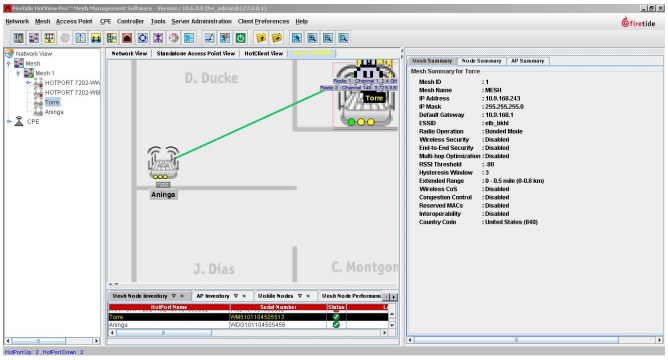


Figure 4. Established radio link between Aninga pole and subestation tower

Figure 5 shows the radio 1 (2.4 GHz) and radio 2 (5.8 GHz) neighbor statistics received in the Aninga mesh equipment from tower mesh equipment in a radio link of 4.36 km. The 2.4 GHZ radio 1 presented a measured RSSI of -66

dBm and a measured SNR of 30 dB. The 5.8 GHZ radio 2 presented RSSI of -79 dBm and SNR of 17 dB.

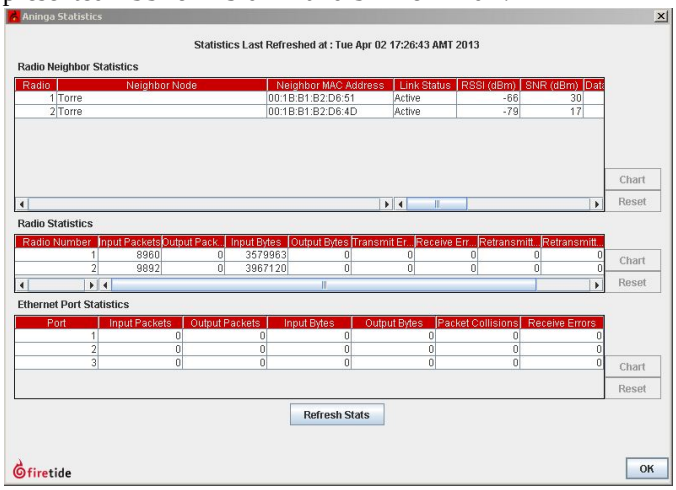


Figure 5. Radio 1 (2.4 GHz) and radio 2 (5.8 GHz) neighbor statistics received by Aninga mesh equipment

Figure 6 shows the coverage prediction result to the 2.4 GHz radio 1 (RX level of -65.2 dBm) very close to the measured RSSI (-66 dBm).

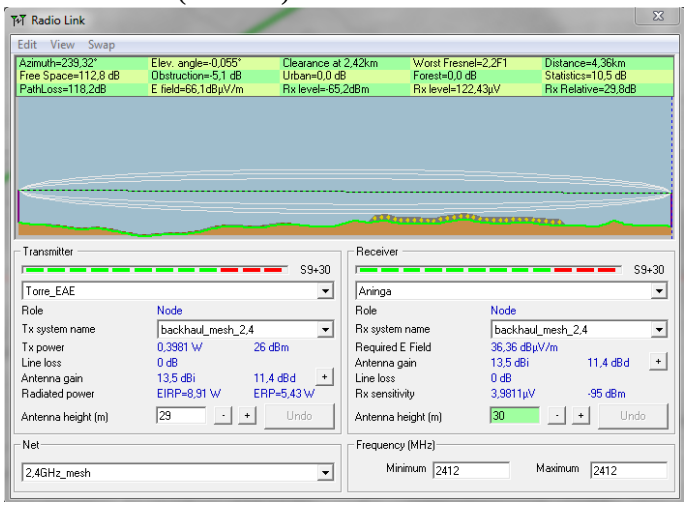


Figure 6. Coverage prediction result to the 2.4 GHz radio 1

Figure 7 shows the coverage prediction result to the 5.8 GHz radio 2 (RX level of -78.8 dBm) very close to the measured RSSI (-79 dBm).

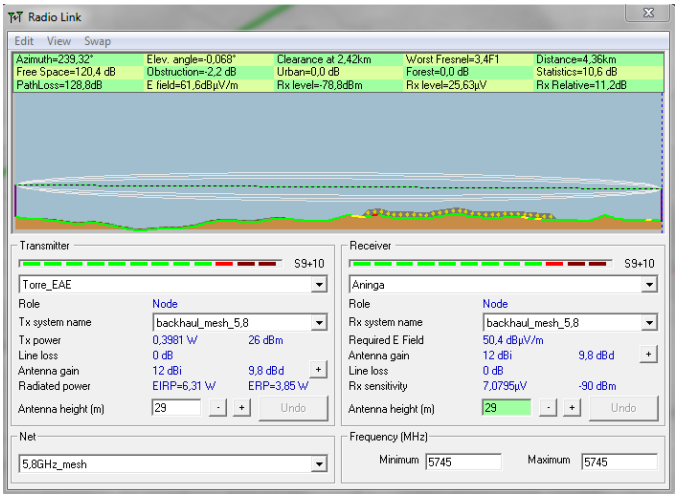


Figure 7. Coverage prediction result to the 5.8 GHz radio 2

Figure 8 shows the radio 1 (2.4 GHz) and radio 2 (5.8 GHz) neighbor statistics received in the tower mesh equipment from Aninga mesh equipment in a radio link of 4.36 km. The 2.4 GHZ radio 1 presented a measured RSSI of -71 dBm and a measured SNR of 25 dB. The 5.8 GHZ radio presented RSSI of -77 dBm and SNR of 19 dB.

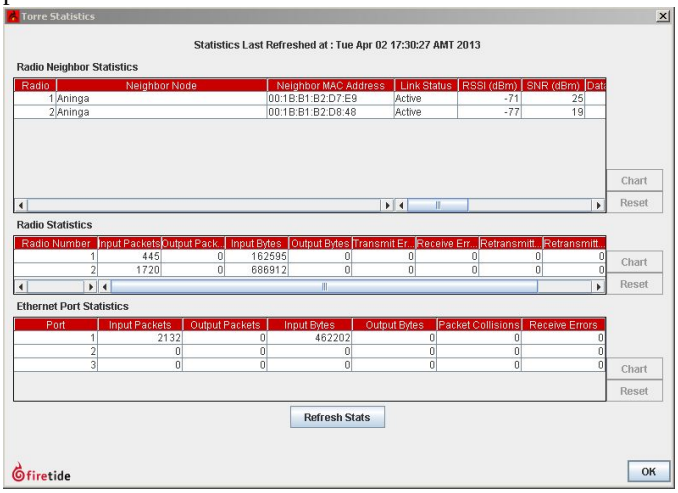


Figure 8. Radio 1 (2.4 GHz) and radio 2 (5.8 GHz) neighbor statistics received by tower mesh equipment

Figure 9 shows the established radio link between Macurani pole and subestation tower.

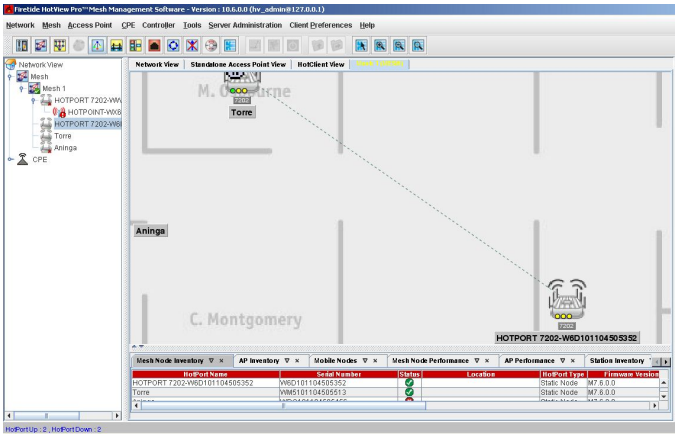


Figure 9. Established radio link between Macurani pole and subestation tower

Figure 10 shows the radio 1 (2.4 GHz) link statistics of the Macurani and tower mesh equipments in a radio link of 2.38 km. The 2.4 GHZ radio 1 of the Macurani mesh equipment presented a measured RSSI of -53 dBm and a measured SNR of 43 dB. The 2.4 GHZ radio 1 of the tower mesh equipment presented a measured RSSI of -56 dBm and a measured SNR of 40 dB.

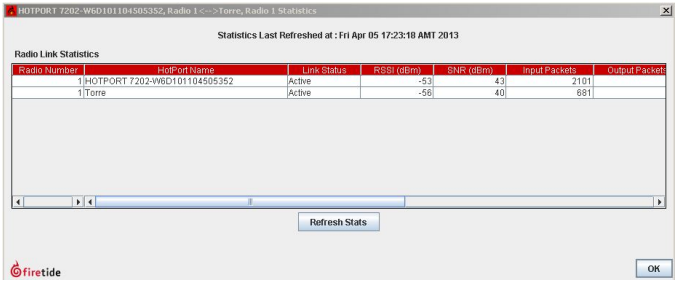


Figure 10. Radio 1 (2.4 GHz) link statistics of the Macurani and tower mesh equipments

Figure 11 shows the radio 2 (5.8 GHz) link statistics of the Macurani and tower mesh equipments in a radio link of 2.38 km. The 5.8 GHZ radio 2 of the Macurani mesh equipment presented a measured RSSI of -68 dBm and a measured SNR of 28 dB. The 5.8 GHZ radio 2 of the tower mesh equipment presented a measured RSSI of -67 dBm and a measured SNR of 29 dB.

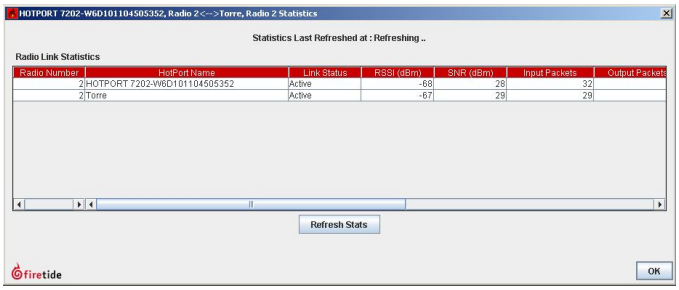


Figure 11. Radio 2 (5.8 GHz) link statistics of the Macurani and tower mesh equipments

VI. THROUGHOUT

The measured row throughout, using iperf tool, of 608 Mbits/sec in the gigabit ethernet port of the master mesh equipment is shown in Figure 12 corresponding half of it to the 2.4 GHz radio and other one to the 5.8 GHz radio.

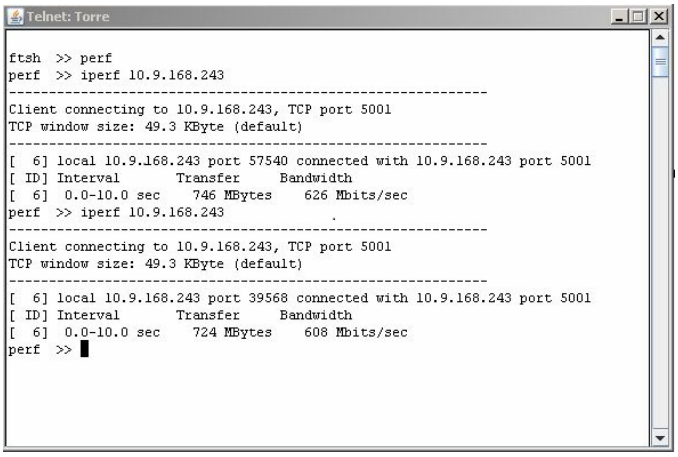


Figure 12. Measured row throughout of 608 Mbits/sec

VII. CONCLUSIONS

In the Parintins Smart Grid Project a mesh backhaul in unlicensed spectrum represented a good strategy considering that: a) The solution cost is much cheaper and it occupies much lesser space than a 4G cellular infrastructure; b) 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.

The coverage prediction was validated by the measured results.

REFERENCES

[1] Oliveira, C. H. R., and Fonseca, E. F. "Wi-Fi Technology in the Smart Grid Backhaul" The 2011 International Conference on Wireless Networks.
[2] http://www.parintins.com/docs/parintins/?p=historiadeparintins

Yet Another Look at the Problem of Detection of Unknown Deterministic Signals over κ–μ and η–μ Fading Channels

A. Olaluwe and 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 present an alternative moment generating function (MGF) method for computing the averaging detection probability over stochastic fading channels. In contrast to the existing approaches based on either the circular contour integral or infinite series representations for the generalized Marcum *Q*-function, $Q_u(.,.)$. We exploit an exponential-type integral for $Q_u(.,.)$ to facilitate the statistical averaging task over channels with independent and non-identical fading statistics. We have chosen κ-μ and η-μ because it provides a unification of a variety of statistical models. Our analytical framework yields prior results with appropriate choice of the fading severity indices. Selected numerical results are also provided for the receiver-operating characteristic (ROC) of Square-law Selection (SLS) and Square-law Combining (SLC) diversity energy detectors over κ-μ and η-μ channels.**

Keywords— **Energy detection, κ-μ and η-μ Fading channels, Square-law selection, Square-law combining,**

I. INTRODUCTION

Detection probability (Pd), false alarm probability (Pf) and missed detection probability (Pm = 1 − Pd) are the key measurement metrics that are used to discuss the performance of an energy detector by plotting the receiver operating characteristics (ROC) or complementary ROC curve. The analytical formulation of detection of an unknown deterministic signal in the presence of additive white Gaussian noise (AWGN) is found in [1] for a flat and band-limited Gaussian channel. Based on the results in [1], Pd and Pf have been derived in closed-form over AWGN channel in [2]. The ROC analysis for Rayleigh, Rice and Nakagami fading channels is discussed in [4] and [2] as two independent works with different analytical approaches. Further, energy detection under different diversity receptions such as maximal ratio combining (MRC), selection combining (SC), switch-and-stay combining (SSC), square-law combining (SLC), square-law selection (SLS) and equal gain combining (EGC) is analysed in [2], [6], [7]. Signal detection is analysed for K and K_G fading models in [8] for the case of multipath fading and the shadowing effect.

The η-μ and κ-μ distributions are more general physical fading models because they involve other fading statistics as special cases [10] [11]. The η-μ fits very well for non-line-of-sight applications and it yields Nakagami-*m* and Nakagami-*q* as special cases. The κ-μ fits line-of-sight applications properly and it involves Nakagami-*m* and Nakagami-*n as* special cases.

In this paper, we analyse the performance of an energy detector under the κ-μ and η-μ fading channel using [13, eq. (9)] expressed in only exponential terms. This enables us to apply MGF approach to solve the detection problem. The average detection probabilities are re-expressed in closed-form using MGF, for cases with and without diversity reception and such as SLS and SLC. This greatly simplifies the problem as we average over the chosen fading statistics.

The rest of the paper is arranged as follows: the system model is discussed in Section II. This features binary hypothesis testing, the channel model, MGF derivative approach, and the exponential-type integral form. Section III deals with square-law diversity combining. Section IV is an expose on square-law selection. Section V and VI feature the computational results and conclusion respectively.

II. SYSTEM MODEL

A. BINARY HYPOTHESIS TESTING

The detection of the existence of the unknown deterministic signal $s(t)$ by the receiver is treated as a binary hypothesis test shown in (1) [1],

$$y(t) = \begin{cases} n(t) & : H_0 \\ hs(t) + n(t) & : H_1 \end{cases} \quad (1)$$

where $s(t)$ is the unknown deterministic signalling waveform, $n(t)$ is the noise waveform (white Gaussian random process), H_0 is the hypothesis 0 (i.e., no $s(t)$ present) and H_1 is the hypothesis 1 (i.e., $s(t)$ present).

Therefore, a sample from noise process n_i is a Gaussian random variable with zero mean and $N_{01}W$ variance; $n_i \sim \aleph(0, N_{01}W)$ [1], where N_{01} is one sided noise power spectral density, W is one-sided bandwidth.

Thus the probability of detection (P_d) and the probability of false alarm (P_f) of an unknown deterministic signal in additive white Gaussian noise (AWGN) is as given by [19].

$$P_d = Q_u\left(\sqrt{2\gamma}, \sqrt{\lambda}\right) \quad (2)$$

$$P_f = \frac{\Gamma\left(u, \frac{\lambda}{2}\right)}{\Gamma(u)} \quad (3)$$

This work is supported by funding from the National Science Foundation (0931679 & 1040207)

where $u = TW$ is the time-bandwidth product, λ is the energy detection threshold, $Q_u(.,.)$ is the u^{th} order generalised Marcum *Q*-function and $\Gamma(.,.)$ is the upper incomplete gamma function which is defined by $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}$ and λ is the decision threshold.

B. CHANNEL MODEL OF η-μ AND κ-μ DISTRIBUTIONS

The MGF of the instantaneous signal-to-noise ratios (SNR) of η-μ and κ-μ distributions given by [12] respectively are:

$$M_{\gamma_{\eta-\mu}}(s) = \left(\frac{4\mu^2 h}{(2(h-H)\mu + s\Omega)(2(h+H)\mu + s\Omega)}\right)^{\mu} \quad (4)$$

and

$$M_{\gamma_{\kappa-\mu}}(s) = \left(\frac{\mu(1+\kappa)}{\mu(1+\kappa)+s\Omega}\right)^{\mu} \exp\left(\frac{-s\mu\kappa\Omega}{\mu(1+\kappa)+s\Omega}\right) \quad (5)$$

Note that (5) is obtained by simplifying algebraically the power of the exponential term in the usual MGF expression for κ-μ distribution.

The behaviour of η-μ distribution is hinged upon η which is also determines the values of two additional quantities *H* and *h*. Thus partitioning this behaviour into two formats.

[a] Format 1: The range of values of η is $0<\eta<\infty$ where η according to [10] is the scattered-wave power ratio between the quadrature and in-phase components of each multipath cluster. Thus $h = \frac{2+\eta^{-1}+\eta}{4}$ and $H = \frac{\eta^{-1}-\eta}{4}$ when $H \geq 0$, then $0<\eta\leq 1$ and when $H\leq 0$ then $0<\eta^{-1}\leq 1$. It is common to express the ratio $\frac{H}{h} = \frac{1-\eta}{1+\eta}$

[b] Format 2: The range of values of η is $-1<\eta<1$. Thus $h = \frac{1}{1-\eta^2}$ and $H = \frac{\eta}{1-\eta^2}$. Also when $H\geq 0$ then $0\leq\eta<1$ when $H\leq 0$ then $-1<\eta\leq 0$. Likewise we can express the ratio $\frac{H}{h} = \eta$.

Consequently, it can be shown with minimum algebraic manipulation that when μ = 0.5, $q^2 = \eta$ (in the case of format 1) or $q^2 = \frac{1-\eta}{1+\eta}$ (in the case of format 2) then η-μ distribution results to Nakagami-*q* or Hoyt (*q* is the Hoyt fading index). When μ = 0.5, η=1 (in the case of format 1) or η=0 (in the case of format 2) then (4) results to MGF of Rayleigh fading. Also, when $\mu = m/2$, $\eta\rightarrow 1$ (format 1) or $\eta\rightarrow 0$ (format 2) then (4) becomes MGF of Nakagami-*m* (*m* being Nakagami fading severity index) fading distribution.

Whereas merely observing (5) reveals that it becomes the MGF of Rayleigh fading; when μ = 1, κ = 0, MGF of Rice fading; when μ = 1, $\kappa = K$, where *K* is the Rice parameter, and MGF of Nakagami-*m*; when μ = m, κ = 0.

C. MGF DERIVATIVE APPROACH

Utilizing the canonical series representation of the generalized Marcum Q-function of real order using [17, eq. (7)],

$$Q_u(a,b) = 1 - \sum_{k=0}^{\infty}\left(\frac{a^2}{2}\right)^k \frac{e^{-\frac{a^2}{2}}}{k!}\frac{G\left(u+k,\frac{b^2}{2}\right)}{\Gamma(u+k)} \quad (6)$$

harnessing the Laplace transforms property of a derivative, given $a = \sqrt{2\gamma}$ and $b = \sqrt{\lambda}$, it can be shown that the average probability of detection is

$$\begin{aligned}\overline{P_d} &= 1 - \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G\left(u+k,\frac{b^2}{2}\right)}{\Gamma(u+k)}\int_0^{\infty}\left(\frac{a^2}{2}\right)^k e^{-\frac{a^2}{2}} f_\gamma(\gamma)\,d\gamma \\ &= 1 - \sum_{k=0}^{\infty}\frac{1}{k!}\frac{G\left(u+k,\frac{\lambda}{2}\right)}{\Gamma(u+k)}\int_0^{\infty}\gamma^k e^{-\gamma} f_\gamma(\gamma)\,d\gamma \quad (7) \\ &= 1 - \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\frac{G\left(u+k,\frac{\lambda}{2}\right)}{\Gamma(u+k)}\varphi_\gamma^k(s)\Big|_{s=1}\end{aligned}$$

where $\phi_\gamma^k(s) = \frac{\partial^k \phi_\gamma(s)}{\partial s^k}$ and $\phi_\gamma(s)$ denotes the MGF of channel SNR for a specified stochastic channel. Moreover, the kth derivative of the MGF of SNR for η-μ and κ-μ distributions are derived using the result in Appendix A (by substituting appropriately for the values of parameters A, B, and C). The resulting expressions are tabulated alongside other channel models' results in Table I for L-branch i.i.d SLC receiver.

It is clear from comparison of (5) with (A.1) that $A = \frac{\mu(1+\kappa)}{\Omega}$, $B = \mu\kappa$, and $C = \mu$ Hence, we can write the kth derivative of (5) utilizing (A.5), viz.,

$$\begin{aligned}\phi_{\gamma_{\kappa-\mu}}^k(s) &= \frac{(-\Omega)^{-k}\,k!\,(L\mu+\kappa-1)!\,(\mu\kappa(1+\kappa))^{L\mu}}{(\mu(1+\kappa)+s\Omega)^{L\mu+k}} \\ &\times \exp\left(-\frac{sL\mu\kappa\Omega}{\mu(1+\kappa)+s\Omega}\right)\sum_{i=0}^{k}\frac{1}{i!(k-i)!(L\mu+i-1)!}\left(\frac{L\mu^2\kappa(1+\kappa)}{\mu(1+\kappa)+s\Omega}\right)^i\end{aligned} \quad (8)$$

In order to evaluate the kth derivative for (4) we invoke Leibnitz differentiation rule:

$$\frac{\partial^k}{\partial s^k}[u(s)][v(s)] = \sum_{i=0}^{k}\binom{k}{i}\frac{\partial^i}{\partial s^i}[u(s)]\frac{\partial^{k-i}}{\partial s^{k-i}}[v(s)] \quad (9)$$

The right hand side of (9) contains the product of ith- and (k-i)th-order derivatives of the of u(*s*) *and* v(*s*) respectively summed over the *i* = 0 to *k* while being multiplied by the binomial coefficient $\binom{k}{i} = \frac{k!}{(k-i)!}$. But we must first rewrite (4) as a product of two terms, which individually resembles the special case (A.6).

TABLE 1: MGF OF SNR WITH L-BRANCH I.I.D SLC RECEIVER AND ITS K^{TH} ORDER DERIVATIVE

Channel Model	**MGF of Combiner Output SNR** $\phi_\gamma(s)=\int_0^\infty f(\gamma)e^{-s\gamma}d\gamma$	**k^{th} order Derivative of the MGF of SNR**, $\phi_{SLC}^{(k)}(s)$
Rayleigh	$(1+s\Omega)^{-L}$	$\frac{(-\Omega)^k\Gamma(L+k)}{(m+s\Omega)^{L+k}\Gamma(L)}$
Nakagami –q $b=\frac{1-q^2}{1+q^2}$	$\left(1+2s\Omega+\frac{4s^2\Omega^2q^2}{(1+q^2)^2}\right)^{-1/2}$	$\frac{k!}{\left[\Gamma(L/2)\right]^2\sqrt{[1+s\Omega(1-b)]^{L/2}[1+s\Omega(1+b)]^{L/2}}}\left[\frac{-\Omega(1+b)}{1+s\Omega(1+b)}\right]^k \times\sum_{w=0}^{k}\frac{\Gamma(L/2+w)\Gamma(L/2+k-w)}{[w!(k-w)!]^2}\left[\frac{(1+b)[1+s\Omega(1+b)]}{(1+b)[1+s\Omega(1-b)]}\right]^w$
Nakagami–n (Rice: K=n^2)	$\left(\frac{1+K}{1+K+s\Omega}\right)^L\exp\left(\frac{-KLs\Omega}{1+K+s\Omega}\right)$	$\frac{(-\Omega)^k k!(L+k-1)!(1+K)^L}{(1+K+s\Omega)^{L+k}}\exp\left(\frac{-sKL\Omega}{1+K+s\Omega}\right)\sum_{i=0}^{k}\frac{1}{i!(L+i-1)!(k-i)!}\left(\frac{KL(1+K)}{1+K+s\Omega}\right)^i$
Nakagami-m	$(1+s\Omega/m)^{-mL}$	$\frac{(-\Omega)^k m^{mL}\Gamma(mL+k)}{(m+s\Omega)^{mL+k}\Gamma(mL)}$
κ-μ	$\left(\frac{\mu(1+\kappa)}{\mu(1+\kappa)+s\Omega}\right)^{L\mu}\exp\left(\frac{-sL\mu\kappa\Omega}{\mu(1+\kappa)+s\Omega}\right)$	$\frac{(-\Omega)^{-k}k!(L\mu+\kappa-1)!(\mu\kappa(1+\kappa))^{L\mu}}{(\mu(1+\kappa)+s\Omega)^{L\mu+k}}\exp\left(-\frac{sL\mu\kappa\Omega}{\mu(1+\kappa)+s\Omega}\right) \times\sum_{i=0}^{k}\frac{1}{i!(k-i)!(L\mu+i-1)!}\left(\frac{L\mu^2\kappa(1+\kappa)}{\mu(1+\kappa)+s\Omega}\right)^i$
η-μ	$\left(\frac{4\mu^2h}{(2(h-H)\mu+s\Omega)(2(h+H)\mu+s\Omega)}\right)^{L\mu}$	$\frac{\left(4\mu^2h\right)^{L\mu}(-\Omega)^{-k}k!\left(h^2-H^2\right)^{L\mu-1}}{[(L\mu-1)!]^2} \times\sum_{i=0}^{k}\frac{(L\mu+i-1)!(L\mu+k-i-1)!}{i!(k-i)!(2(h-H)\mu+s\Omega)^{L\mu+i}(2(h+H)\mu+s\Omega)^{L\mu+k-i}}$

$$M_{\gamma_{\eta-\mu}}(s)=\left(\frac{C_O}{C_1C_2}\right)^\mu\left(\frac{C_1}{C_1+s}\right)^\mu\left(\frac{C_2}{C_2+s}\right)^\mu \tag{10}$$

where the first term in (10) constitutes constants defined as follows; $C_o=\frac{4\mu^2h}{\Omega^2}$, $C_1=\frac{2(h-H)\mu}{\Omega}$, and $C_2=\frac{2(h+H)\mu}{\Omega}$.

Hence, we evaluate i[th]and (k-i)[th]derivatives of u(s) and v(s) respectively and substitute in (9), where

$u(s)=\left(\frac{C_1}{C_1+s}\right)^{L\mu}$ and $v(s)=\left(\frac{C_2}{C_2+s}\right)^{L\mu}$ using (A.7) and after little simplification we obtain,

$$\phi_{\gamma_{\eta-\mu}}^k(s)=\frac{\left(4\mu^2h\right)^{L\mu}(-\Omega)^{-k}k!\left(h^2-H^2\right)^{L\mu-1}}{[(L\mu-1)!]^2} \times\sum_{i=0}^{k}\frac{(L\mu+i-1)!(L\mu+k-i-1)!}{i!(k-i)!(2(h-H)\mu+s\Omega)^{L\mu+i}(2(h+H)\mu+s\Omega)^{L\mu+k-i}} \tag{11}$$

D. EXPONENTIAL-TYPE INTEGRAL FORM FOR $Q_u(a,b)$

Using the contour integral representation for the generalized Marcum Q-function, [19, p. 885]

$$Q_u(a,b)=\frac{1}{2\pi j}\oint_\Gamma\frac{e^{g(z)}}{z^u(1-z)}dz \tag{12}$$

where $g(z)=\frac{a^2}{2}\left(\left(1/z\right)-1\right)+\frac{b^2}{2}(z-1)$. If we use $z=e^{jq}$ in (12) with the contour in figure 1 we obtain an exponential integral representation for $Q_u(a,b)$ viz.,

$$\begin{aligned}Q_u(a,b)&=\frac{1}{2\pi j}\oint_\Gamma\frac{e^{g(z)}}{z^u(1-z)}dz+\frac{1}{2\pi j}\oint_{\Gamma_1}\frac{e^{g(z)}}{z^u(1-z)}dz\\&=\frac{1}{2}+\frac{1}{2\pi}\int_0^{2\pi}\mathrm{Re}\left\{\frac{e^{-\frac{b^2}{2}\left(1-e^{j\theta}\right)}}{e^{j(u-1)\theta}\left(1-e^{j\theta}\right)}e^{-\frac{a^2}{2}\left(1-e^{-j\theta}\right)}\right\}d\theta\end{aligned} \tag{13}$$

This expression holds for any arbitrary values of a and b. The average detection probability over any fading channel can be obtained by finding the statistical expectation of (13) over the fading channel.

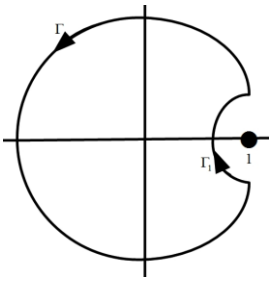


Fig.1 Contour Integral for $Q_u(a,b)$ for arbitrary a and b

III. SQUARE-LAW DIVERSITY COMBINING

This section discusses the energy-detection performance with SLC. In this scheme, the total decision statistics g_T is the sum of square of random variables, which follows a central chi-square distribution. This is expressed as follows:

$$\gamma_T = \sum_{k=1}^{L} \gamma_k \tag{14}$$

where γ_k , represents the output of the k^{th} square device. The corresponding average probability of detection $\overline{P}_{d,SLC}$ when a signal is present can be determined by averaging (13) over the density function of the ensuing fading channel. In terms of the product of the MGF of the individual branch. Given $a = \sqrt{2\gamma_T}$ and $b = \sqrt{\lambda}$, we can write the average probability of detection for an L-branch SLC receiver viz.,

$$\overline{P}_{d_{SLC}} = \frac{1}{2} + \frac{1}{2\pi}\int_0^{2\pi} \mathrm{Re}\left\{ \frac{e^{-\frac{\lambda}{2}\left(1-e^{j\theta}\right)}}{e^{j(Lu-1)\theta}\left(1-e^{j\theta}\right)} \prod_{k=1}^{L} \varphi_{\gamma_k}\left(1-e^{-j\theta}\right) \right\} d\theta \tag{15}$$

Note: $\phi_{\gamma_k}(s)$ is the MGF of the k$^{\text{th}}$ branch of the SLC receiver structure. The probability of false alarm is given viz,

$$P_{f_{SLC}} = \frac{\Gamma\left(Lu, \frac{\lambda}{2}\right)}{\Gamma(Lu)} \tag{16}$$

It is worth noting that (15) is tractable since it is in terms of a single integral and can thus be easily evaluated in the instances where the MGF $\phi_{\gamma_k}(s)$ is defined. In this work we chose to investigate the application of (15) to the named distributions.

IV. SQUARE-LAW SELECTION

This selects the branch with maximum statistics; the probability of false alarm $P_{f_{SLS}}$ is thus evaluated by using the CDF of this statistics [2] i.e. $y_{SLS} = \max\{y_1, y_2, \ldots, y_L\}$ whose CDF in the absence of signal H_0 yields, $F_{y_{SLS}}[y_{SLS} \mid H_o]$.

$$P_{f_{SLS}} = 1 - \left[1 - \frac{\Gamma\left(u, \frac{\lambda}{2}\right)}{\Gamma(u)}\right]^L \tag{17}$$

Similarly we can write the average probability of detection when there is signal, H_1 in a more compact form as

$$\overline{P}_{d\,SLS} = 1 - \prod_{k=1}^{L}\left[1 - \int_0^{\infty} Q_u\left(\sqrt{2\gamma_k}, \sqrt{\lambda}\right) f_\gamma(\gamma)\, d\gamma\right]$$

$$= 1 - \prod_{k=1}^{L}\left[\frac{1}{2} - \frac{1}{2\pi}\int_0^{2\pi} \mathrm{Re}\left\{ \frac{e^{-\frac{\lambda}{2}\left(1-e^{j\theta}\right)}}{e^{j(u-1)\theta}\left(1-e^{j\theta}\right)} \varphi_{\gamma_k}\left(1-e^{-j\theta}\right) \right\} d\theta\right] \tag{18}$$

where $\phi_{\gamma_k}(\bullet)$ is the MGF of k$^{\text{th}}$ branch of the SLS receiver.

V. COMPUTATIONAL RESULTS

We hereby present the numerical results, computations and graphs to substantiate the applications of our approach. In order to investigate the performance of energy detector utilizing our model we generate some complementary ROC curves, probability of a miss, $P_m = 1 - \overline{P}_d$ versus false alarm probability P_{fa}) as well as some P_d versus P_f curves under the η-μ and κ-μ faded channels.

Fig. 2 depicts that for the same signal energy, the smaller the order u, which is dependent on the number of, samples the better the performance of the detector for both extreme fading distributions. Hence, there is bound to be high sensitivity to a loss in energy compared to increase in noise energy.

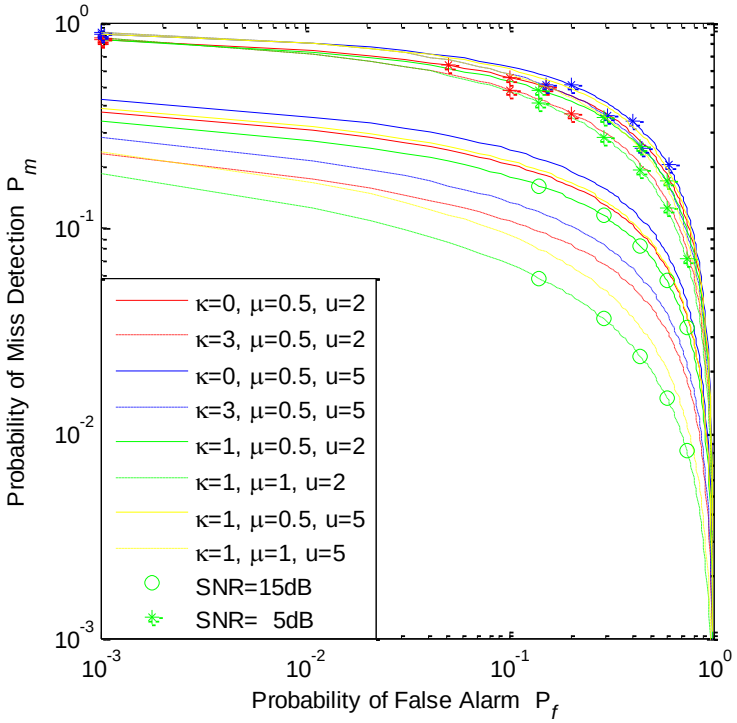


Fig. 2 Complementary ROC curve for single k - n faded channel reception

The figure above also reveals that there is improvement in performance with increasing values of mean channel SNR Ω,κ andμ, as well the time-bandwidth product u.

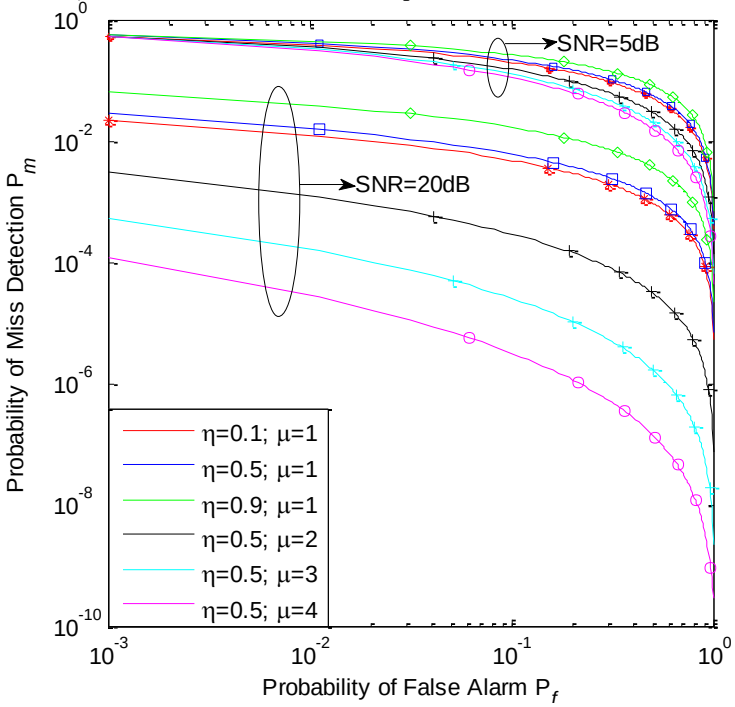


Fig. 3 Complementary ROC curves for single η–μ (format 2) faded channel reception

Fig. 3 shows that for a fixed μ, the performance is best when η is least. Similarly, the performance at a fixed η increases with increase in μ. It is clear from the figure that increases in channel mean SNR also improves the detection performance.

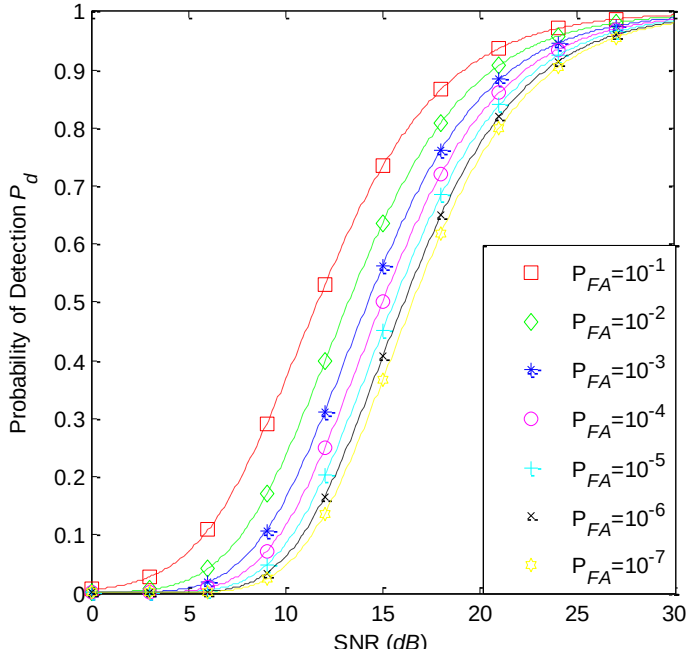


Fig. 4 Detection performance for single η–μ faded channel for a fixed $P_{FA}=10^{-1}$ to 10^{-7} η = 0.5 and μ=0.6.

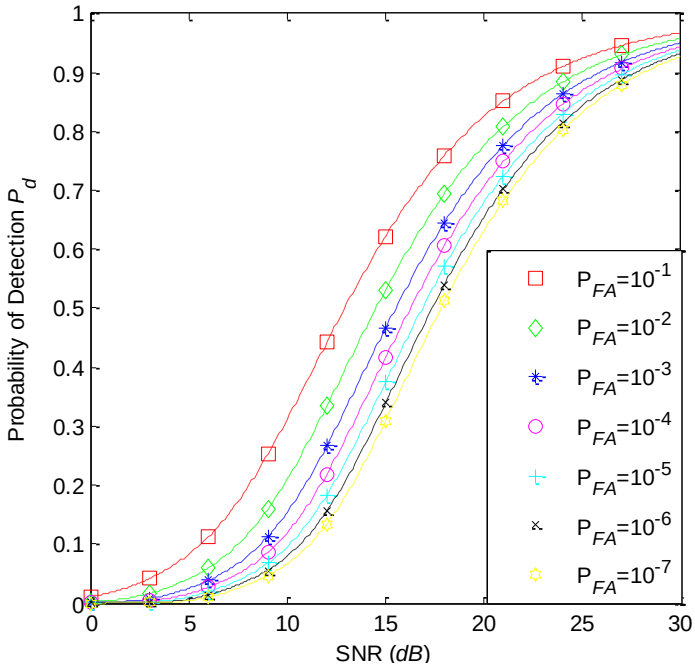


Fig. 5 Detection performance of single κ–μ faded channel for a fixed $P_{FA}=10^{-1}$ to 10^{-7} μ=0.75

It can be seen in Fig. 4 and 5 that in order to improve detection performance, you either increase the probability of false alarm or increase the mean channel SNR.

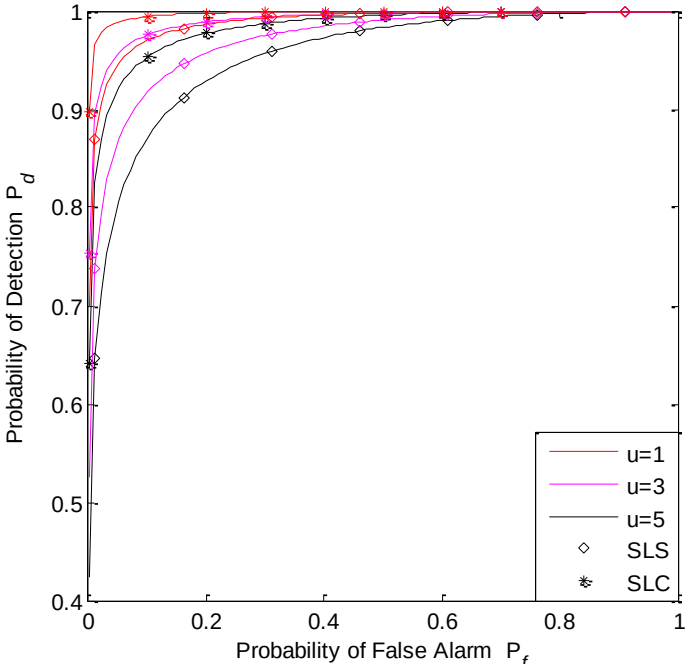


Fig. 6 Detection Performance of dual diversity κ–μ fading channel Ω_1 = 5db, Ω_2 = 10db, μ=3.5, and κ=1

In Fig. 6 we showed the variation of detection performance with changing *u*. It is clear that the performance degrades as u increases

for a fixed energy in the channel. The figure also confirms the SLC gives a better performance than SLS. It is also evident from Fig. 6 and Fig. 7 below that the performance degrades as *u* increases.

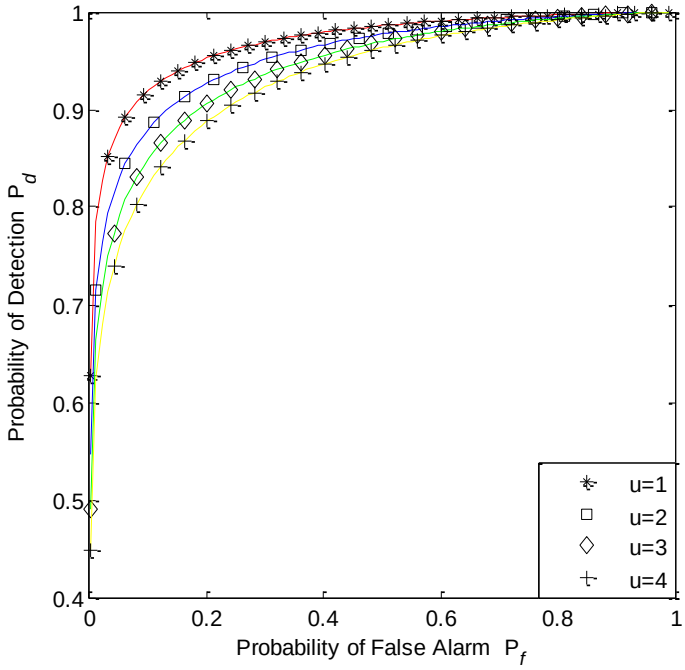


Fig. 7 Detection Performance of dual diversity η–μ (format 1) fading channel. Ω_1 = 5db, Ω_2 = 10db, μ_1= 0.5, μ_2 = 1, η_1= 2 and η_2 = 3

VI. CONCLUDING REMARKS

In this work we have taken a deep dive into detection performance on the scale of two extremes fading environments for both i.i.d and i.n.d cases. Consequently informed decision can be made on their effects on detection schemes and decision rules to be used. Elegant ways of expressing the generalized Marcum *Q*-function in terms of the MGF was introduced using previously obtained results. The chosen fading statistics though extreme but are general in that they are scalable to popular statistics as special cases. We have achieved result that agrees with prior works [2][4] on detection of unknown deterministic signals in an intuitive manner resulting in nicely looking expressions, which are easily evaluated. The approach here deployed has neither been adopted nor considered to the best our knowledge in earlier literary works on the subject.

APPENDIX A

In this appendix, we derive a closed-form expression for the *n*-th order derivative (with respect to parameter *s*) of an auxiliary function in the form of (A.1). This development is of interest because the *n*-th order derivative of the MGF of SNR in different fading environments such as Rayleigh, Rice, Nakagami-m or a Nakagami-q channels can be readily obtained using (A.5) by simple observation (or with minimal algebraic manipulations). For instance, $F(m/\Omega,0,m)$ and $F((1+K)/\Omega,K,1)$ correspond to the MGF of SNR in a single channel reception case over Nakagami-m and Rice channels, respectively. Thus, the corresponding *n*-th order derivative of the MGF of SNR in Rayleigh, Rice and Nakagami-m channels are given by (A.5) with appropriate substitutions for the values *A*, *B* and *C*. In fact, many of the entries in Table 1 for i.i.d. MRC diversity receiver are obtained by observation alone!

$$F(A,B,C)=\left(\frac{A}{A+s}\right)^C \exp\left(\frac{-sB}{A+s}\right), \quad C>0 \\ =\frac{A^C\exp(-B)}{(A+s)^C}\exp\left(\frac{AB}{A+s}\right) \tag{A.1}$$

By replacing the exponential term in Eq. (A.1) with its power series representation, we can carry out the *n*-th derivative of $F(A,B,C)$ term by term as

$$\frac{d^n}{ds}F(A,B,C)=A^C\exp(-B)\sum_{k=0}^{\infty}\frac{(AB)^k}{k!}\frac{d^n}{ds}\left(\frac{1}{(A+s)^{k+C}}\right) \\ =A^C\exp(-B)\sum_{k=0}^{\infty}\frac{(-1)^n(C+k)_n}{(A+s)^{k+C+n}} \tag{A.2}$$

where $(a)_n=a(a+1)...(a+n-1)$ denotes the Pochhammer symbol. Next, using the identity (A.3) into (A.2),

$$(C+k)_n=\frac{\Gamma(C+k+n)}{\Gamma(C+k)}=\frac{(C+n)_k\Gamma(C+n)}{(C)_k\Gamma(C)}, \tag{A.3}$$

we obtain

$$\frac{d^n}{ds}F(A,B,C)=\frac{(-1)^nA^C\exp(-B)\Gamma(C+n)}{(A+s)^{C+n}\Gamma(C)}\sum_{k=0}^{\infty}\frac{(C+n)_k}{k!(C)}\left(\frac{AB}{(A+s)}\right)^k \\ =\frac{(-1)^nA^C\exp(-B)\Gamma(C+n)}{(A+s)^{C+n}\Gamma(C)}{}_1F_1\left(C+n;C;\frac{AB}{A+s}\right) \tag{A.4}$$

where ${}_1F_1(.;.;.)$ denotes the confluent hypergeometric function. Applying Kummer's transformation formula [14, Eq. (9.212.1)] in (A.4), and recognizing that ${}_1F_1(-a;b;z)$ reduces to a finite series if $b>0$ and for integer $a\geq 0$, (A.4) simplifies into

$$\frac{d^n}{ds}F(A,B,C)=\frac{(-1)^nA^C\Gamma(C+n)}{(A+s)^{C+n}\Gamma(C)}\exp\left(\frac{-sB}{A+s}\right){}_1F_1\left(-n;C;\frac{-AB}{A+s}\right) \\ =\frac{(-1)^nA^C\Gamma(C+n)}{(A+s)^{C+n}}\exp\left(\frac{-sB}{A+s}\right)\sum_{k=0}^{n}\frac{\binom{n}{k}}{\Gamma(C+k)}\left(\frac{AB}{A+s}\right)^k \tag{A.5}$$

where the notation $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ in (A.5) corresponds to the binomial coefficient.

As a special case, when B = 0 then (A.1) and (A.5) become:

$$F(A,0,C)=\left(\frac{A}{A+s}\right)^C \tag{A.6}$$

and

$$\frac{d^n}{ds^n}F(A,0,C)=\frac{(-1)^nA^C\Gamma(C+n)}{(A+s)^{C+n}\Gamma(C)} \tag{A.7}$$

respectively.

References

[1] H. Urkowitz, "Energy detection of Unknown deterministic Signals," in Proc IEEE, vol. 55, no. 4, pp. 523-531, Apr 1967.

[2] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," in Proc. IEEE Int. Conf. Communications (ICC), 2003, pp. 3575–3579.

[3] ——, "On the energy detection of unknown signals over fading channels," IEEE Trans. Commun., vol. 55, no. 1, pp. 21–24, Jan. 2007.

[4] V. I. Kostylev, "Energy detection of a signal with random amplitude," in Proc. IEEE Int. Conf. Communications (ICC), 2002, pp. 1606–1610.

[5] S. Atapattu, C. Tellambura, and H. Jiang, "Relay Based Cooperative Spectrum Sensing in Cognitive Radio Networks," Proc. IEEE GLOBECOMC'09, pp. 1-5.

[6] W. Zhang and K. B. Letaief, "Cooperative spectrum sensing with transmit and relay diversity in cognitive radio networks," IEEE Trans. Wireless Commun., vol. 7, no. 12, pp. 4761–4766, Dec. 2008.

[7] S. P. Herath and N. Rajatheva, "Analysis of equal gain combining in energy detection for cognitive radio over Nakagami channels," in Proc. IEEE Global Telecomm. Conf. (GLOBECOM), 2008.

[8] Atapattu, Saman; Tellambura, Chinthananda; Jiang, Hai,, "Energy detection of primary signals over η-μ fading Channel," International Conference on Industrial Information System, pp 118-122 Dec. 2009.

[9] ——, "Analysis of area under the ROC curve of energy detection," IEEE Trans. Wireless Commun., vol. 9, no. 3, pp. 1216-1225, Oct. 2010.

[10] M. D. Yacoub, "The η-μ distribution: A general fading distribution," Proc. IEEE Veh. Technol. Conf. (VTC Fall). 2000, pp. 872 – 877

[11] ——, "The κ-μ distribution and the η-μ distribution," IEEE Antennas and Propagation Magazine, vol. 49, pp. 68–81, Feb. 2007.

[12] N. Ermolova, "Moment generating functions of the generalized η- μ and κ-μ distributions and their applications to performance evaluation s of communication systems," IEEE Commun. Lett., vol. 12, no. 7, pp. 502–504, July 2008.

[13] C. Tellambura, A. Annamalai, and V. K. Bhargava, "Closed form and infinite series solutions for the MGF of a dual-diversity selection combiner output in bivariate Nakagami fading," IEEE Trans. Commun. vol. 51, no. 4, pp. 539–542, Apr. 2003.

[14] I.S. Gradshteyn and I.M.Ryzhik, Table of Integrals, Series, and Products 6th ed. San Diego, CA: Academic, 2000.

[15] E. A. Neasmith and N. C. Beaulieu, "New results on selection diversity," IEEE Trans. Commun., vol. 46, no. 5, pp. 695–704, May 1998.

[16] A. H. Nuttall, "Some integrals involving the *Q*-function," Naval Underwater Syst. Center (NUSC) Tech. Rep., May 1974.

[17] A. Annamalai, O. Olabiyi, S. Alam, O. Odejide, and D. Vaman, "Unified Analysis of Energy Detection of Unknown Signals over Generalized Fading Channels," in Proc. of IEEE IWCMC 2011 Conference, Sept, 2011, Turkey, Istanbul.

[18] J. Shen, T. Jiang, S. Liu, and Z. Zhang, "Maximum channel throughput via cooperative spectrum sensing in cognitive radio networks," IEEE Trans. Wireless Commun., vol. 8, no. 10, pp. 5166–5175, Oct. 2009.

[19] J.G. Proakis, Digital Communications, 3rd ed. New York; McGraw Hill, 1995.

Evaluation of the Capacity of MIMO-OFDM Free Space Optical Communication System in Strong Turbulent Atmosphere

Manish Sharma[1], D. Chadha[2], Vinod Chandra[3]

[1,2,3] Department of Electrical Engineering, IIT Delhi, New Delhi-110016, India

Abstract - *In wireless communication Free Space Optical (FSO) communication is an emerging technology to provide high bandwidth and high quality communication. In this work we have evaluated the capacity of the Free Space Optical communication system with Multiple Input Multiple Output (MIMO) Orthogonal Frequency Division Multiplexing (OFDM) technique under strong atmospheric turbulence conditions. The capacity of MIMO-OFDM FSO is evaluated in the presence of Inter symbol Interference (ISI) and is compared with MIMO FSO with ISI. It is found that the effect of ISI is mitigated by using OFDM and the capacity is enhanced. The performance of the system is also compared for Gamma Gamma channel and Negative Exponential channel. It is found that Gamma Gamma channel reduces to Negative Exponential channel in very strong turbulence atmospheric condition.*

Keywords: FSO, MIMO, OFDM, Gamma Gamma distribution, Negative Exponential channel

1 Introduction

Free Space Optical (FSO) Communication represents one of the promising approaches for addressing the emerging broadband access market and its last mile bottleneck [1]. However, to exploit all potentials of FSO communication, the designers have to overcome the challenges introduced by the atmosphere. FSO offers many features like cost effectiveness, protocol independence, high speed connectivity and license free operation [2]. It is a secure communication which can bring superior quality, wideband services to home or businesses.

The performance of FSO communication can be improved by using Multiple-Input Multiple-Output (MIMO). When MIMO systems are used significant capacity gain can be observed as it exploits spatial multiplexing by having several transmit and receive antennas. At higher bit rates the problem of Inter Symbol Interference (ISI) is predominant. Due to the effect of ISI the capacity of the MIMO system severely degrades. The OFDM provides a solution to this problem. In OFDM the original signal is modulated by orthogonal tones. The entire stream is divided into many parallel substreams. Hence the symbol duration is increased and the effect of ISI is mitigated. OFDM has high spectral and power efficiency and simple frequency domain equalization. The implementation of OFDM can be easily carried out by using IFFT at the transmitter and FFT at the receiver. Thus it provides a cost effective solution to combat the fading caused due to atmospheric turbulence in FSO transmission. The use of OFDM in conjunction with MIMO is an attractive technology to meet up the needs of future broadband wireless communication. However there is a difference in using MIMO and OFDM in optical communication. The signals are unipolar by nature so we use either DC biased OFDM or clipped OFDM in the optical domain. Thus the combination MIMO-OFDM FSO provides robustness to the system and is an effective solution for next generation FSO systems.

In this work we have evaluated the capacity (bits/sec/Hz) of MIMO-OFDM FSO and MIMO FSO in the presence of ISI. The effect of increasing the order of MIMO is also shown in the results. To model the atmospheric turbulence Gamma Gamma distribution function and Negative Exponential channel models are used. The performance of both the models is also compared in very high strong turbulence region.

2 MIMO-OFDM FSO SYSTEM

Figure 1 shows a typical block diagram of the MIMO-OFDM FSO system. To evaluate the ergodic capacity [2], of MIMO-OFDM FSO communication system, we are using an $M_t \times M_r$ MIMO with OFDM.

2.1 Transmitter

The transmitter part of the system has M_t FSO transmitters. The bits from information source are first modulated by QPSK/QAM modulation scheme and mapped into symbols (not shown in the figure). These symbols are then demultiplexed according to number of transmitters. The message at each transmitter is now parallel substreams of the original signal. These substreams are the inputs of serial to parallel converter which takes N (Number of Subcarriers of OFDM) of these symbols as input and produces N output symbols corresponding to the original signal. These N output signals

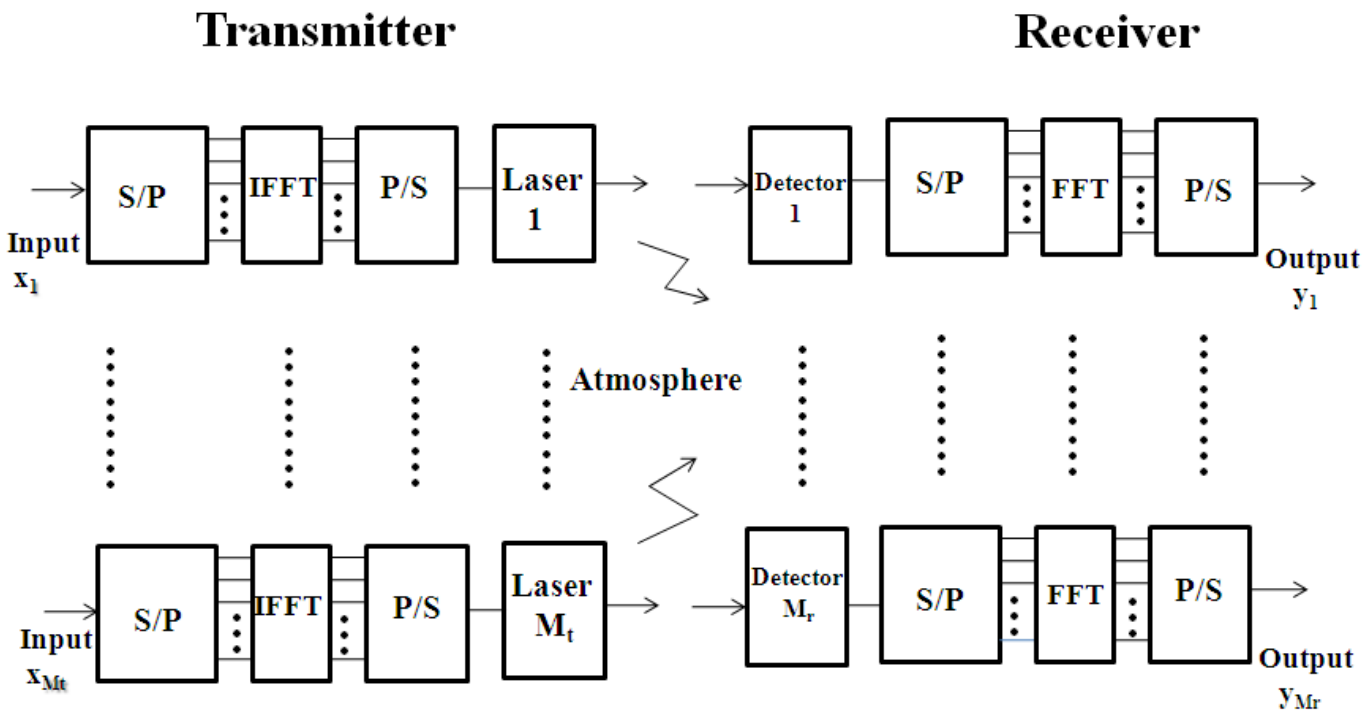


Fig.1. Block Diagram of MIMO-OFDM FSO system

are put through IFFT for OFDM modulation. The output of the OFDM modulator is the time domain symbol corresponding to the frequency domain input symbol. These symbols are then converted from parallel to serial and transmitted by all the M_t by FSO transmitters. The FSO transmitter's uses either laser diode or light emitting diodes to transmit the optical signal.

2.2 Atmospheric Channel

The signals propagate through the atmosphere where they get distorted due to the effect of turbulence. The Gamma Gamma distribution model is used for a wide range (weak to strong) atmospheric turbulence conditions [1, 4]. In this model the variations of the medium are assumed as individual cells of air or eddy of different diameters and refractive indices. The intensity fluctuations are called scintillation, one of the most important factors that limit the performance of an atmospheric FSO communication link. The most widely accepted theory of turbulence is attributed to Kolmogorov. This theory assumes that kinetic energy from large turbulent eddies, characterized by the outer scale L_0, is transferred without loss to the eddies of decreasing size down to sizes of a few millimetres characterized by the inner scale l_0. The inner scale represents the cell size at which energy is dissipated by viscosity. The refractive index varies randomly across the different turbulent eddies and causes phase and amplitude variations to the wavefront. Turbulence can also cause the random drifts of optical beams which are called beam wandering and can induce beam defocusing.

The irradiance of optical field is defined as the product of two random processes, i.e. $I = I_x I_y$, where I_x arises from large scale turbulent eddies and I_y from small-scale eddies. The probability density function (pdf) of the Gamma Gamma channel is given by (1)

$$f(I) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}} I^{\left(\frac{\alpha+\beta}{2}\right)-1} K_{(\alpha-\beta)}(2\sqrt{\alpha\beta I})}{\Gamma(\alpha)\Gamma(\beta)} \quad (1)$$

where I is the signal intensity, α and β are parameters of the pdf, $\Gamma(.)$ is the Gamma function and $K_{(\alpha-\beta)}(.)$ is the modified Bessel function of the second kind of order $(\alpha-\beta)$. Here, α and β are the effective number of small-scale and large scale eddies of the scattering environment. These parameters can be directly related to atmospheric conditions as

$$\alpha = \left(exp\left[\frac{0.49\sigma_R^2}{(1+1.11\sigma_R^{12/5})^{7/6}}\right] - 1\right)^{-1} \quad (2)$$

$$\beta = \left(exp\left[\frac{0.51\sigma_R^2}{(1+0.69\sigma_R^{12/5})^{5/6}}\right] - 1\right)^{-1} \quad (3)$$

where σ_R^2 is the Rytov variance given by

$$\sigma_R^2 = 1.23 C_n^2 k^{7/6} L^{11/6} \quad (4)$$

Here k is the optical wave number given by $k=2\pi/\lambda$; λ is the wavelength and C_n^2 is the atmospheric structure parameter.

The other model which is used in very high strong turbulence region is a Negative Exponential channel model. In the limit of strong irradiance fluctuations (i.e., in saturation regime and beyond) where the link length spans several kilometers, the number of independent scatterings becomes large [4]. The amplitude fluctuation of the field traversing the turbulent medium in this situation is generally believed and experimentally verified to obey the Rayleigh distribution implying negative exponential statistics for the irradiance [4]. That pdf of Negative Exponential channel is given by (5)

$$p(I) = \frac{1}{I_0}\exp(-I/I_0), \qquad I_0 > 0 \tag{5}$$

where $E\ [I] = I_0$ is the mean radiance which is often normalized to unity. The Gamma Gamma turbulence model also gives the negative exponential in the limit of strong turbulence.

2.3 Receiver

After the signal passes through the channel, the faded signal with superimposed noise on the M_t transmitted signals is received by the FSO receiver array. At the receiver the signals are detected by optical detectors like an avalanche photo diode. These detected signals after converting into parallel subchannel are put through Fast Fourier Transform (FFT) for OFDM demodulation. The fading caused by atmospheric turbulence produces ISI. This ISI can be taken into account by changing the delay spread into a number of interfering symbols.

3 Capacity Analysis

The conventional MIMO and MIMO-OFDM channel are described by (6), (7) respectively

$$\tilde{y} = \tilde{H}\tilde{x} + \tilde{n} \tag{6}$$

$$\tilde{y} = \tilde{E}_2\tilde{H}\tilde{E}_1\tilde{x} + \tilde{E}_2\tilde{n} = \tilde{H}'\tilde{x} + \tilde{n}' \tag{7}$$

where $\tilde{E}_1$ is IFFT transformation matrix given by $\tilde{E}_1 = [e_{i,j}] \otimes I_{M_t}$; $e_{i,j} = exp(j\,2\pi ij/N)$; $i,j=0,1,2.....N\text{-}1$, $\tilde{E}_2$ is FFT transformation matrix given by $\tilde{E}_2 = [e_{i,j}] \otimes I_{M_r}$; $e_{i,j} = exp(-j\,2\pi ij/N)$; $i,j=0,1,2....N\text{-}1$. N is the number of subcarriers, I_{Mt} represents the $M_t x M_t$ identity matrix and I_{Mr} represents an $M_r x M_r$ identity matrix. The characteristic of the channel is taken into account by changing the delay spread into a number of interfering symbols represented by L. The channel matrix is given by [5].

$$\tilde{H} = \begin{bmatrix} H^1 & 0 & \cdots & 0 \\ \vdots & H^1 & \cdots & 0 \\ H^L & \vdots & \ddots & 0 \\ 0 & H^L & \ddots & H^1 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & H^L \end{bmatrix}$$

The channel capacity of conventional MIMO and MIMO-OFDM [3, 5] are given by (8), (9)

$$C = E\left[log_2\, det\left(I_m + \frac{\rho}{M_t}\tilde{H}\tilde{H}^{\dagger}\right)\right] \tag{8}$$

$$C = E\left[log_2\, det\left(I_m + \frac{\rho}{M_t}\tilde{H}'\tilde{H}'^{\dagger}\right)\right] \tag{9}$$

where $E\ [.]$ denotes expectation, m is min (M_r, M_t), ρ is an average signal to noise ratio (SNR) at each receive antenna and $\tilde{H}'$ is the channel matrix which incorporates the subcarriers of OFDM, number of interfering symbols and random numbers generated according to the channel model i.e. Gamma Gamma distribution and Negative Exponential channel. †denotes conjugate transpose of a matrix.

4 Results

In this section, simulation results for the capacity of MIMO-OFDM FSO with Gamma Gamma channel and Negative Exponential channel are presented. In the simulation 64 numbers of subcarriers were used. The simulation was carried out in MATLAB environment.

Figure 2 compares the performance of MIMO FSO with ISI and MIMO-OFDM FSO with ISI for a *2x2* antenna system. The Gamma Gamma channel model for strong turbulence (σ_R=2) is used here and it is evident from the figure that in the presence of ISI the performance of the MIMO-OFDM FSO system is superior to MIMO FSO. The effect of using OFDM is depicted in the figure as it reduces ISI caused by the atmospheric turbulence generated fading. The increase in the capacity at E_b/N_0 =10 dB is approximately 2 bits/sec/Hz.

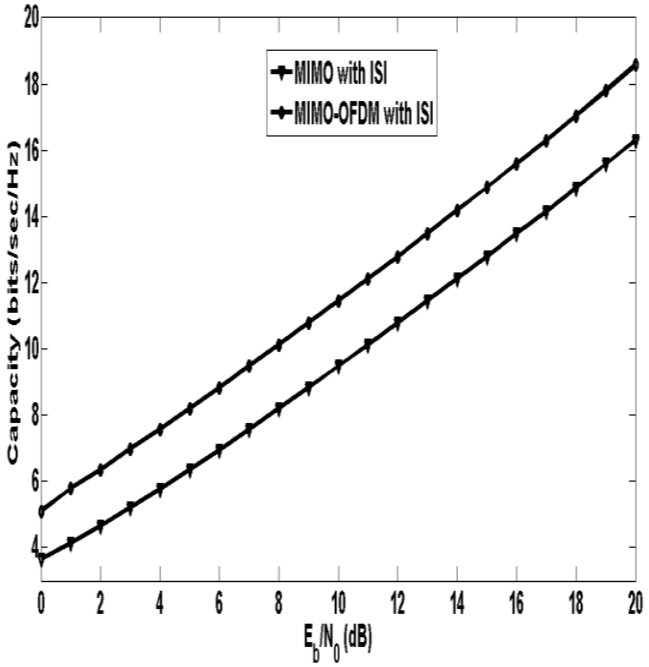


Fig. 2. Capacity v/s E_b/N_0 for MIMO and MIMO-OFDM with ISI for Gamma Gamma Channel

Figure 3 compares the performance of MIMO-OFDM FSO with MIMO FSO with ISI for different number of receiving antennas at σ_R=2. The advantage of using MIMO is observable from the figure since the capacity in both the cases is increased as the number of receiving antennas is increased. These curves are drawn at E_b/N_0=2 dB.

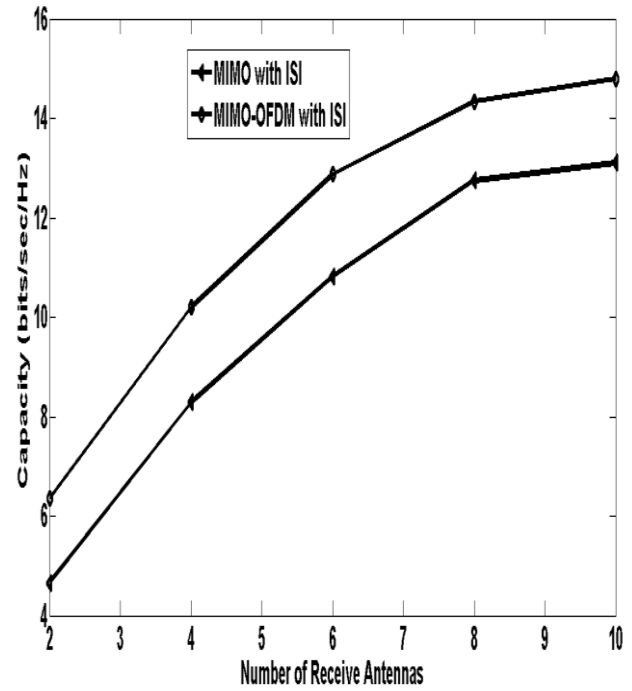


Fig. 3. Capacity v/s Number of Receive Antennas for Gamma Gamma channel

Figure 4 compares the performance of Gamma Gamma and Negative Exponential channel in very strong turbulent atmospheric condition (σ_R=4) for a *2x2* antenna system. The graph reveals that the Gamma Gamma channel reduces to negative Exponential channel in very strong turbulence condition. The increase in the capacity at E_b/N_0 =10 dB is approximately 2.5 bits/sec/Hz for Gamma Gamma channel. If this is compared with the numerical value of the increase in capacity as obtained from figure 2 at same E_b/N_0 the advantage of using OFDM with MIMO is evident. Since fading is increased; hence, the capacity of MIMO FSO with ISI suffers more than the MIMO-OFDM FSO.

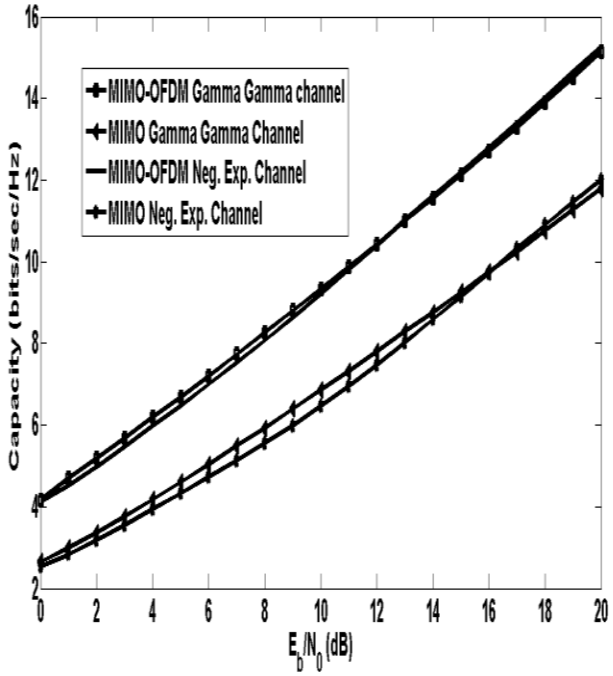


Fig. 4. Capacity v/s Eb/N0 for Gamma Gamma channel and Negative Exponential channel

5 Conclusions

In this paper the capacity of MIMO-OFDM FSO is compared with MIMO FSO in the presence of ISI in strong and very strong turbulent atmospheric regions. It is found that the use of OFDM improves the performance of the system by mitigating the effect of ISI and capacity is increased. Since MIMO uses spatial diversity to combat fading hence the capacity is enhanced as the number of transmitting and receiving antennas is increased. The performance of MIMO-OFDM FSO is also compared with MIMO FSO in the presence of ISI for Gamma Gamma distribution model and Negative Exponential channel model. The performance of Gamma Gamma channel is almost equal to the Negative Exponential channel in the very strong turbulence region.

6 References

[1] Ehsan Bayaki, Robert Schober and Ranjan K.Mallik, "Performance analysis of MIMO Free-Space Optical systems in Gamma-Gamma Fading,";IEEE Transactions on Communications, Vol.57 No.11, Pages:3415-3424,November 2009.

[2] Ngoc T. Dang and Anh T. Pham, "Performance improvement of FSO/CDMA systems over dispersive turbulence channel using multi-wavelength PPM signaling,"; Optics Express, Vol. 20, Issue 24, Pages:26786-26797, 2012.

[3] V. K. Varma Gottumukkala and Hlaing Minn, "Capacity Analysis and Pilot-Data Power Allocation for MIMO-OFDM with Transmitter and Receiver IQ Imbalances and Residual Carrier Frequency Offset,"; IEEE Transactions on Vehicular Technology, Vol. 61 No. 2, Pages: 553- 565, February 2012.

[4] Wasio O. Popoola and Zabih Ghassemlooy, "BPSK Subcarrier Intensity Modulated Free-Space Optical Communications in Atmospheric Turbulence,"; Journal of Lightwave Technology, Vol. 27 No. 8, Pages: 967-973, April 15,2009.

[5] P. Uthansakuand and M. E. Bialkows,"Multipath signal effect on the capacity of MIMO, MIMO-OFDM and spread MIMO-OFDM,"; Microwaves, Radar and Wireless Communications,2004. MIKON-2004. 15th International Conference, Vol.3, Pages: 989- 992, 17-19 May 2004.

SESSION

POSTERS AND SHORT PAPERS

Chair(s)

TBA

Using a ZigBee Wireless Network for Greenhouse Control over the Internet

Yuriy V. Eero, Mohamed A. Osman, Scott Hudson

Washington State University, 2710 Crimson Way, Richland, WA, 99354, USA

Abstract – *There are many reasons to improve efficiency and quality of a greenhouse. In areas where resources are scarce and conditions are not suitable for plant growth, a greenhouse can provide an efficient and carefully balanced environment. In the last few years there were a number of improvements made by adding ZigBee wireless network devices at the Washington State University greenhouse, and until now these improvements were only accessible for use from inside of the wireless network. This paper briefly describes how communication can be established over the internet with a number of ZigBee wireless networks to further improve a greenhouse.*

Keywords: wireless sensor network; greenhouse; ZigBee; MySQL; control over the internet

1 Introduction

Managing a greenhouse could be a difficult task. Depending on the size, the variety of plants, and availability of resources such as water, it can quickly become a complex controls system. In [5][3] motivation is given for having an efficiently operating greenhouse. A ZigBee wireless network is presented as an ideal solution to this problem.

Washington State University (WSU) – Tri-Cities camps is located in a steppe climate. Even locations in close proximity to the Columbia River are not ideal for unirrigated vegetation. This introduces challenges in efficiently operating the WSU greenhouse, and it is not surprising that a number of improvements to the greenhouse have already been made utilizing ZigBee technology; however, all of the previous projects added solutions which could only be used from the inside of the greenhouse. This presented an inconvenience to someone having to be inside the structure to collect measurements or change the conditions. The original system required constant monitoring and intervention to achieve proper operation of the greenhouse. By connecting ZigBee wireless network to a website and MySQL servers, it became possible to monitor sensor readings via a browser connected to the internet. A password protected section was added to the website so commands could be sent to the greenhouse such as watering a plant for a period of time. In this paper, it will be briefly explained how this system was designed at the WSU greenhouse.

2 Materials and methods

The challenge of assembling a system capable of securely communicating with wireless devices over the internet comes from composing code in a variety of programming languages, designing and interconnecting hardware, understanding elements of a secure network, and most importantly visualizing how all of these components fit together.

The overall network consisted of the ZigBee wireless network, website and MySQL servers, a central computer responsible for a connection between the ZigBee coordinator and MySQL server, and a gateway for network security. All communication over the internet except communication with the browser was accomplished through a VPN connection. This setup allowed the gateway and servers not to be in the same physical location with the wireless network. The only open port of the gateway for outside of the network communication was the website port. The website server provided the point of access for viewing the sensor data via any internet connected browser.

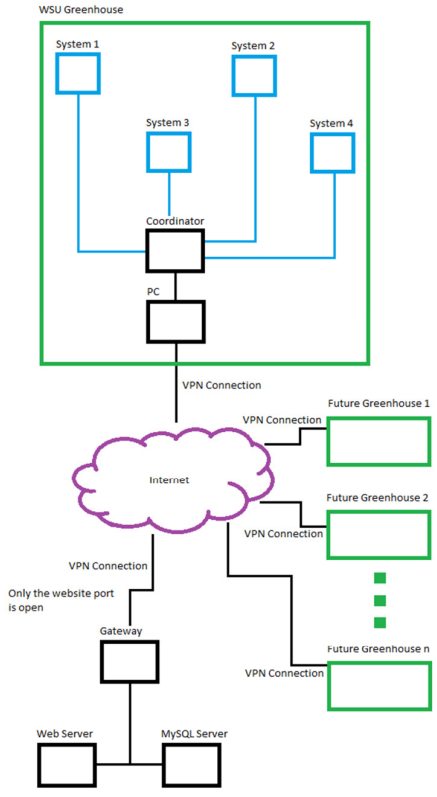


Fig. 1: Wireless network configuration for WSU greenhouse.

All of the information regarding sensors and instruction status was stored on a MySQL database server. The network configuration can be seen from Fig. 1. Systems 1, 2, 3, and 4

are wireless ZigBee devices, and both the website and MySQL servers are installed on a single machine. The future greenhouses are shown to illustrate ease of expansion to the current network. The website server made the greenhouse controls user friendly and accessible via an internet browser. The reason for introduction of the MySQL server was to store useful information in the database tables. One of the database tables was used to store instructions for the ZigBee devices, and another held the sensor readouts. The ZigBee wireless network was put together using XBee ZB transceivers made by Digi. An XBee ZB transceiver is designed to work in a mesh wireless network configuration, but it still requires one central command point called the coordinator. The coordinator is yet another XBee ZB transceiver setup specifically for this function. The rest of the XBee ZB devices were programmed as fully functional devices. The entire wireless network was configured to operate using API protocol. Hard-wire communication between the coordinator and the PC consisted of a USB interface connector and the applet written in Python. The applet is the focal point of the network, as it extracts information from the coordinator about the status of the network and stores that information on a remote MySQL server. Additionally, it monitors the same server for any instructions from an authorized remote browser and relays those instructions to the appropriate local devices for execution.

An important element of any secure network is a correctly configured gateway, so for this project, one was setup using ClearOS, as it proved to be a very intuitive and cost effective secure WAN interface. Both the website and MySQL servers were setup separately from the gateway using Ubuntu OS. This decision was made to provide additional security to the network. MySQL is an open source database and is a good solution for information management in this type of application.

The programming languages used for this project included Python, Java, PHP, HTML, and AVR machine language. The hardware portion consists of the XBee ZBs, a USB interface connector for PC interaction with a wireless network coordinator, and an Atmel ATmega48P microcontroller used for controlling servos. Two separate boards were designed with CadSoft EAGLE PCB Design Software and commercially fabricated. These boards greatly accelerated testing, troubleshooting, and ease of connection between the XBee ZB and ATmega48P.

The communication established between website and MySQL servers was done with server side programs written using PHP. The most important reason for use of a PHP based program is that the source is not accessible from a remote browser, and it runs directly on the server [2]. This is important from a security stand point, as it allows sensitive information to be hidden.

While an XBee ZB comes standard with four 8 bit analog to digital converters (ADC) and a number of on/off programmable pins, its greatest feature is its capability to communicate with other ICs using Universal Asynchronous Receiver/Transmitter (UART). It allows control of additional hardware connected to an XBee ZB transceiver. In this case, the ATmega48P microcontroller was used to operate a servo connected to a water valve of the irrigation system.

3 Results

Currently the wireless network consists of four devices. There is a station with three sensors for measuring ambient temperature, ambient humidity, and humidity of flowerpot soil. There are also three smaller devices for measuring temperature in different locations of the greenhouse. Whenever the Python applet is running, the information is available on the website. By reading the difference between ambient and soil humidity, it's easy to determine if a flowerpot needs to be watered, and a command could be sent from the browser to enable irrigation. This project is a critical step forward to providing full remote control to a greenhouse.

4 Conclusions

Introducing ZigBee network to WSU greenhouse combined with monitoring through the internet is convenient. It is now possible to keep an eye on plants from half way around the world. Although this network was designed for a greenhouse, there are many other applications which could benefit from the concept. Since ZigBee devices are relatively inexpensive and require little maintenance, it is also affordable and easy to implement [1][4]. The next step in greenhouse improvement is going to be combining all of the previous research into one easy to use system. This includes power consumption monitoring, lights control, and gathering useful information from a variety of sensors. The previously mentioned Python applet can be easily modified to handle many tasks automatically. Tasks like controlling air conditioning units based on greenhouse temperature or shading based on available light. The ability to send a complex set of information via UART to a microcontroller through XBee ZB transceiver opens up possibilities for implementation of wireless robotics on the network. These robots could be designed for tasks like removing weeds or fertilizing plant soil. Lastly, adding internet access to a greenhouse creates room for many more improvement possibilities.

5 References

[1] A. Elahi and A. Gschwender, *ZigBee Wireless Sensor and Control Network*. Upper Saddle River, N.J.: Prentice Hall, 2010
[2] J. Valade, *PHP & MySQL*. Hoboken, N.J.: Wiley Publishing, Inc., 2006
[3] Song J., "Greenhouse Monitoring and Control System Based on ZigBee Wireless Senor Network," International Conference on Electrical and Control Engineering (ICECE 2010), IEEE Press, Jun. 2010, pp. 2785-2788, doi: 10.1109/iCECE.2010.680
[4] Zigbee Allience, Zigbee Specification 2012, http://www.zigbee.org
[5] Y. Zhou, X. Yang, X. Guo, M. Zhou, and L. Wang, "A Design of Greenhouse Monitoring & Control System Based on ZigBee Wireless Sensor Network," in Proc. Int. Conf. WiCom, Sep. 21–25, 2007, pp. 2563–2567

Efficient Processing of Sensor Network Data using Object-Oriented Databases

Kyung-Chang Kim[1], **and Choung-Seok Kim**[2]
[1]Department of Computer Engineering, Hongik University, Seoul, South Korea
[2]Department of Information Technology, Silla University, Pusan, South Korea

Abstract - *Many wireless sensor network (WSN) applications require join of sensor data belonging to various sensor nodes. For join processing, it is important to minimize the communication cost since it is the main consumer of battery power. Most join techniques for sensor data assumes that the sensor data are either stored in OS files or in relational databases. In this paper, we introduce a join technique for sensor networks based on column-oriented databases. A column-oriented database store table data column-wise rather than row-wise as in traditional relational databases. The proposed algorithm is energy-efficient since, unlike relational databases, only relevant columns are shipped to the join region for final join processing.*

Keywords: Wireless sensor network, sensor data, join technique, column-oriented database, communication cost.

1 Introduction

Many sensor network applications require correlation of sensor readings scattered among sensor nodes. For example, in an object tracking system, one may be interested in objects that travelled from one designated region to another designated region to monitor the traffic volume and speed of particular objects. The sensor network can be modeled as a distributed database and the sensor readings are collected and processed using queries. A join is an important operation in finding the correlation of sensor readings.

One of the most important performance criteria in processing a join operation for sensor networks is to minimize the total communication cost. A total communication cost is the total data transfer between neighboring sensor nodes. Minimizing the communication cost is important because each sensor node has limited battery power and data communication is the main consumer of battery energy.

In this paper, we propose a novel join technique for wireless sensor networks to minimize the total communication cost and improve performance. Our approach is based on a column-oriented databases rather than relational databases to store sensor data. Recent years have seen an increased attention and research work on column-oriented databases. A column-oriented database store data in column order (i.e. column-wise) and not in row order as in traditional relational databases. They are more I/O efficient for read-only queries since they only access those columns (or attributes) required by the query. The read-only queries are common in workloads and applications found in data analysis, semantic web and sensor networks. The paper is organized as follows. Related works are mentioned in Section 2. Section 3 discusses the proposed algorithm. Conclusion is made in Section 4.

2 Related Works

Query optimization techniques for column-oriented databases were introduced in the literature. Materialization strategies, both early and late, are important in column-oriented databases since tuple reconstruction is required to produce and display query result using conventional ODBC interface [1]. Our earlier join algorithm for sensor networks based on column-oriented database was proposed in [2]. The algorithm is based on an early materialized strategy in column-oriented database. The invisible join [3] is another research result to improve star schema queries using a column-oriented database. Several energy-efficient query processing algorithms were proposed for wireless sensor networks [4-5].

3 Proposed Algorithm

In this paper, we assume that the sensor data are stored in a column-oriented database rather than a relational database. There are two materialization strategies, early and late, in column-oriented databases [2]. Materialization, also known as tuple stitching or tuple construction, is a process of combining single-column projections into wider tuples. The reason for materialization in column-oriented databases is because it needs to output row-style tuples to support standards-compliant relational database interface such as ODBC and JDBC. In early materialization strategy, each column is added to the intermediate query result to form tuples if the column is needed. In late materialization strategy, the accessed columns do not form tuples until after some part of the query plan has been processed. We propose an in-network join technique based on early and late materialized strategy for column-oriented databases.

We assume that the join query involves the join of sensor data in R region and S region of the sensor network. In addition, the actual join is performed in the join region since no single node can perform join due to resource limitations.

The data in R (S) region is stored in R (S) table in column-order.

The proposed join algorithm executes in three phases, namely the ***selection*** phase, the ***join*** phase, and the ***result*** phase. In the selection phase, the semi-join of join columns of R and S is performed to determine the qualified columns and column values for the given query. The result of the semi-join is used to create bitmaps to be shipped back to region R and region S. A Bitmap(R) contains one bit for every tuple in R table. That bit is set to 1 if it is in the semi-join result. The qualified column values in R and S are shipped to the join region for the actual join. In the join phase, the qualified column values of R and S are joined using either a nested-loop or a sort-merge join algorithm. In the result phase, the result of join is sent to the query sink as the query result.

In the late materialized strategy, the qualified column values of R and S are stitched together to construct tuples in the result phase. In the early materialized strategy, the columns in R and S participating in the query result are stitched to form tuples before shipping to the join region for final join. The difference between the early and late materialized strategy is the timing of the tuple stitching. In late materialized strategy, it is performed in the result phase while in the early materialized strategy it is performed in the selection phase.

In this paper, we also introduce a hybrid strategy to reduce the communication cost. Both the early and late materialized strategy performs the semi-join in the semi-join region which is not the join region. In addition, the actual join is also performed in the join region for the reduced data. In the hybrid case, no distinction is made between the semi-join region and the final join region. Both the semi-join and the final join are performed in the join region. Hence, no semi-join region is required.

4 Conclusions

We introduced a join technique that can be used in wireless sensor networks to reduce communication cost during sensor query processing. Communication cost is the main query performance criteria in sensor networks. Our algorithm is based on column-oriented databases. To the best of our knowledge, no other research results for sensor networks based on column-oriented databases were published in the literature. The advantage of using column-oriented database is that only relevant columns of qualified records, not the whole records, need to be shipped to the join region for final join reducing the amount of data shipped. Our algorithm is based on both the early and late materialized strategies. We also introduced a hybrid approach in which the semi-join and the join processing both occur in the same region of the sensor network. It is easy to observe that the communication cost is reduced during join processing based on column-oriented databases compared to data storage based on relational databases.

5 Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number 2012-0007012)

6 References

[1] Abadi, D.J., Myers, D.S., DeWitt, D.J. and Madden, S.R. Materialization strategies in a column-oriented DBMS. In Proceedings of the International Conference on Data Engineering (ICDE), Istanbul, Turkey, 2007.

[2] Kim, K.C and Kim C. S. An Energy-Efficient Technique for Processing Sensor Data in Wireless Sensor Networks. In: Proceedings of Ubiquitous Computing and Multimedia Applications (UCMA), Bali, Indonesia 2012

[3] Abadi, D.J. and Madden, S.R., Hachem, N. Column-Stores vs. Row-Stores: How Different Are They Really? In Proceedings of ACM SIGMOD, Vancouver, Canada, 2008.

[4] S. J. Lim and M. S. Park. Energy-Efficient Chain Formation Algorithm for Data Gathering in Wireless Sensor Networks. International Journal of Distributed Sensor Networks, Vol. 2012, 2012.

[5] Ren, Qingchun and Liang, Qilian. Energy and Quality Aware Query Processing in Wireless Sensor Database Systems. Information Sciences. Vol. 177, pp. 2188-2205, 2007

GPS and GSM based points of interest localization system for blind people

W. Gelmuda[1]**, and A. Kos**[1]
[1] AGH University of Science and Technology, Department of Electronics, Krakow, Poland

Abstract - *GPS/GSM based localization system is designed to help blind people navigate and find important for them points of interest (POI) in urban areas. When the GPS localization is not for some time or the system has been just turned on in an unfamiliar place for the blind user, where one cannot get a GPS fix, the GSM based localization is being used in order to find a specific POI or to alert a family member or emergency services about the blind user's approximate location. In the following paper we present a system overview, its prototype and some tests results.*

Keywords: blind people;GPS;GSM;localization;MOBIAN

1 Introduction

Navigation devices and systems of different kinds have been used since ages. Technology advance makes it possible to develop low-priced, yet accurate devices for localization purposes. Since most of these devices use some kind of a display to present information, they are not suited for the visually impaired users. However, there are some systems being developed specifically for blind people [1]. Unfortunately, these kinds of systems usually have limited functionality, like for instance, informing users about a specific POI, pedestrian crossings [2]. Nonetheless, using a GPS based application is only proper when the receiver has a clear sky view. Inside of buildings and/or underground areas, the GPS signal is not received, thus getting a position fix is problematic. There are some solutions, which include other sensors, for localization and navigation when a GPS fix was accessible and due to entrance to a building, is no longer available [3], but in situations where a blind user turns on the device for the first time in places where the GPS signal is too weak to get a fix it is hard to get a reference to a global position. That is why we present GPS based system overview with GSM localization capability.

2 Related work

Electronic navigation assistants for blind people are difficult to develop. GPS based devices, even with some position correction functionality [4], do not provide high enough reliability to safely navigate a blind person through an urban area. Even a 5 m position error while locating a pedestrian crossing can be dangerous for visually impaired users. That is why information about specific, important for blind people POI, should be considered rather as a tip than a navigation guide like in car navigation assistants. When the GPS signal is too weak to get a localization fix, a GSM based localization comes in handy. There are many techniques to get a pinpoint location by using GSM network [5]. RSSI based systems are accurate, but they need a fingerprint map or an architecture map to correlate RSSIs from GSM base stations with pinpoint location and these are not always available for some areas [6]. There are some systems which provide high reliability and were created especially for blind people. However, these systems often employ components, which have to be embedded into existing infrastructure, for instance RFIDs [7-8], thus these systems are expensive if one wants to have them installed on large areas. Another technique uses the fact that every GSM base station can be identified with special codes like: Mobile Network Code (MNC), Mobile Country Code (MCC), Location Area Code (LAC) and Cell ID (CID). There are databases, in which these identifiers can be correlated with base station longitude, latitude and approximate signal range [9].

3 System overview

The system block diagram is presented in Fig. 1a. The system runs ARM Cortex-M3 based low-energy EFM32 microcontroller and embeds a GPS module for a pinpoint location. Navigation assistants usually use POI database to inform about some specific locations, e.g. a gas station, restaurants, traffic accidents, etc. POI which blind users could benefit from should be far more extended than a widely available POI used in common databases, like Google Places. Apart from the standard places, this database should include pedestrian crossings adapted for blind people usage, hospitals, special centers for visually impaired, bus stops and many others. Also, there should be a way to easily include specific, customized records, like places from blind user environment, e.g. university campus buildings. A small prototype database was embedded into microcontroller memory. With a clear sky view, the GPS localization accuracy is sufficient for guiding a blind user. When entering a building, where GPS localization is no longer available, a GSM Cell ID based localization is used. Thus, a GSM/GPRS module is embedded. Cell ID based localization technique does not provide great localization accuracy, nevertheless blind people navigation inside buildings is difficult, even with the 1 m accuracy, without external sensors, cameras, building architecture, etc., due to users'

visual impairment. Therefore, Cell ID localization is being used just to inform the blind user about specific POI in the nearest area in the selected range and to send possible location via a distress SMS in case of some emergency or an accident. Small head phones are used for feeding users voice commands and information about POI.

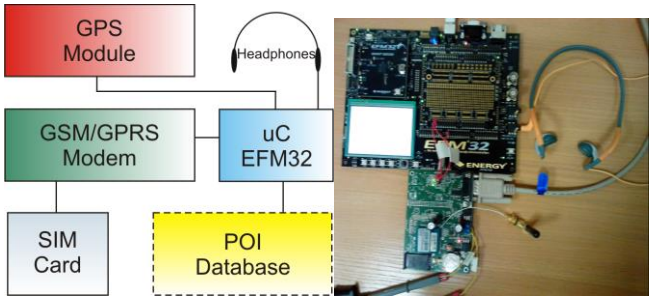


Figure 1. a) system block diagram
b) device prototype

4 Acquiring GSM stations localization

Practically, every GSM modem has the ability to check base stations codes, desired for GSM localization (MNC, MCC, LAC, CID). The system for blind people checks the codes not only from a base station, which it is currently connected to, but also base stations in the area (with signal range). This is done due to the fact that checking only the main base station (mainly with the strongest RSSI) would give accuracy of hundreds of meters or even a couple of kilometers. By checking all the available base stations codes, the system is able to increase the localization accuracy.

There are many ways to link base stations codes with their global position and range. The easiest method is to create a database of base stations with their codes and global localization and compare the data from a GSM modem against it. It would work without any Internet connection, however would consume large quantities of memory. Since the system has a GSM/GPRS modem onboard, it can check external services. One of these services is run by Google, so cell phones even without any GPS module can obtain their approximate position and users can run Google Maps and other applications. Google does not provide GSM based localization by itself as an official API, but still it can be accessible [9]. Other service which helps to locate a GSM base station is the OpenCellID project [10]. It is an open source and has documented API for queries.

5 Tests and results

The system prototype was designed and is presented in Fig. 1b. In a building, where the GPS fix could not have been established, the GSM localization service was enabled. Afterwards, base station codes were correlated with base stations positions and ranges via Google service. Sample results are presented in Fig. 2. Every visible base station is presented with a blue circle. Circle center is a base station location and circle radius is a range. The black dot is an actual device location in the building. The red shape is the area where all the signal areas from base stations overlay. That is the likely GSM localization based fix. The last step for the device is to check if there are any important POI for blind users.

Figure 2. GSM base stations locations in the nearest area and real device position.

6 Conclusions

The working prototype of the described system provides platform for future tests with blind people. The GPS position fix is accurate enough to inform visually impaired people about nearest POI. In cases when the GPS signal is too weak, the GSM based localization is being used, providing information POI. This information is useful for blind people, who are lost in unfamiliar area and they cannot get assistance from people nearby. The option of sending users' position via SMS is also very helpful. In the future, other localization algorithms will be tested to achieve better localization error.

7 References

[1] D. Dakopoulos, N. G. Bourbakis, "Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey". IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews 1/2010, pp. 25-35 (2010).

[2] M. S. Uddin, T. Shioyama, "Detection of Pedestrian Crossing and Measurement of Crossing Length - an Image-Based Navigational Aid for Blind People". IEEE Conference Publications of the IEEE Intelligent Transportation Systems (2005).

[3] H. S. Kim, "Advanced indoor localization using ultrasonic sensor and digital compass". International Conference on Control, Automation and Systems (2008).

[4] K. Sung, H. Kim, "Bayesian Navigation System with Particle Filtering and Dead Reckoning in Urban Canyon Environments". 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (2012)

[5] M. Ibrahim, M. Youssef, "CellSense: An Accurate Energy-Efficient GSM Positioning System". IEEE Transactions on Vehicular Technology, vol. 61, no. 1, pp. 286-296 (2012).

[6] I. Ahriz, Y. Oussar, B. Denby, G. Dreyfus, "Full-Band GSM Fingerprints for Indoor Localization Using a Machine Learning Approach". International Journal of Navigation and Observation, vol 2010.

[7] J. Chen, Z. Li, M. Dong, X. Wang, "Blind Path Identification System Design Base on RFID". IEEE Conference Publications of the International Conference on Electrical and Control Engineering (2010).

[8] S. Chumkamon, P. Tuvaphanthaphiphat, P. Keeratiwintakorn. "A Blind Navigation System Using RFID for Indoor Environments". IEEE Conference Publications of the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (2008).

[9] B. Landoni, "GSM Localizer". Elettronica In, pp. 54-65, Septemper 2009.

[10] http://www.opencellid.org/ accessed 4th April 2012.

A Study on Signal Interference in Clustered WiFi Networks

Jie Zhang, Hwa Jong Kim, and Goo Yeon Lee
Dept. of Comp. Eng., Kangwon National Univ. Chuncheon, Kangwon-Do, Korea

Abstract - *These days, many WiFi devices are trying to generate high throughput and have longer signal coverage which also bring unnecessary signal interference to neighboring wireless networks, and result in decreased network throughput. Signal interference could be minimized by reducing signal coverage of wireless devices. On the other hand, small signal coverage means low transmission power and low data throughput. In the paper, we analyze the relationship among signal strength and network throughput by simulation.*

Keywords: Signal interference, Throughput, WiFi network, Signal coverage

1 Introduction

Many WiFi networks are used these days, and they usually need high power for larger transmission range and higher data throughput. However, with rapid increasing number of wireless users such as smart phones, large signal coverage of WiFi causes higher signal interference among the WiFi devices which causes network quality deterioration especially in densely populated areas.

Signal interference is a common problem in wireless network, which may decrease throughput and cause security problems. The signal interference is inevitable due to wireless characteristics but could be reduced by controlling network configuration. [1] and [2] suggested topology control methods with changing transmission power in order to reduce signal interference in wireless ad-hoc network.

In WiFi networks, the interference can be managed by controlling the sender's transmission power. However, small signal power results in low throughput. There is a trade-off between signal interference and network throughput. Therefore, in the paper, we analyze the relationship among signal transmission power and network throughput with simulations to find a better way to improve quality of WiFi network.

2 Related work

Signal interference deteriorates network throughput and also wastes power. Reducing signal interference is widely researched in wireless ad-hoc network in order to minimize power consumption. Martin Burkhart et als. compared several topology control methods which claim to resolve interference, and proposed an interference-minimal method in wireless ad-hoc network with connectivity-preserving and spanner construction [1]. N. M. Karagiorgas introduced a multicost routing that constructs route with variable transmission power to reduce interference in ad-hoc network [2]. Sutep Tongngam [4] proposed a reducible transmission range approach for wireless network, which optimizes broadcasting latency. Ilenia Tinnirello and Giuseppe Bianchi analyzed the interference effects in WiFi networks [5]. Anand Kashyap et al. presented a passive monitoring of wireless traffic to estimate interference in WiFi networks [6].

3 Signal Strength and Network Throughput in WiFi network

We performed simulations to investigate signal strength and network throughput. The simulation topology is illustrated in Figure 1. A 120m x 120m area is divided into 9 cells, each cell contains one AP at the center and 6 devices are evenly located in each cell. Network throughput for downlink (from AP to clients) is measured for different transmission powers of 18dBm, 13dBm and 8dBm.

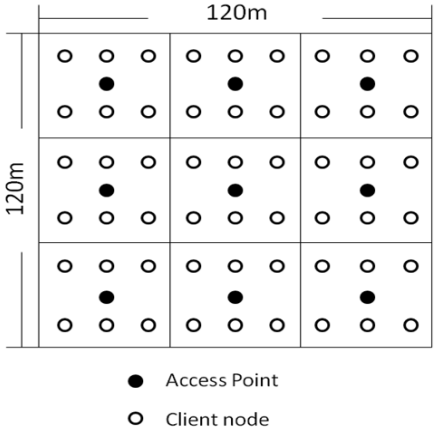


Fig. 1. Simulation Topology with 9 cells and 54 clients

In the simulation, we assumed all clients use same channel and channel association time is equally shared by all client devices no matter how the actual throughput is. We used Java for the simulations. Figures 2~4 show network throughput for each clients.

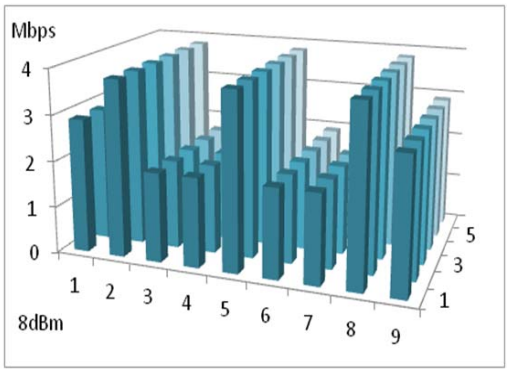


Fig. 2. Throughput distribution for 6 x 9 clients with transmission power 8dBm

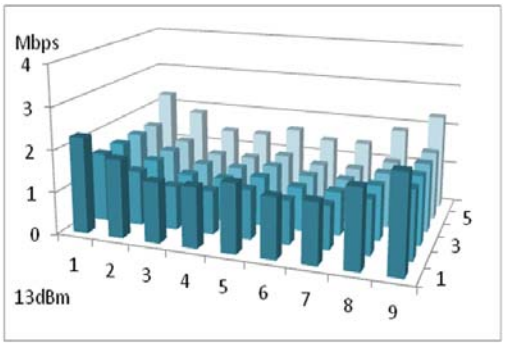


Fig. 3. Throughput distribution for 6 x 9 clients with transmission power 13dBm

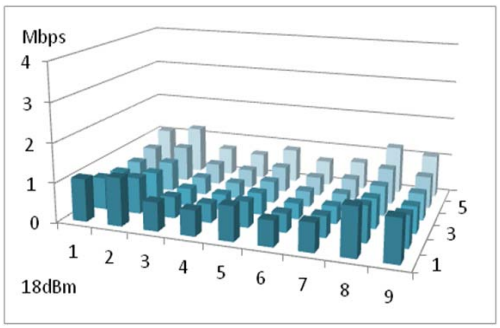


Fig. 4. Throughput distribution for 6 x 9 clients with transmission power 18dBm

Network throughput with transmission power of 8dBm showed the best performance because of the minimum interference. In Figure 2, clients at [2, y], [5, y], [8, y] (where [x, y] denotes client's position (x, y) in Figure 2) show best throughput because it is close to AP, which means high received signal strength. Edge clients at [1, y] and [9, y] show better throughput than clients in [3, y], [4, y], [6, y] and [7, y] which have same distance to APs but have less interference from other APs. Figures 3 and 4 also show that clients have higher throughput at the corners of the cell area.

High transmission power increases throughput in interference-free environments, however, today's wireless devices rarely show non-interference cases. For instance, from our smart phone's WiFi setting menu, the number of reachable APs just means the interference we may take from. Moreover, the number does not even include interference devices such as neighbor smart phones.

Low transmission power gives small signal coverage and less signal interference, resulting in higher throughput as shown in the above simulations. However, smaller signal coverage may also make shadow zone where clients may not reach APs.

4 Conclusions

Performance of local area WiFi network does not only depend on signal strength of AP but also on signal interference. In the paper, we analyzed the network throughput effects of signal interference via simulations. We found that network throughput could be maximized if transmission power of APs in WiFi network are properly controlled.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. NRF-2011-0013951), and also supported by Kangwon National University

5 References

[1] Martin Burkhart, Pascal von Rickenbach, Roger Wattenhofer, Aaron Zollinger, "Does topology control reduce interference?", Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing, 2004

[2] Karagiorgas, N.M., Kokkinos, P.C., Papageorgiou, C.A., Varvarigos, E.A, "Multicost Routing in Wireless AD-HOC Networks with Variable Transmission Power", Personal, Indoor and Mobile Radio Communications, IEEE 18th International Symposium, PP. 1 – 5, 2007

[3] Giuseppe Bianchi, "Performance analysis of the IEEE 802.11 Distributed Coordination Function", Selected Areas in Communications, Vol. 18, PP. 535 – 547, 2000

[4] Sutep Tongngam, "A Reducible Transmission Range Approach for Interference-Aware Broadcasting in Wireless Networks", International Conference on Future Information Technology, PP. 144 – 148, 2011

[5] Tinnirello, I. Bianchi, G., "Interference Estimation in IEEE 802.11 Networks", Control Systems, Vol. 30, PP. 30-43, 2010

[6] Anand Kashyap, Utpal Paul, Samir R. Das, "Deconstructing Interference Relations in WiFi Networks", Sensor Mesh and Ad Hoc Communications and Networks, 7th Annual IEEE Communications Society Conference, PP. 1 – 9, 2010

Key technologies of the Earth integrated information network

A. DING Ying [1], B. JIANG Hui-lin [1], C. YU Hai-yang [1], D. HU Yuan [1], E.ZHU Yi-feng[1], and F. CONG Li-gang[1]

[1] Key Laboratory of Education Ministry Optoelectronics Measurement & Control and Optical Information Transfer Technology, Changchun University of Science and Technology, Changchun, Jilin, China

Abstract - *With the rapid development of computer network technology, aerospace technology and communications technology, the demands of earth integrated information network become increasingly stronger in terms of national defense information and the development of the national economy.According to the course of foreign development process and the interpretation of networking mode,as well as the research on domestic-related advantages research unit, the paper elaborates the key technologies of the earth integrated network which need to be broken through. The key technologies involves architecture, routing, protocol, management, switch and so on, thus will provide a useful reference for our earth integrated information network.*

Keywords: earth integrated information networks; key technologies ;satellite communication; network architecture

1 Introduction

Since the concept of the Internet be brought forward, our lives has undergone enormous changes. Human beings have faced more challenges than ever before when they enjoy the high-speed information times. First and foremost, in one side of national blueprint and the people's livelihood, human is faced with the challenge of environmental disaster warning, resource exploration, deep space exploration perception, and the idea of moved to the planet. What is more, in another side of national security, the battlefield that human scramble for has constantly extended into outer space, and the advantage of air and space become a precondition for the information war. Last but not the least, in the other side of national economy, the movement towards informationization for our country still needs the support of the spatial information infrastructure and information services system, thus ensuring sustainable development of the economic society. Therefore, the earth integrated network , as an important means to promote sustained growth and the key to winning military information warfare, is especially essential.

Link of earth integrated network is mainly set up between base stations and satellite-to-satellite, satellite-to-sky or satellite-to-ground nodes. The earth integrated network can connect the user, aircraft and communications platforms among the land, sea, air and deep space, and it adopts high-speed intelligent- processing unit and supports exchange technology. According to the utilizing principle of maximum effective information resources, the information network can exact information accurately and process information fast and transmit communication efficiently, so it called space-based and sky-based and ground-based integrated network[1]. Achieving the network will not only improve the capacity and time effectiveness of communication, but also boost the reliability and survivability of information network, so a variety of applications integrated into an organic whole. Earth integrated network can provide powerful information support in terms of communications, navigation, positioning, monitoring, early alarm, remote sensing, detection and so on. It is a huge project to benefit the nation ,the people and the army[2].

In this paper, we analyse the key technologies to building earth integrated network. They all need to break through. Based on the unique characteristics of the space network, we provide a useful reference for the final implementation of the earth integrated network.

2 Research Situation of Interplanetary Internet

The Iridium System of the Motorola Company became successful in 1999. It broke the original convention of satellite communication system using GEO and transparent transponder communication. Since then, the space-based mobile communication network has gone through a narrow-band satellite communications network, the broadband satellite communications network and the space-based Internet[3]. NASA established a kind of integrated network architecture of communications satellites and other spacecraft in orbit [4-5](Figure 1).NASA carried out every parts of design and plan in detail.The architecture is represented by four architectural elements. They are backbone elements, access elements, proximity elements and other spacecraft.

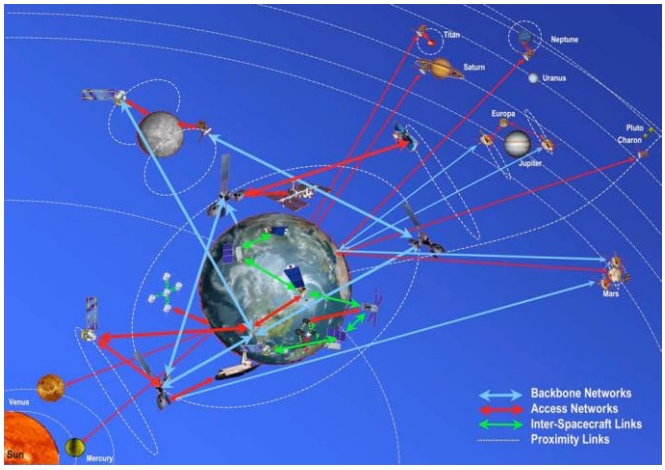


Figure 1: Integrated space communication architecture.

NASA defines that the communication network will use the OSI (Open System Interconnection) hierarchical model of the bottom layer to achieve the routing of IP data, shown in Figure 2. OSI satisfies end-to-end data routing capability. It is common protocols and interfaces at these layers that have ability to make links for all the nodes in the network.

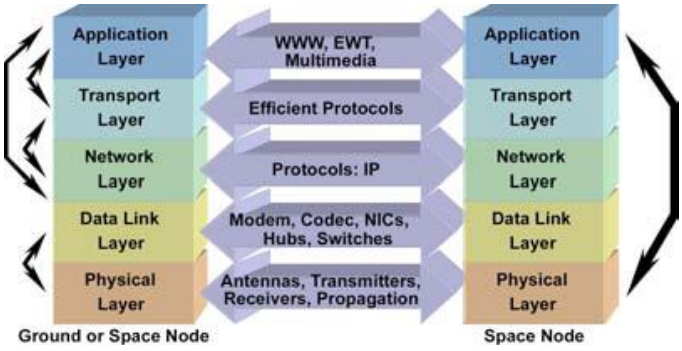


Figure 2:Internet protocol layers used in integrated communication architecture.

In addition, The Goddard Space F Center (GSFC) also makes a contribution to developing the space communications. It carried out another research project, as shown in Figure 3, called "Operating Mission as Nodes on the Internet (OMNI)". It aims to use commercial communication protocols on the ground. Then the COTS (commercial off-the-shelf) accorded with the ground commercial communications protocol. The all things achieve standardization of space communications.

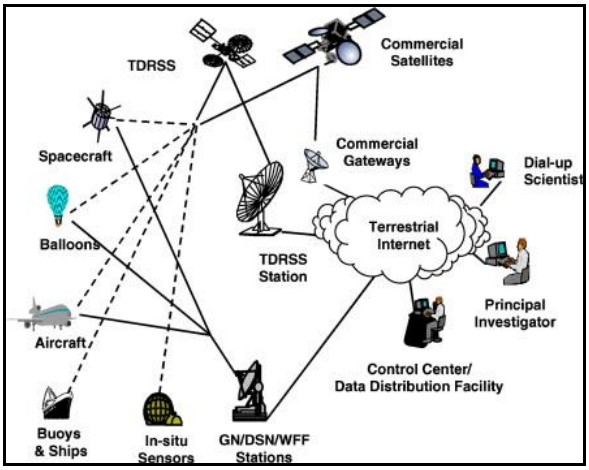


Figure 3:Operating Mission as Modes on the Internet (OMNI)

Other institutes like ESA also developed terrestrial transmission protocol called SLE (Space link extension) which based on the CCSDS (Consultative Committee for Space Data Systems).It has been incorporated into the architecture of CCSDS, so the Space Transfer Protocol of CCSDS will extend to the aerospace facilities on the ground.

3 Key Technologies

We want to combine space-based systems and ground-based system effectually, to achieve capability of information processing on satellite, to establish a reliable interstellar communication link and to form the earth integrated network of autonomous operation. What we need to do focus on the key technologies to breakthroughs[5-10]:

3.1 Optimized design of architecture

The network architecture design of the system is the top of the system construction and key issues. Earth integrated network involving space-based network, ground-based network and other multiple networks. It has the characters include large scale, complex structure, a wide range of business, and quickly changesd network topology. So it is an urgent breakthrough technology to design a network system which has a strong reliability, reasonable number of satellites and reasonable standard of orbital path. In order to ensure the network long-term stability, we should research the arrangement and set of network node, subnetwork operation process and the carrying capacity of business. Therefore, we also need to analyze the space mission and future plans in detail.As a core of a communications satellite, we take a special considerations for the distribution of mobile satellite node and the relationship of node. Constellation design, topology design and protocol specification help establish a mathematical model of the network architecture. The characters of earth integrated network conduce to establish constraint function and infer the optimal evaluation parameters.

It redound to building the earth integrated network that is help to accord with conditions and needs of common people.

3.2 Routing and protocol on the satellite

MEO/LEO satellite communications network have obvious advantages in improving the communication performance of the system, reducing the loss of communication and promoting the miniaturization of the mobile terminal. But the single satellite coverage is limited, and it can not be covering the same place for a long time.We need to achieve continuous global coverage seamless communication by the "relay" through multiple satellites. Routing and protocol on the satellite could realize the business data forwarded in a multi-hop routing, which is the key on the MEO/LEO satellite network with inter-satellite links. Due to the sustained high-speed movement of the satellite nodes, the satellite network topology changes periodically over time, the ground IP routing protocol can not be directly applied to the earth integrated network, and satellite communication is still used in the form of transparent forwarding but no inter-satellite link.The information distribution and processing are executed on the ground. Terminal data between space-based and ground-based is transmitted self-definedly by establishing a link through the frame form, and we received all information packet parsing and forwarding on the ground. Therefore, it is not only necessary to fully take into account the characteristics of space communication and space missions, but also as far as possible or existing agreements and technical to draw on the ground Internet, so it is essential that how to develop and reasonable allocate the satellite network layer routing protocols.

3.3 Network Management Technology

The management methods and management forms are very different of space-based and ground-based network. Due to the limited load and processing power of the satellite nodes, it is impossible to establish a the similar ground network management system; At the same time, integrated space-based component of the network is a dynamic wireless network, mobile users can be covered by any satellite to access satellite node via a wireless link, and space-based network communication with other users, which greatly increasing the difficulty of resource management. Therefore, we need to not only consider the traditional configuration management functionality, fault management capabilities, performance management, security management, and user management function, but also focus on considering network radio resource management and user mobility management technology.

In addition, due to the electromagnetic radiation in the space environment, satellite position with limited resources. The space-based satellite network is constantly moving and posture changing. Taking into consideration the factors of the network, the network management ensure premise of the normal operation, especially for the large complex environment.

3.4 Space switching technology

In order to avoid relay of multi-hop between earth station and satellite, and reach the goal of wide span single-hop, which can connection to reduce the transmission delay. That is a better solution of simplifing equipment and operating broadband communications satellite on-board processing and space switching technology. To achieve the exchange between the satellites, we need modulation, decoding, routing and other functions on the satellite estimating and allocation of network resources, which help us to establish the adaptive communication link.

The current space exchange is to achieve information sharing in the form of circuit-switched. To ensure data security and efficiency at the same time, achieved the automatic forwarding of the space-based routing ATM switch can improve the spectrum utilization, reduce the transmission delay and reduce the transmission error rate. So it is better to improve the communication quality of transmission, to enhance spatial link invulnerability. It plays an important role to improve the performance of the integrated space network. In addition, due to the satellite system updates hard, it is difficult to make changes and upgrades the system when it is determined Bandwidth and traffic control require further study at the same time.

3.5 System simulation and demonstration technology

Researching and constructing the earth integrated information network is a complex system engineering. For such a large, complex system, it adopts a lot of new technology. To fight only on paper is not enough, the ground simulation and verification is able to large system design and potential problems for effective verification and examination. It is not easily applied, so the digital and semi-physical simulation for the whole system and each technology is necessary.Using application software of network simulation, conduct a network simulation environment of the earth integrated information network by the way of an advanced simulation methods. For the spatial network architecture, routing technology, exchange of technology, access to technology, network management and other key technologies, we carry out simulation, demonstration and evaluation test. It will lay the foundation of building the earth integrated information network.

4 Conclusion

In this paper, we have described the key technologies of earth integrated information network, which can meet the needs of the national army and the ordinary people. The network can not only support mobile communication, disaster relief , intellectual traffic and so on, but also offer

transmission command, navigation, etc. These help us fully understand the current main direction, providing a reference for the network construction of integration of earth network. It is important and benefit to the sustainable development and has large strategic significance. With the deeply study and profounder understand, the technologies of earth integrated information network will come true.However, building the earth integrated information network is a systemic huge project. The movement of achievement it still has a long way to go.

5 References

[1] Shen Rong-jun. The mind of earth integrated network and spaceflight [J].China Engineering Science, 2006, 8(10):19-30.

[2] FARSEROTU J,PRASAD R.A.Survey of future broadband multimedia satellite systems,issues and trends[J].IEEE Communications Magazine,2000,(6):128-133.

[3] GUI Qi-shan, ZHAO Xin-guo, GUO Wei-min, GU Xiao-xia. Study on US Army Space Information System Establishment[J]. Network and Information Technology.2008,27（1）63-65.

[4] Kul Bhasin,Jeffrey Hayden. Space Internet Architectures and Technologies for NASA Enterprises,"Int. J. Satell.Commun. 2002; 20, 311-332.

[5] Kul Bhasin,Jeffrey Hayden. Developing Architectures and Technologies for an Evolvable NASA Space Communication Infrastructure[R]. National Aeronautics and Space Administration:Washington, DC 20546–0001.2004,7.

[6] Peng Chang-yan.The space network security key technology research[D].Changsha：National Defense Science and Technology University.2010.

[7] Zhang Jun.Space-based mobile communication network[M].National Defense Industry Press，2011.

[8] Zhi Ying-jian,Zhu Zi-jian.The empty world information network integrated management of preliminary study[C].XiAn：2010 Proceedings of frontier defence space information technology BBS，2010.

[9] Wang Zhen-yong. Structural design and analysis of multi-layered satellite network[D].Harbin：Harbin Institute of Technology, 2007.

[10] Space Communications and Navigation (SCaN) Network Architecture Definition Document (ADD) Volume 1: Executive Summary[R], Revision 2, 2011,10, NASA / SCaN.

SESSION

LATE BREAKING PAPERS: SENSOR NETWORKS, SECURITY AND EDUCATION, FADING CHANNELS ISSUES

Chair(s)

Prof. Hamid Arabnia
University of Georgia

Minimum Latency Aggregation Scheduling in Interference-Aware 3-Dimensional WSNs

Min Kyung An, Nhat X. Lam, D. T. Huynh, and Trac N. Nguyen
Department of Computer Science, University of Texas at Dallas, Richardson, Texas 75080
Emails: {mka081000, lxnhat, huynh, nguyentn}@utdallas.edu

***Abstract*—In this paper, we study the *Minimum Latency Aggregation Scheduling (MLAS)* problem in Wireless Sensor Networks (WSNs) adopting the two interference models: the *graph model* and the more realistic *physical interference model* known as Signal-to-Interference-Noise-Ratio (SINR). The main issue of the MLAS problem is to compute schedules with the minimum number of timeslots, that is, to compute the minimum latency schedules, such that data can be aggregated without any collision or interference. While existing works studied the problem in 2-dimensional (2D) WSNs only, we investigate the problem in the more general 3-dimensional (3D) WSNs, and introduce two approximation algorithms with $O(1)$-approximation ratios that yield schedules whose latency is bounded by $O(\Delta + R)$, where Δ is the maximum node degree and R is the network radius. To the best known of our knowledge, our results are the first results of the MLAS problem in 3D WSNs.**

***Keywords*-Data Aggregation, Interference-Aware, Signal-to-Interference-Noise-Ratio, 3-dimensional Network, Approximation Algorithm**

I. Introduction

A Wireless Sensor Network (WSN) consists of a number of sensor nodes which monitor nearby environmental conditions and gather data periodically. The gathered data is forwarded to a destination called the *sink* node. This type of application is commonly known as *data aggregation* in the literature, and it is one of the most crucial applications of WSNs.

Although recent advances in WSNs have led to the development of sensor nodes, the small-sized sensors still have limited energy resources. Therefore, researchers have focused on the issue of prolonging the network lifetime by reducing energy consumption which is caused by the unnecessary retransmission using sensors' limited power. An interesting approach is to assign *timeslots* to sensor nodes to obtain a good *schedule* by which data can be aggregated without any collision or interference. Since the data collection occurs periodically, reducing the *latency* of the schedule, that is, constructing schedules with a minimum number of timeslots, has been a fundamental issue.

In the literature, the problem of constructing minimum latency data aggregation schedules, namely the *Minimum Latency Aggregation Scheduling* problem, has been widely investigated by several researchers in two interference models: the *graph model* and the *physical interference model*. In the *collision-free graph model*, [1] proved the NP-hardness of the problem, and showed the $(\Delta - 1)$-approximation algorithm. Later, [2] introduced the first constant factor approximation algorithm whose latency is bounded by $23R + \Delta - 18$ which was improved by [3] and [4] to $16R + \Delta - 14$. In [5], Wan et al. proposed three approximation algorithms whose latency is bounded by $15R + \Delta - 4$, $2R + O(\log R) + \Delta$ and $\left(1 + O(\frac{\log R}{\sqrt[3]{R}})\right) R + \Delta$, respectively.

While these works considered only collision, some researchers have studied the problem taking into consideration interference as well in the *collision-interference-free graph model*. [5] and [6] proposed constant factor approximation algorithms whose latency is bounded by $O(\Delta + R)$, and [6] also proved an $\Omega(\log n)$ approximation lower bound in the metric model. Recently, [7] introduced a constant factor approximation algorithm whose latency is bounded by $O(R + \log n)$ assuming that multiple power levels are present, and the maximum power level is bounded, while [5], [6] assumed the uniform power model.

Recently, several researchers have started investigating data aggregation scheduling in the more realistic *physical interference model* known as Signal-to-Interference-Noise-Ratio (SINR). Unlike the graph model, the SINR model captures real world phenomena adequately by considering the *cumulative interference* caused by all the other concurrently transmitting nodes. The first investigation of the Minimum Latency Aggregation Scheduling problem in the SINR model was done by Li et al. [8]. [8] introduced a constant factor approximation algorithm whose latency is bounded by $O(\Delta + R)$ under the uniform power model. [9] extended it to the dual power model, and introduced two constant factor approximation algorithms whose latency is bounded by $O(\Delta + R)$. [9] also showed not only an $\Omega(\log n)$ approximation lower bound in the metric SINR model, but also its NP-hardness in the geometric SINR model. [10] proposed an algorithm that yields $O(\log^3 n)$-latency which was improved by [11] to $O(\log n)$. Recently, [7] introduced a constant factor approximation algorithm whose latency is bounded by $O(R + \log n)$. Note that [10] and [11] assumed the unlimited power model and [7] assumed that multiple power levels are present, and the maximum power level is bounded.

While these studies have been concerned with data aggregation, some other researchers have focused on related applications such as *broadcast* and *gossiping*. The broadcast problem is to distribute a unique message from a source (sink) node to all the other nodes, whereas the gossiping problem, which is also known as *all-to-all broadcast*, is to distribute the

message of each node to all the other nodes in the network. For the problem of broadcast, NP-hardness was proved by [12] which holds for the collision-free graph model, but not for the other models. For the gossiping problem, [13] proved its NP-hardness which holds for the SINR model under assumption that a node can combine messages and there is no limit on the length of the combined message.

Among works that studied the problems in 2-dimensional WSNs, [14] was the only one that has investigated the broadcast problem in both 2- and 3-dimensional WSNs. In 3-dimensional (3D) WSNs adopting the graph model, [14] introduced a constant factor approximation algorithm where the 3D space is partitioned into several truncated octahedrons.

In this paper, we continue the study of the Minimum Latency Aggregation Scheduling problem in 3D WSNs adopting both the graph model and geometric SINR model. While existing works studied the problem in 2-dimensional (2D) WSNs only, we investigate the problem in the more general 3-dimensional (3D) WSNs, and introduce two constant factor approximation algorithms that yield schedules whose latency is bounded by $O(\Delta + R)$. Our approximation algorithms for the problem are the first results, to the best of our knowledge, for 3D WSNs adopting both interference models.

This paper is organized as follows. Section II describes our network models and defines the Minimum Latency Aggregation Scheduling (MLAS) problem. In Section III, we show our 3D-space-filling and labeling techniques. Section IV introduces two constant factor approximation algorithms for the MLAS problem, and we analyze them in Section V. Finally, Section VI contains some concluding remarks.

II. Preliminaries

A. 3D Network Models

In this paper, a wireless sensor network (WSN) consists of a set V of sensor nodes deployed in a 3-dimensional (3D) space, and each node $u \in V$ is assigned a transmission power level $p(u)$. Accordingly, a directed edge (u, v) exists from node u to node v, if v resides in the *transmission ball* with radius $p(u)$ of u, i.e., $d(u, v) \leq p(u)$, where $d(u, v)$ denotes the Euclidian distance between u and v.

1) Graph Model: In the graph model, let $B^u_{p(u)} = \{v \,|\, v \in V,\, d(u, v) \leq p(u)\}$ denote the set of all nodes that can be reached by u with the power level $p(u)$. If two nodes u and v reside in the transmission ball of each other, i.e., $u \in B^v_{p(v)}$ and $v \in B^u_{p(u)}$, then u and v can communicate. However, we also need to consider the collision or interference. Given a power level $p(u)$ of u, the *interference ball* of u is defined as a ball with radius $\rho \cdot p(u)$, where $\rho \geq 1$ is the interference factor. (Seel Figure 1.) Given $\rho \geq 1$, let $I^u_{p(u)} = \{v \,|\, v \in V,\, d(u, v) \leq \rho \cdot p(u)\}$ denote the set of all nodes in the interference ball of u. Then, *collision* (or *conflict*) is said to occur at a receiver node w if there exist other concurrently sending nodes u and v such that $w \in B^u_{p(u)} \cap I^v_{p(v)}$, where $\rho = 1$. On the other hand, *interference* is said to occur at w if there exist other concurrently sending nodes u and v such that $w \in B^u_{p(u)} \cap I^v_{p(v)}$, where $\rho > 1$. In the literature, the graph model concerning only collision (i.e., when $\rho = 1$) is called the *collision-free graph model*, whereas the graph model concerning both collision and interference (i.e., when $\rho \geq 1$) is called the *collision-interference-free graph model*.

In the graph model, the communication graph can be modeled as a bidirectional ball graph $G(V, E)$, where $E = \{(u, v) \,|\, u, v \in V,\, d(u, v) \leq p(u) \text{ and } d(v, u) \leq p(v)\}$.

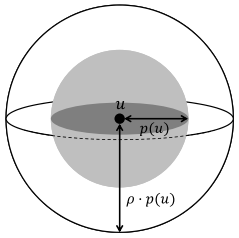


Fig. 1: Transmission ball and interference ball of u

2) SINR Model: In the physical interference model (SINR) [15], if a node u transmits with its power level $p(u)$, then the received power at a receiver v is $p(u) \cdot d(u, v)^{-\alpha}$, where $\alpha \in [2, 6]$ is the *path loss exponent*. In order that the receiver v can receive the data transmitted by the sender u, the ratio of the received power at v to the interference caused by all the other concurrently transmitting nodes and background noise must be beyond an SINR threshold $\beta \geq 1$. Formally, node v can successfully receive data via the communication edge (u, v) only if

$$SINR_{(u,v)} = \frac{\frac{p(u)}{d(u,v)^\alpha}}{N + \sum_{w \in X - \{u,v\}} \frac{p(w)}{d(w,v)^\alpha}} \geq \beta \qquad (1)$$

where $N > 0$ is the background noise, and X is the set of other concurrently transmitting nodes. Observing that u can send its data to the nodes within the distance $(\frac{p(u)}{N\beta})^{\frac{1}{\alpha}}$, the network can be modeled as a directed ball graph $G(V, E)$, where $E = \{(u \to v) | u, v \in V, d(u, v) \leq (\frac{p(u)}{N\beta})^{\frac{1}{\alpha}}\}$.

Note that in the literature, a communication graph is called a *ball graph (BG)* if all nodes are assigned various power levels, and if all nodes are assigned the same power level, it is called a *unit ball graph (UBG)*.

B. Problem Definition

The Minimum Latency Aggregation Scheduling (MLAS) problem is defined as follows. Considering a set of nodes in a 3D space, we assign these nodes a number of *timeslots* such that nodes scheduled to send data at the same timeslot can send data to its receivers *simultaneously* without any collision or interference. A schedule is defined as a sequence of such timeslots. Formally, at each timeslot t, we have an *assignment vector* $\pi_t = \langle (s_{t_1}, p(s_{t_1})), \cdots, (s_{t_m}, p(s_{t_m})) \rangle$ in which s_{t_i} is assigned to send data with its power level $p(s_{t_i})$, $1 \leq i \leq m$, and

- (Graph Model) neither collision nor interference occurs at any receiver r, or

- (SINR Model) the SINR threshold inequality is satisfied for all receivers r,

where (s_{t_i}, r) is an edge in the communication graph $G(V, E)$.

A *schedule* is a sequence of assignment vectors $\Pi = (\pi_1, \pi_2, \cdots, \pi_M)$, where M is the length of the schedule which is also called its *latency*. A schedule Π is *successful* if all data of each node $v \in V$ is aggregated to a sink node $s \in V$.

Input. A set V of nodes in a 3-dimensional (3D) space, a sink node $s \in V$.

Output. A successful minimum latency schedule.

C. NP-Hardness of the MLAS Problem in 3D WSNs

Note that [1], [6] and [9] showed the NP-hardness of the MLAS problem in the 2-dimensional (2D) collision-free graph model, collision-interference-free graph model, and SINR model, respectively. Observing that a 2D WSN is a special case of 3D WSNs, the MLAS problem in 3D WSNs is also NP-hard.

III. 3-Dimensional Space Filling

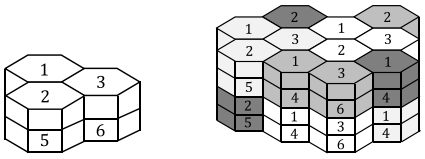
In this section, we introduce our 3-dimensional (3D) space-filling technique. It has been known that there are only five space-filling convex polyhedra [16]: *triangle prism*, *cube*, *hexagonal prism*, *truncated octahedron*, and *gyrobifastigium*. In this paper, we fill the 3D space with the cube, which is the only platonic solid, and the hexagonal prism.

A. Space Filling with Cubes

We partition the 3D space containing the network into cubes whose side length a is $\frac{r}{\sqrt{3}}$, and space diagonal is r. Each cube is labeled with the label $CL(x, y, z)$ if (x, y, z) is its vertex with the smallest $x-$, $y-$ and $z-$ coordinates.

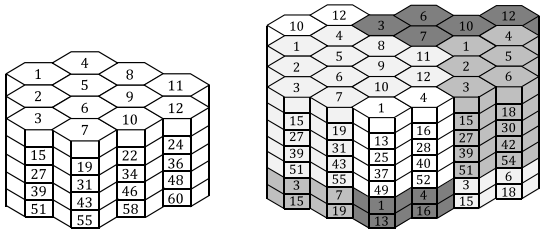
B. Space Filling with Hexagonal Prisms

The 3D space containing the network is tessellated with hexagonal prisms whose side length a is $\frac{r}{\sqrt{5}}$, and space diagonal is r. The hexagonal prisms are labeled using $(9k^3 - 3k^2)$-labeling, where k is a positive integer. Figure 2(a) shows a (3×2)-labeling when $k = 1$. The (3×2)-labeling consists of 2 layers each of which consists of 3 hexagonal prisms. Figure 2(b) shows an example of filling a 3D network space using (3×2)-labeling. In this (3×2)-labeling, we can observe that the distance between two hexagonal prisms with the same label is $\frac{r}{\sqrt{3}}$. Here, the distance between any two hexagonal prisms, denoted by hex_i and hex_j, respectively, is defined as the distance between two closest vertices p in hex_i, and p' in hex_j. Next, Figure 3(a) shows (12×5)-labeling when $k = 2$, and it consists of 5 layers each of which consists of 12 hexagonal prisms. Figure 3(b) shows an example of hexagonal-prism tessellation on a 3D network using (12×5)-labeling. In this (12×5)-labeling, the distance between two hexagonal prisms with the same label is $4 \cdot \frac{r}{\sqrt{3}}$. When $k = 3$, we have (27×8)-labeling that consists of 8 layers each of which consists of 27 hexagonal prisms. When a 3D network is filled with the hexagonal prisms with (27×8)-labeling, the distance between the hexagonal prisms is $7 \cdot \frac{r}{\sqrt{3}}$.



(a) (b)

Fig. 2: (a) 6-labeling ($k = 1$) (b) Space-filling with 6-labeling



(a) (b)

Fig. 3: (a) 60-labeling ($k = 2$) (b) Space-filling with 60-labeling

In general, we have a $K = (3k^2 \times (3k-1))$-labeling, and the distance between two hexagonal prisms with the same label is $(3k - 2) \cdot \frac{r}{\sqrt{3}}$.

IV. Constant Factor Approximation Algorithms

In this section, we introduce two constant factor approximation algorithms for the MLAS problem in graph model and the physical interference (SINR) model in a 3-dimensional (3D) wireless sensor network. We assume the uniform power model where all nodes are initially assigned a uniform power level P. We further make the following assumptions:

- For the *graph model*, we set the maximum link length $r = P$, and assume that the undirected unit ball graph $G = (V, E)$, where $E = \{(u, v) \,|\, d(u, v) \leq r\}$, is connected and the interference factor $\rho \geq 1$.
- For the *SINR model*, notice that if node u on link (u, v) of the maximum link length $r_{max} = (\frac{P}{N\beta})^{\frac{1}{\alpha}}$ is transmitting, then the node u can be the only sending node, i.e., none of remaining nodes can transmit concurrently with u. Thus, we consider only links (u, v), where $d(u, v) \leq \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$ for some constant $\delta \in (0, 1)$ as considered in [8]. Thus, in the SINR model, we set $r = \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$, and assume that the undirected unit ball graph $G = (V, E)$, where $E = \{(u, v) \,|\, d(u, v) \leq r\}$, is connected and the path loss exponent $\alpha > 3$.

A. Data Aggregation Tree Construction

We start this section by introducing some standard notations that are used subsequently:

- *Graph Center*: Given a communication graph $G = (V, E)$, we call node c a *center* node if the distance from c to the farthest node from c is minimum.

- *Maximal Independent Set* (MIS): A subset $V' \subseteq V$ of the graph G is said to be *independent* if for any vertices $u, v \in V'$, $(u, v) \notin E$. An independent set is said to be *maximal* if it is not a proper subset of another independent set.
- *Connected Dominating Set* (CDS): A *dominating set* (DS) is a subset $V' \subseteq V$ such that every vertex v is either in V' or adjacent to a vertex in V'. A DS is said to be *connected* if it induces a connected subgraph.

Our algorithm assigns timeslots to nodes based on a data aggregation tree whose construction is based on that of a virtual backbone tree in [17]. The construction of a virtual backbone tree in [17] also works for 3D networks although it was originally introduced for 2D networks. In order to build a data aggregation tree T, we first choose a center node c, then construct a breadth-first-search (BFS) tree (cf. [18]) on G rooted at node c so that the latency can be bounded in terms of the network radius R rather than its diameter. Based on the BFS tree obtained, we construct a data aggregation tree as done in [17]. [17] first finds an MIS layer by layer, which is the depth on the BFS tree. Let us call the nodes in the MIS *dominators*, and the others *dominatees*. Next, [17] obtains a CDS of G by connecting the dominators using some upper level *connectors* that were originally dominatees, and connecting the connectors to their upper level or same level dominators. If there exist some remaining dominatees that are not connected to the CDS, then each of such dominatees is connected to its neighboring dominator. We denote the newly formed tree by T, and use it as the data aggregation tree in our algorithm.

B. Data Aggregation Scheduling

1) Cube-based Aggregation Scheduling Algorithm: The first algorithm called the *Cube-based Aggregation Scheduling (CBAS)* algorithm starts by partitioning the network into cubes as described in Section III-A, and a number of iterations are performed to find a schedule based on T. Assigning timeslots is also based on the constant value C that guarantees that any two senders can send data to their receivers at the same time if they are C cubes apart from each other. Now, the constant C is set as follows in the two different interference models:

- Graph model: $C = \lceil \rho \cdot \sqrt{3} + 3 \rceil$
- SINR model: $C = \lceil (\frac{P \cdot 4\pi^2}{N(\delta^{-\alpha}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} \cdot \frac{\sqrt{3}}{\delta} (\frac{N\beta}{P})^{\frac{1}{\alpha}} + 2 \rceil$

Then, the CBAS algorithm can be used not only for the graph model, but also for SINR model where the constant C is defined accordingly.

Algorithm 1 shows the details of the CBAS algorithm. It schedules the nodes in T starting with the dominatees so that they can send data without any collision or interference to their dominators (Steps 4 – 5 in Algorithm 1). While scheduling dominatees, the CBAS algorithm (Algorithm 1) uses Algorithm 2 as a subroutine to assign the same timeslots to the dominatees if they are C cubes away from each other. Note that while scheduling, we pick only one dominatee in each cube arbitrarily (Step 2 in Algorithm 2), and repeat the scheduling procedure (Algorithm 2) until all dominatees are scheduled.

Algorithm 1 Cube-based Aggregation Scheduling (CBAS)

Input: A set V of nodes in a 3D space
Output: Length of Schedule

1: Fill the space with cubes whose edge length is $\frac{r}{\sqrt{3}}$.
2: Construct a data aggregation tree T using the algorithm in [2] rooted at a center node c.
3: Set the first timeslot $t \leftarrow 1$
4: $S_d \leftarrow$ the set of dominatees
5: $t \leftarrow$ TA-C(S_d, t)
6: **for** $i = R$ to 1 **do**
7: $S_d^i \leftarrow$ the set of dominators at level i in T
8: **if** $S_d^i \neq \emptyset$ **then** $t \leftarrow$ TA-C(S_d^i, t)
9: $S_c^i \leftarrow$ the set of connectors at level i in T
10: **if** $S_c^i \neq \emptyset$ **then** $t \leftarrow$ TA-C(S_c^i, t)
11: **end for**
12: Send the aggregated data from the center node c to the sink node s via a shortest path f.
13: **return** $(t-1)+$ length of f

Algorithm 2 TimeSlot Assignment (TA-C)

Input: A set S of sender nodes and a starting timeslot t
Output: Timeslot t

1: **while** $S \neq \emptyset$ **do**
2: Pick one node $v_s \in S$ in each cube. Let $S' \subseteq S$ be the set of such nodes.
3: **for** $t_1 = 0, ..., C$, $t_2 = 0, ..., C$, $t_3 = 0, ..., C$ **do**
4: $S'' \leftarrow \emptyset$, $S'' \leftarrow \{v_s | v_s \in S'$ with $CL(x, y, z)$ such that $t_1 = x \mod (C+1)$, $t_2 = y \mod (C+1)$ and $t_3 = z \mod (C+1)\}$
5: **if** $S'' \neq \emptyset$ **then**
6: **for** each $v_s \in S''$ **do**
7: $TS(v_s) \leftarrow t$
8: **end for**
9: $t \leftarrow t+1$, $S \leftarrow S - S''$, $S' \leftarrow S' - S''$
10: **end if**
11: **end for**
12: **end while**
13: **return** t

After all dominatees are scheduled, several iterations are performed (Steps 6 – 11 in Algorithm 1) to schedule the remaining dominators and connectors level by level until all of them are scheduled, as follows. At each iteration for level i of T, if there exist dominators that have just received data from their lower level dominatees or connectors at level $i+1$, then the dominators are scheduled to send the aggregated data to their upper level connectors at level $i-1$. Otherwise, if there exist connectors that have just received data from their lower level dominators at level $i+1$, then the connectors are scheduled to send the aggregated data to their upper level dominators at level $i-1$. While scheduling, the CBAS

algorithm (Algorithm 1) also uses Algorithm 2 as a subroutine to assign the same timeslots to them if they are C cubes away from each other. Notice that there exists only one dominator in a cube, whereas there may exist several connectors in a cube. Therefore while scheduling connectors, at each level i, we pick only one connector in each cube arbitrarily (Step 2 in Algorithm 2), and repeat the scheduling procedure (Algorithm 2) until all connectors at level i are scheduled. Once all data is aggregated to the center node c, c sends the aggregated data to the sink node s via a shortest path (Step 12 in Algorithm 1).

2) Hexagonal-Prim-based Aggregation Scheduling: The second algorithm called Hexagonal-Prim-based Aggregation Scheduling (HPBAS) starts by partitioning a network into hexagonal prisms which are labeled using $K = (9k^3 - 3k^2)$-labeling as described in Section III-B, and a number of iterations are performed to find a schedule based on T obtained as the CBAS algorithm. Assigning timeslots is also based on the constant value K that guarantees that any two senders can send data to their receivers at the same time if they are located in the hexagonal prisms with the same label according to the K-labeling. Let us set the constant K as follows in the two different interference models:

- Graph model: $K = 9k^3 - 3k^2$, where $k = \lceil \frac{\sqrt{5}(\rho+2)+2}{3} \rceil$
- SINR model: $K = 9k^3 - 3k^2$, where $k = \lceil \frac{1}{3}\{ \frac{\sqrt{5}}{\delta} \cdot (\frac{N\beta}{P})^{\frac{1}{\alpha}} \cdot (\frac{P \cdot 4\pi^2}{N(\delta^{-1}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} + \sqrt{5} + 2\} \rceil$

The HPBAS algorithm can be used not only for the graph model, but also for the SINR model where the constant K is defined accordingly.

Algorithm 3 shows the details of the HPBAS algorithm. Similar to CBAS, the HPBAS algorithm first schedules the dominatees (Steps 4 – 5 in Algorithm 3), and then several iterations are performed to schedule the remaining dominators and connectors (Steps 6 – 11 in Algorithm 3). While scheduling, HPBAS uses Algorithm 4 as a subroutine to assign the same timeslots to node which are located in the hexagonal prisms with the same label. As the final step, the aggregated data at the center node c is sent to the sink node s via a shortest path (Step 12 in Algorithm 3).

Algorithm 3 Hexagonal-Prism-based Scheduling

Input: A set V of nodes in a 3D space
Output: Length of Schedule

1: Fill the space with hexagonal prims whose side length is $\frac{r}{\sqrt{5}}$, and label the prisms using K-labeling.
2: Construct an aggregation tree T using an algorithm in [2] rooted at a center node c.
3: Set the first timeslot $t \leftarrow 1$
4: $S_d \leftarrow$ the set of dominatees of V.
5: $t \leftarrow$ TA-X(S_d, t)
6: **for** $i = R$ to 1 **do**
7: $S_d^i \leftarrow$ the set of dominators at level i in T
8: **if** $S_d^i \neq \emptyset$ **then** $t \leftarrow$ TA-X(S_d^i, t)
9: $S_c^i \leftarrow$ the set of connectors at level i in T
10: **if** $S_c^i \neq \emptyset$ **then** $t \leftarrow$ TA-X(S_c^i, t)
11: **end for**
12: Send the aggregated data from the center node c to the sink node s via a shortest path f.
13: **return** $(t-1)+$ length of f

Algorithm 4 TimeSlot Assignment (TA-X)

Input: A set S of sender nodes and a starting timeslot t
Output: Timeslot t

1: **while** $S \neq \emptyset$ **do**
2: Pick one node $v_s \in S$ in each cube. Let $S' \subseteq S$ be the set of such nodes.
3: **for** $i = 1$ to K **do**
4: $S'' \leftarrow \emptyset$, $S'' \leftarrow \{v_s | v_s \in S'$ with $HL(v_s) = i\}$
5: **if** $S'' \neq \emptyset$ **then**
6: **for** each $v_s \in S''$ **do**
7: $TS(v_s) \leftarrow t$
8: **end for**
9: $t \leftarrow t+1$, $S \leftarrow S - S''$, $S' \leftarrow S' - S''$
10: **end if**
11: **end for**
12: **end while**
13: **return** t

V. Analysis

In this section, we analyze the Cube-Based Aggregation Scheduling (CBAS) and Hexagonal-Prim-Based Aggregation Scheduling (HPBAS) algorithms (Algorithms 1 and 3).

A. Analysis of CBAS Algorithm

First, we analyze the CBAS algorithm, and bound the latency of the schedule produced by it. We first prove that any two senders can send data at the same time without any collision and interference if they are C cubes apart.

Lemma 1 (Graph Model)**.** Let $C = \lceil \rho \cdot \sqrt{3} + 3 \rceil$, where $\rho \geq 1$ is the interference factor. Then any two sender nodes that are at least C cubes apart from each other can concurrently send data without any collision and interference.

Proof: Consider a sender node v_i trying to send data to its receiver v_j, and the farthest sender node v_k that interferes with v_j. Then, $d(v_k, v_j) \leq \rho \cdot r$.

Next, letting z denote the number of cubes between v_j and v_k, we bound z as follows. Consider the straight line between v_j and v_k. Then, as $\frac{r}{\sqrt{3}} \cdot z \leq d(v_k, v_j)$, we have $z \leq d(v_k, v_j) \cdot \frac{\sqrt{3}}{r}$ which implies that $z \leq \rho \cdot \sqrt{3}$. Therefore, there are at most $\lceil \rho \cdot \sqrt{3} \rceil$ cubes between v_j and v_k, and any other sender must be at least $\lceil \rho \cdot \sqrt{3} + 1 \rceil$ cubes apart from the node v_j not to cause interference. Now, observing that the number of cubes between v_i and v_j is at most 1, and considering the cube in which v_j is located, we can set $C = \lceil \rho \cdot \sqrt{3} + 3 \rceil$. ■

Lemma 2 (SINR Model)**.** For SINR threshold $\beta \geq 1$, path loss exponent $\alpha > 3$, background noise $N > 0$, and some constant $\delta \in (0, 1)$, let

$$C = \lceil (\tfrac{P \cdot 4\pi^2}{N(\delta^{-\alpha}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} \cdot \sqrt{3} \cdot (\delta(\tfrac{P}{N\beta})^{\frac{1}{\alpha}})^{-1} + 2 \rceil$$

Then any two sender nodes that are at least C cubes away from each other can send data at the same time.

Proof: Consider a sender node v_i trying to send data to its farthest possible receiver v_j, i.e., $d(v_i, v_j) = \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$. In order that the receiver v_j receives data from the sender v_i without interference, for all other concurrently sending nodes, the following must be satisfied:

$$\frac{P\left(\delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}\right)^{-\alpha}}{N + \sum_{v_i' \notin \{v_i, v_j\}} P \cdot d(v_i', v_j)^{-\alpha}} \geq \beta$$

which implies

$$\frac{P \sum_{v_i' \notin \{v_i, v_j\}} \frac{1}{d(v_i', v_j)^{\alpha}}}{N(\delta^{-\alpha} - 1)} \leq \frac{P \int_0^{2\pi} \int_0^{2\pi} \int_x^{\infty} \frac{y^2}{y^{\alpha}} \, dy \, d\theta \, d\varphi}{N(\delta^{-\alpha} - 1)} \quad (2)$$

$$= \frac{P \cdot 4\pi^2 \int_x^{\infty} y^{2-\alpha} \, dy}{N(\delta^{-\alpha} - 1)} \quad (3)$$

$$= \frac{P \cdot 4\pi^2 \cdot x^{3-\alpha}}{N(\delta^{-\alpha} - 1)(\alpha - 3)} \leq 1 \quad (4)$$

where x is the shortest distance between v_j and one of the other concurrently sending nodes. From inequality (4), we get $x \geq (\frac{P \cdot 4\pi^2}{N(\delta^{-\alpha}-1)(\alpha-3)})^{\frac{1}{\alpha-3}}$. Thus $X := (\frac{P \cdot 4\pi^2}{N(\delta^{-\alpha}-1)(\alpha-3)})^{\frac{1}{\alpha-3}}$ is a lower bound for x.

Next, let us bound the number of cubes between v_j and a closest concurrently sending node to v_j, say v_i'. Let z be the number of cubes between v_i' and v_j. We bound z as follows. Consider the straight line between v_i' and v_j, and the cubes lying on the line. We have $X \leq z \cdot \frac{1}{\sqrt{3}} \cdot \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$ which implies $X \cdot \sqrt{3} \cdot (\delta(\frac{P}{N\beta})^{\frac{1}{\alpha}})^{-1} \leq z$. Therefore, v_j and v_i' should be at least $Z := \lceil X \cdot \sqrt{3} \cdot (\delta(\frac{P}{N\beta})^{\frac{1}{\alpha}})^{-1} \rceil$ cubes apart. Now, observing that the number of cubes between v_i and v_j is at most 1, and considering the cube in which v_j is located, we can set $C = \lceil (\frac{P \cdot 4\pi^2}{N(\delta^{-\alpha}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} \cdot \sqrt{3} \cdot (\delta(\frac{P}{N\beta})^{\frac{1}{\alpha}})^{-1} + 2 \rceil$. ■

Lemma 3. The number of connectors in a cube is at most $7^3 - 1$.

Proof: Consider a dominator v in a cube, denoted by $cube(v)$, and its connectors. As the connectors connect dominators which are 2-hops away from v in the CDS, the number of connectors in one cube cannot exceed the number of dominators that are at most 2 hops away from v. Thus, it is sufficient to bound the number of such dominators.

Consider a ball, whose radius is $2r$, that is totally contained within a cube, denoted by $cube'$, with the side length $4r$ (See Figure 4). Then, the number of 2-hop away dominators cannot exceed the number of cubes whose side length is $\frac{r}{\sqrt{3}}$ within $cube'$. As there exist at most $\lceil 4\sqrt{3} \rceil^3$ cubes whose side length is $\frac{r}{\sqrt{3}}$ within $cube'$, there are at most $\lceil 4\sqrt{3} \rceil^3$ dominators in $cube'$. This implies that there exist at most $7^3 - 1$ connectors for v, and therefore at most $7^3 - 1$ connectors in a cube. ■

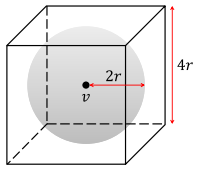


Fig. 4: The inscribed ball that touches each face of the cube.

Lemma 4 (Lower Bound, Graph Model)**.** [4] In order to produce a successful schedule, any data aggregation scheduling algorithm requires

- $\geq \max\{\Delta, \log R\}$ timeslots, for $\rho = 1$,
- $\geq \max\{\frac{\Delta}{\phi}, R\}$ timeslots, where $\phi = \frac{2\pi}{\lfloor \arcsin \frac{\rho-1}{2\rho} \rfloor}$, for $1 < \rho < 3$, and
- $\geq \max\{\Delta, R\}$ timeslots, for $\rho \geq 3$.

Lemma 5 (SINR Model)**.** [8] For any node, at most $\omega = \frac{r^{\alpha}}{\beta} - 1$ neighboring nodes can send data at the same time, where $r = \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$.

Corollary 6 (Lower Bound, SINR Model)**.** In order to produce a successful schedule, any data aggregation scheduling algorithm requires $\geq \max\{\frac{\Delta}{\omega}, R\}$ timeslots.

Theorem 7. *The CBAS algorithm produces a successful schedule whose latency is bounded by $O(\Delta + R)$, and it is therefore a constant-factor approximation algorithm.*

Proof: First, consider the Steps 4 – 5 in Algorithm 1 that schedules dominatees. In a cube, there exist at most Δ dominatees sharing one dominator v. Thus, gathering data from all the dominatees to the corresponding dominators takes at most $\Delta \cdot (C + 1)^2$ timeslots.

Next, consider Steps 6 – 11 in Algorithm 1 that schedule the remaining dominators and connectors. For each iteration that schedules nodes at level i, we consider the following cases:

1) Assigning timeslots to dominators at level i to send data to their connectors at level $i - 1$: In this case, there exists at most 1 dominator in a cube, and therefore gathering data from all the dominators at level i to the corresponding connectors at level $i - 1$ takes at most $(C + 1)^2$ timeslots.
2) Assigning timeslots to connectors at level i to send data to their dominators at level $i - 1$: In this case, there exist at most $7^3 - 1$ connectors in a cube (Lemma 3), and therefore gathering data from all the connectors at level i to the corresponding dominators at level $i - 1$ takes at most $(7^3 - 1)(C + 1)^2$ timeslots.

As Steps 6 – 11 repeat at most R times, it takes at most $7^3(C + 1)^2 \cdot R$ timeslots.

Finally, as Step 12 in Algorithm 1 takes at most R timeslots, the latency of CBAS is bounded by $(C+1)^2 \cdot \Delta + 7^3(C+1)^2 \cdot R = O(\Delta + R)$. Thus, it is a constant factor approximation in both the graph model and the SINR model by Lemma 4 and Corollary 6. ■

B. Analysis of HPBAS Algorithm

We now analyze the HPBAS algorithm, and bound the latency of the schedule produced by it. We first prove that any two senders can send data at the same time without any collision and interference if they are located in the hexagonal prisms with the same label according to the K-labeling.

Lemma 8 (Graph Model)**.** Let $k = \lceil \frac{\sqrt{5}(\rho+2)+2}{3} \rceil$, where $\rho \geq 1$ is the interference factor. Then any two sender nodes which are located in the hexagonal-prisms with the same label according to the K-labeling can send data at the same time.

Proof: Consider a sender node v_i sending its data to its farthest possible receiver v_j, i.e., $d(v_i, v_j) = r$. The proof of Lemma 1 showed that the farthest distance between v_j and the other sender node that interferes with v_j, say v_k, is at most $\rho \cdot r$. As the maximum distance between any two nodes is r, if any two senders are $\rho \cdot r + 2r$ distance apart from each other, then they can send data at the same time.

Next, the distance between two hexagonal prisms with the same label is $\frac{r}{\sqrt{5}}(3k-2)$ in $K = (9k^3 - k^2)$-labeling. Letting $\frac{r}{\sqrt{5}}(3k-2) \geq \rho \cdot r + 2r$, we can set $k = \lceil \frac{\sqrt{5}(\rho+2)+2}{3} \rceil$. ■

Lemma 9 (SINR Model)**.** For SINR threshold $\beta \geq 1$, path loss exponent $\alpha > 3$, background noise $N > 0$, and some constant $\delta \in (0, 1)$, let

$$k = \lceil \tfrac{1}{3} \{ \tfrac{\sqrt{5}}{\delta} \cdot (\tfrac{N\beta}{P})^{\frac{1}{\alpha}} \cdot (\tfrac{P \cdot 4\pi^2}{N(\delta^{-1}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} + \sqrt{5} + 2 \} \rceil$$

Then any two sender nodes which are located in the hexagonal-prisms with the same label according to the K-labeling can send data at the same time.

Proof: Consider a sender node v_i sending its data to its farthest possible receiver v_j, i.e., $d(v_i, v_j) = r = \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$. The proof of Lemma 2 showed that the shortest distance between v_j and the other concurrently sending node is at least $(\frac{P \cdot 4\pi^2}{N(\delta^{-1}-1)(\alpha-3)})^{\frac{1}{\alpha-3}}$. Since the maximum distance is $\delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$, any two sender nodes can send data at the same time if they are at least of distance $D := (\frac{P \cdot 4\pi^2}{N(\delta^{-1}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} + \delta(\frac{P}{N\beta})^{\frac{1}{\alpha}}$ apart from each other.

Next, the shortest distance between two hexagons with the same label is $\frac{r}{\sqrt{5}}(3k-2)$ in $K = (9k^3 - k^2)$-labeling. Letting $\frac{r}{\sqrt{5}}(3k-2) \geq D$, we can set $k = \lceil \frac{1}{3} \{ \frac{\sqrt{5}}{\delta} \cdot (\frac{N\beta}{P})^{\frac{1}{\alpha}} \cdot (\frac{P \cdot 4\pi^2}{N(\delta^{-1}-1)(\alpha-3)})^{\frac{1}{\alpha-3}} + \sqrt{5} + 2 \} \rceil$. ■

Lemma 10. The number of connectors in a hexagonal prism is at most $7^3 - 1$.

Proof: Omitted. ■

Theorem 11. *The HPBAS algorithm produces a successful schedule whose latency is bounded by $O(\Delta + R)$, and it is therefore a constant-factor approximation algorithm.*

Proof: Using an argument similar to the one the proof of Theorem 7, the latency of the schedules produced by the HPBCA algorithm (Algorithm 3) can be bounded by $K \cdot \Delta + 7^3 \cdot K \cdot R = O(\Delta + R)$. Thus, it is a constant factor approximation in both the graph model and the SINR model by Lemma 4 and Corollary 6. ■

VI. Conclusion

In this paper, we studied the Minimum Latency Aggregation Scheduling (MLAS) problem adopting the two interference models: the graph model and the more realistic physical interference model known as Signal-to-Interference-Noise-Ratio (SINR). While existing works studied the problem in 2-dimensional (2D) WSNs only, we investigated the problem in the more general 3-dimensional (3D) WSNs, and introduced two approximation algorithms with $O(1)$-approximation ratios that yield schedules whose latency is bounded by $O(\Delta + R)$, where Δ is the maximum node degree and R is the network radius. As to future work, we plan to study the other related problems such as broadcast as well as gossiping in 3D WSNs adopting both interference models.

References

[1] X. Chen, X. Hu, and J. Zhu, "Minimum Data Aggregation Time Problem in Wireless Sensor Networks," in *MSN*, 2005, pp. 133 – 142.

[2] S. C. H. Huang, P.-J. Wan, C. T. Vu, Y. Li, and F. Yao, "Nearly Constant Approximation for Data Aggregation Scheduling," in *INFOCOM*, 2007, pp. 6 – 12.

[3] X. Xu, S. Wang, X. Mao, S. Tang, P. Xu, and X.-Y. Li, "Efficient Data Aggregation in Multi-hop WSNs," in *Globecom*, 2009, pp. 3916 – 3921.

[4] X. Xu, X. Y. Li, X. Mao, S. Tang, and S. Wang, "A Delay-Efficient Algorithm for Data Aggregation in Multihop Wireless Sensor Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 1, pp. 163 – 175, 2011.

[5] P.-J. Wan, S. C.-H. Huang, L. Wang, Z. Wan, and X. Jia, "Minimum-Latency Aggregation Scheduling in Multihop Wireless Networks," in *MobiHoc*, 2009, pp. 185 – 194.

[6] M. K. An, N. X. Lam, D. T. Huynh, and T. N. Nguyen, "Minimum Data Aggregation Schedule in Wireless Sensor Networks," *I. J. Comput. Appl.*, vol. 18, no. 4, pp. 254 – 262, 2011.

[7] N. Lam, M. K. An, D. Huynh, and T. Nguyen, "Scheduling problems in interference-aware wireless sensor networks," in *ICNC*, 2013, pp. 783 – 789.

[8] X.-Y. Li, X. Xu, S. Wang, S. Tang, G. Dai, J. Zhao, and Y. Qi, "Efficient Data Aggregation in Multi-hop Wireless Sensor Networks under Physical Interference Model," in *MASS*, 2009, pp. 353 – 362.

[9] M. K. An, N. X. Lam, D. T. Huynh, and T. N. Nguyen, "Minimum Latency Data Aggregation in the Physical Interference Model," *Computer Communications*, vol. 35, no. 18, pp. 2175 – 2186, 2012.

[10] H. Li, Q. S. Hua, C. Wu, and F. C. M. Lau, "Minimum-latency Aggregation Scheduling in Wireless Sensor Networks under Physical Interference Model," in *MSWiM*, 2010, pp. 360 – 367.

[11] M. M. Halldórsson and P. Mitra, "Wireless Connectivity and Capacity," in *SODA*, 2012, pp. 516 – 526.

[12] R. Gandhi, S. Parthasarathy, and A. Mishra, "Minimizing Broadcast Latency and Redundancy in Ad Hoc Networks," in *MobiHoc*, 2003, pp. 222 – 232.

[13] M. K. An, N. X. Lam, D. T. Huynh, and T. N. Nguyen, "Minimum Latency Gossiping in Wireless Sensor Networks," in *ICWN*, 2012.

[14] R. Tiwari, T. N. Dinh, and M. T. Thai, "On Approximation Algorithms for Interference-Aware Broadcast Scheduling in 2D and 3D Wireless Sensor Networks," in *WASA*, 2009, pp. 438 – 448.

[15] P. Gupta, S. Member, and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Trans. on Information Theory*, vol. 46, pp. 388 – 404, 2000.

[16] [Online]. Available: http://mathworld.wolfram.com/Space-FillingPolyhedron.html

[17] S. C.-H. Huang, P.-J. Wan, J. Deng, and Y. S. Han, "Broadcast Scheduling in Interference Environment," *IEEE Trans. Mob. Comput.*, vol. 7, no. 11, pp. 1338 – 1348, 2008.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. MIT Press and McGraw-Hill, 2009.

A Comparative Analysis of Hands-on Firewall Configuration Exercises for the Undergraduate Classroom

A. Evan Damon[1], **B. Jens Mache**[1], **C. Miles Crabill**[1], **D. Kaleb Ganz**[1], **E. Claire Humbeutel**[1]
[1]Department of Mathematical Sciences, Lewis & Clark College, Portland, OR, USA

Abstract - *Teaching cybersecurity through hands-on, interactive exercises is a good way to engage students, especially undergraduates. In this paper, we compare three firewall configuration exercises: FireSim, DETERLab, and RAVE. We found that they are usable but could be improved upon. They each have strengths and weaknesses, and an exercise combining their strengths would be ideal. There were three main strengths: simplicity, extensibility, and competition. Each of these elements serve to make the exercises engaging and educational.*

Keywords: security, firewalls, education, exercises, networks

1 Introduction

As computer security continues to develop as a field, one thing remains an important element in keeping networks and computers secure: firewalls. Firewalls are often one of the main lines of defense in a system. Despite this, most firewall education is done by employers, as different employers use different firewalls and/or syntax. We believe that teaching a conceptual understanding of firewalls to undergraduate students will help them to gain a more complete knowledge of computer security. While detailed knowledge of every element of security would be ideal, it is less than feasible. However, conceptual knowledge is easily achievable through interactive exercises.

Multiple interactive firewall exercises already exist. In an effort to find one suitable for teaching conceptual knowledge to undergraduates, we used and evaluated each of three exercises. The three exercises were: FireSim,[1] a Java applet-based simulation of a firewall environment that pits players against one another; DETERlab,[2] a remote environment, which involves the creation of firewall rules on real, working machines; and RAVE,[3] a virtual environment complete with lab exercises, including the configuration of a firewall. We chose these exercises because they were readily available, and, as firewall education tools, fit the purpose of our evaluation well.

2 FireSim

One teaching exercise we examined was created by Professor Ken Williams, of North Carolina A&T State University. The firewall simulator (or FireSim for short) was created using Java applets supplemented with XML files. The setup involves downloading a group of necessary files to a networked machine and running the Java .jar file. This machine serves as the host machine. Players then connect to the host machine via web browser and choose a username. Once the administrator/ instructor (on a separate computer) uses the GUI to start the game, players are able to conduct attacks against one another in a competitive setting. Through a GUI, players select an attack and direct it at a particular opponent. If the attack is successful, the attacker receives a point while the defender loses a point. After 60 seconds, the same attack can be launched against the same opponent.

This is where the firewall education takes place; in order to prevent successful attacks from other players, you must write firewall rules. These rules use the proprietary Cisco firewall syntax and are simply typed into a special applet window. The goal is to create a list of firewall rules that prevent other players from successfully attacking you. This can take many forms, but most rules block access to certain ports, depending on IP addresses.

To introduce complexity, FireSim allows the administrator/ instructor to give players new "tasks". Each task may require players to create additional firewall rules that mitigate the new situation or vulnerability. The task is announced to the players, and after 60 seconds, players may begin to attack each other using it.

FireSim does many of these things well. The firewall rules interpret correctly and are as intuitive as the syntax allows. FireSim's network map includes many servers at specific IP addresses and does a good job of simulating a real network. This is one of FireSim's greatest strengths. The underlying network simulation gives the exercise a great deal of potential for further development.

FireSim has weaknesses as well. It is, after all, a work in progress. The tasks are less clear than they could be. Some are "trick" tasks that do not require additional firewall rules. These are intended to show the user that, when creating a firewall, extending your whitelist too far is dangerous. However, this is not obvious to the user. If a player does nothing and is impervious to attack, they may not learn this lesson. On the other hand, players who incorrectly extend the whitelist may have difficulty understanding why they lose points. Essentially, the level of feedback provided by FireSim is too little for some of its exercises to be successful.

This is the only real shortcoming of FireSim: the scenarios get users to write firewall rules in a competitive setting, but they could explain more to promote a real understanding of what is going on under the hood. The scenarios are often about blocking access to a particular service, which is easy enough. However, the zero-sum nature of the point system may promote an offensive strategy, where players spend time to attack other players as much and as soon as possible. For some players, this may take emphasis away from the firewall configuration and places much of it on attacking.

Additionally, the setup was not entirely stable. Sometimes users could not connect to the game, or when connected, the game did not function. The administrator password is also non-functional, allowing players access to administrator privileges which usually breaks the game.

Overall, FireSim does many things well, but some things need improvement. The exercise is very interactive but does not provide much feedback. FireSim is also competitive, and the structure of the competition does well to engage students.

3 DETERlab

DETERlab is the second exercise we examined, and the first to use a remote-access setup. The DETERlab exercise that covers firewalls also covers Posix file permissions. The tasks for the two are independent of each other, but the reading and information for the two are intermixed. This can make it quite challenging to find relevant information for each of the tasks on the walkthrough webpage. This section of the paper will focus exclusively on the firewall aspects of the exercise.

The firewall portion of this exercise starts out with a description of stateless and stateful firewalls, what the difference between them is, and a little bit of history behind the development of the two. Next, there is a description of how a firewall policy should be designed and different ways to view the problem of designing a firewall. After that, the lab gives the user a crash course on iptables. In it, they tell the user what iptables is, briefly describe the syntax and then give a few example rules using the syntax. The lab then goes on to describe four different network tools. These are nmap, ifconfig, telnet and netcat. Each of these has about a sentence to a paragraph long description and an example of what it returns in the terminal when used.

The interactive portion of the lab is done on two remotely accessed nodes. Accessing these nodes requires the user to connect to their DETERlab account through SSH and then from there SSH into the two control nodes. Each of these two nodes serves a different function. One node is a server and is where the user implements their firewall rules. The other is the system on which the user tests the firewall rules they have implemented on their server node.

The firewall rules are easily established since the server node has a firewall script provided, in which the user only needs to write the rules and then run the script to get a firewall up and running. It is also easy for the user to turn off all rules and remove them by running another script that is provided on the server node. The tasks that the lab asks you to do are: to write a rule that prevents spoofing; allow access to OpenSSH, Apache and MySQL on their standard ports; allow UDP access to specified ports; allow ICMP ping requests; allow all established and related traffic; and lastly to drop all other traffic to any unspecified port. Since the user is remotely connected to the the server, it is possible for them to write firewall rules that lock them out of the server node, requiring the user to reboot the instance through the DETERlab website.

At any time while working on the firewall rules, the user can begin to follow the tests that are provided for each of the assigned tasks. These tests mostly involve using telnet to see if the ports are open and responding as they are expected to. However, there is no test provided that allows the user to determine if their anti-spoofing rules are functional. Also, the recommended test for seeing if all the expected ports are open (and all others are closed) is to write a script that tests every port. This could be difficult for some undergraduate students.

The best feature of using this DETERlab exercise is that it mirrors a real life firewall admin situation since it uses a commonly used firewall syntax, the student must remotely connect to the environment to make changes, and they have a broad spectrum of network testing tools at their command. It is also possible for the instructor to create new scenarios that require the students to modify their firewall to fit this new situation. Many of the advantages of using this exercise for learning/ teaching firewall rules have the potential to also be disadvantages, depending on the proficiency of the students and instructor. Iptables has a complicated syntax that can be hard to learn without a certain amount of experience. Also, the possibility that a student may lock themselves out of the environment is both an advantage and a disadvantage in that it teaches the student about the possibility of doing so in a real IT position, but it takes time to reboot the system and can be frustrating for the student. Another disadvantage is that some overhead is required in setting up the experiment and waiting for the systems to become available for the user to connect to them.

4 RAVE

The third exercise, the Rave lab, is a series of instructions and questions in the book *Principles of Computer Security*[4] that (alongside an installed VMware vSphere client) teach students how to configure a firewall in Linux. The client allows the student to have access to different virtual machines that they can configure. Students follow the steps of the lab, enter commands and rules, and gauge how the different firewall configurations affect the system. The goal of this lab is to teach students how to use iptables (using the UFW syntax) and the effects of different rules through hands on work.

The advantages of this approach are mostly found in its simplicity, resistance to mistakes, and opportunity to practice what you are learning about in the related book. The instructions are easy to understand and follow, and progress in a logical manner. The pictures that are provided make it easy to check how your lab is progressing, and also to find any part of the GUI that the lab is requiring you to use.

It is resistant to mistakes because you are using virtual desktops that can be restored from snapshots. That way, if a student makes a relatively damaging mistake and changes any program for the worse, the problem can be corrected relatively simply, even if the whole desktop needs to be restored.

Finally, students can learn through trying an action, not just reading about it. By doing this, they gain a greater understanding of the subject they are currently working on.

The major disadvantage is the necessity of having to have a connection to the web in order to work. This can cause connection problems if a number of people are using the same account at the same time or the servers are under load. Also, if the web connection is not working, it is impossible to get to the virtual desktop, and thus impossible to do the lab exercise. Finally, if updates or any other changes are applied, the usernames and passwords will change as well, effectively making it impossible to do any work until you have gotten new ones.

5 Comparison

The conclusions made in the table below were informed by our experience with each exercise in an undergraduate classroom setting, and the subsequent application of the skills learned when the students participated in the Collegiate Cyber Defense Competition (CCDC).[5] The comparison is based on each tool's usefulness in learning network administration and security.

Comparison of Firewall Exercises

	FireSim	**DETERlab 7**	**Rave/Nestler 7.3**
Syntax	Cisco	iptables	UFW
Scenarios	ftp, dns, http, snmp, ntp, instant messaging, netbios	ssh, http, sql, mail, ping, udp	ssh, http, ftp
Setup	LAN (need at least n+2 computers for n players)	Cloud (access through ssh)	Cloud (access through vSphere)
Complexity for students	Simple	Complex	Simple (step-by-step instructions)
Documentation	Present but not extensive	A long webpage (interleaved with Unix permissions exercise)	Lab book[3]
Extensibility	Possible (XML file supplements)	.ns files to build environments	Maybe possible (virtual machine images)
In-class vs. Homework	LAN setup (most likely requires classroom usage)	Possible as homework, but might require help from instructor	Easy steps make for a good homework assignment
Best feature	Competitive	Realism	Step-by-step
Disadvantage	Lack of feedback, bugs	Realism (places student in large, breakable environment)	Slow (needs bandwidth for remote desktop), limited support for Mac OS X
Syntax	Cisco	iptables	UFW

6 EDURange

The EDURange project was inspired by a pressing need for hands-on security education in the undergraduate classroom.[6,7,8,9,10,11] EDURange seeks to implement the core ideas of the "hacker curriculum" in conjunction with interactive and competitive exercises that challenge undergraduates to both compromise and secure systems.[12] Designed with a focus on active learning and inquiry, EDURange aims to improve on the scenarios and exercises described above by providing an extensible framework for the development of interactive and competitive scenarios. EDURange is currently implemented in Amazon Web Services' Elastic Compute Cloud,[13] and is configured to quickly and painlessly create virtual machine instances that participants can connect to via SSH. EDURange players will be in direct control of a virtual instance, an approach that minimizes abstraction between participant and scenario. EDURange will implement systems that enable the simple configuration and design of scenarios, allowing for customized scenarios. EDURange looks to combine the best of each of the reviewed exercises, allowing for a competitive experience that accurately represents the responsibilities of a network administrator tasked with the configuration of firewalls.

7 Conclusion

Based on undergraduate students trying three different firewall configuration exercises, we are concluding that all three are workable firewall education tools, but they can all be improved upon. FireSim is competitive, interactive, and entertaining, but the educational aspects are dampened by the lack of feedback and guidance. A similarly engaging tool with additional guidance would make an excellent exercise for undergraduate security students. DETERlab was the most realistic of the exercises. The setup was also well thought-out. However, accessing and following the exercise was rather involved, and could be daunting for introductory-level students. Finally, the RAVE labs are efficient and accessible. The exercises are well-written, easy to follow, and educational. Running on virtual machines, students can make any number of catastrophic mistakes without preventing further use, since the system can be restored via snapshots. However, the virtual structure can cause service-related issues, and accessing user credentials is sometimes a problem. A web connection is also required, as well as a significant amount of setup to begin using RAVE.

An ideal exercise would combine the strengths of these three tools. FireSim's competitive aspect engages students on a level that the other two cannot. DETERlab's realism and breadth is also excellent, contributing to its extensibility. RAVE's simple approach and resistance to mistakes make it a good platform for teaching introductory-level students. In constructing a new firewall education tool, educators should seek to include these elements. A competitive, engaging tool with an extensible structure will be an excellent tool for teaching undergraduates about firewalls. If it can be as simple and approachable as the RAVE exercise, as well, the exercise will be the most useful tool of those currently available.

8 Acknowledgements

Partial support for this work was provided by the National Science Foundation's "Transforming Undergraduate Education in Science, Technology, Engineering and Mathematics (TUES)" program under Award No. 1141314, by the John S. Rogers Science Research Program of Lewis & Clark College and by the James F. and Marion L. Miller Foundation. We would also like to thank Richard Weiss and Stefan Boesen.

[1] http://williams.comp.ncat.edu/FireSim/index.htm accessed 5/31/2013
[2] https://education.deterlab.net/mod/resource/view.php?id=367 accessed 5/31/2013
[3] Vincent Nestler, Gregory White, WM. Arthur Conklin, "Principles of Computer Security: CompTIA Security+ and Beyond", Lab Manual. McGraw Hill, 2011.
[4] Arthur Conklin, Gregory White, Vincent Nestler, "Principles of Computer Security", ISBN 0071786198, McGraw-Hill, 2012.
[5] http://nationalccdc.org/ accessed 5/31/2013
[6] Richard Weiss, Jens Mache, Erik Nilsen, "Top 10 Hands-on Cybersecurity Exercises", Journal of Computing Sciences in Colleges, Volume 29, Issue 1, 2013 (to appear)
[7] Richard Weiss, Jens Mache, Vincent Nestler, Ronald Dodge, Brian Hay, "Hands-on Cybersecurity Exercises and the RAVE Virtual Environment", Workshop 9 at the 44rd ACM Technical Symposium on Computer Science Education (SIGCSE), 2013, http://dx.doi.org/10.1145/2445196.2445505
[8] Richard Weiss, Michael E. Locasto, Jens Mache, Blair Taylor, Beth Hawthorne, "Teaching Security Using Hands-On Exercises", Birds of a Feather at the 44rd ACM Technical Symposium on Computer Science Education (SIGCSE), 2013, http://dx.doi.org/10.1145/2445196.2445490
[9] Richard Weiss, Jens Mache, Vincent Nestler, Ronald Dodge, Brian Hay, "Teaching Cybersecurity Through Interactive Exercises Using a Virtual Environment", Journal of Computing Sciences in Colleges, Volume 28, Issue 1, 2012, http://dl.acm.org/citation.cfm?id=2379703.2379735

[10] Richard S. Weiss, Michael E. Locasto, Jens Mache, "Hacking and the Security Curriculum: Building Community", Proceedings of the 43rd ACM Technical Symposium on Computer Science Education (SIGCSE), 2012, http://dx.doi.org/10.1145/2157136.2157432

[11] Richard Weiss, Jens Mache, "Teaching Security Labs with Web Applications, Buffer Overflows and Firewall Configurations" , Journal of Computing Sciences in Colleges , Volume 27 Issue 1, 2011, http://dl.acm.org/citation.cfm?id=2037151.2037185

[12] Bratus, Sergey, Anna Shubina, and Michael E Locasto. "Teaching the principles of the hacker curriculum to undergraduates." *Proceedings of the 41st ACM technical symposium on Computer science education* 10 Mar. 2010: 122-126.

[13] http://aws.amazon.com/ec2/ accessed 5/31/2013

New Results on Asymptotic Analysis of Digital Communications in Generalized Fading Channels

A. Annamalai and E. Adebola

Center of Excellence for Communication Systems Technology Research

Department of Electrical and Computer Engineering
Prairie View A&M University, Texas 77446, U.S.A.

Abstract — **In this article, we develop new closed-form asymptotic approximations for the average symbol error rate (ASER) and the outage probability performance metrics of digital communication systems impaired by additive white Gaussian noise and fading. Specifically, our expressions generalize some of the known asymptotic results to a wide range of fading environments and digital modulation schemes. We also demonstrate that the consideration of only the first term and/or the first two terms of a Maclaurin series expansion of the probability density function (PDF) of the received signal-to-noise ratio (SNR) random variable in conjunction with the Mellin transform of the conditional error probability (CEP) formulas may not be sufficient to yield accurate predictions of the ASER performance and/or outage capacity over a wide range of SNR values.**

Keywords — Asymptotic approximation, Mellin transform, generalized fading channels, diversity methods.

I. INTRODUCTION

Regardless of the branch of science or engineering, theoreticians have always been fascinated with the notion of expressing their results in the form of a closed-form expression, although an exact solution may not always be possible. Even when an exact closed-form solution exists, quite often the elegance of this solution is overshadowed by the complexity of its form (i.e., may not provide direct insight into the key parameters that govern the system performance) and the difficulty in evaluating it numerically. This motivates the search for a solution that is both simple in form and likewise easy to evaluate, and that can be used to develop insights as well as suitable for applications such as cross-layer system design/optimization. A further motivation is that the method used to derive these alternative simple forms should also be sufficiently general for unified analysis over generalized fading channels and/or modulation schemes as well as applicable in situations where the exact closed-form solutions are ordinarily unattainable.

Wireless systems performance measures such as the average bit, symbol, or block error probabilities and the ergodic channel capacities typically involve taking the statistical expectation of the CEP or a logarithmic function with respect to the random variables that characterize the multichannel/multipath fading. For at least six decades, researchers have studied problems of these types (i.e., finding statistical expectations of mathematical functions such as $\ln^p(1+\gamma)$ and $Q^p(\sqrt{\gamma})$ with respect to the SNR random variable γ when the constant p is a positive integer) and wireless system engineers have used both the theoretical and numerical results reported in the literature to guide the design of their systems. However, analytical difficulties associated with computing the statistical expectations of the Gaussian-Q function or its integer powers with respect to their arguments (especially for diversity systems) have led to development of various bounds and approximations (e.g., [1]-[13]). Among these various known solutions, the analytical methods based on either the asymptotic (i.e., large mean SNR) approximation of the probability density function of γ in a variety of fading environments/diversity techniques [1]-[5] or tight exponential type approximations [10]-[11] for the CEPs of different modulation schemes are of significant interest owing to their simplicity, generality and also because they offer insights into how channel and modulation related parameters determine the "diversity gain" and the "coding gain" for various diversity combining/modulation techniques. But the accuracy of the asymptotic analysis approach degrades rapidly the mean SNR decreases, while the gap between the exact and approximate curves for the exponential-type CEP approximation technique widens as the channel experience more severe fading and/or for higher order modulations. Since current wireless systems operate at mean SNR of 0 – 15 dB with average error rates as high as 10^{-2}, a more precise or better closed-form ASER approximations than [1]-[6] are desirable.

More recently, [12] had attempted to tackle the above problem and subsequently proposed highly accurate uniform approximations for the ASER and outage probability over a wider range of mean SNR values, at the expense of requiring additional information (i.e., fractional moments of the fading SNR random variables) and increased computational complexity. It also appears that this method is less versatile compared to [2]-[4] and [10]-[11]. Ref. [13] proposed another class of closed-form ASER/outage probability approximations which require the knowledge of the first two non-zero terms of the Maclaurin series expansion of PDF of γ. Although [13] yields accurate approximations over a wider range of mean SNRs values in Rayleigh and Nakagami-m channels

This work is supported in part by funding from the National Science Foundation (0931679 & 1040207).

TABLE I: MELLIN TRANSFORM FOR THE CEP OF BINARY AND M-ARY DIGITAL MODULATIONS.

Modulation	Conditional Error Probability $g(\gamma)$	Mellin Transform $\psi(t_j+1)=\int_0^\infty g(\gamma)\gamma^{t_j}d\gamma$
	$\rho e^{-k\gamma}$	$\rho(k)^{-(t_j+1)}\Gamma(t_j+1)$
	$\rho Q(\sqrt{k\gamma})$	$\frac{\rho 2^{t_j}\Gamma\left(t_j+1.5\right)}{(k)^{t_j+1}(t_j+1)\sqrt{\pi}}$
BDPSK/BFSK with L^{th} order SLC diversity	$\frac{1}{2^{2L-1}}\exp(-q\gamma)\sum_{i=0}^{L-1}C_i(q\gamma)^i$	$\frac{(q)^{-(t_j+1)}}{2^{2L-1}}\sum_{i=1}^{L-1}C_i\Gamma(i+t_j+1)$
Orthogonal M-FSK	$\sum_{n=1}^{M-1}(-1)^{n+1}\binom{M-1}{n}\frac{1}{n+1}\exp\left(\frac{-n\gamma}{n+1}\right)$	$\Gamma(t_j+1)\sum_{n=1}^{M-1}(-1)^{n+1}\binom{M-1}{n}\frac{1}{n+1}\left(\frac{n}{n+1}\right)^{-(t_j+1)}$
Coherent M-PSK (exact)	$\frac{1}{\pi}\int_0^{(M-1)\pi/M}\exp\left(\frac{-\gamma\sin^2(\pi/M)}{\sin^2\theta}\right)d\theta$	$\frac{\Gamma(t_j+1)B_{(\sin[(M-1)\pi/M])^2}(t_j+1.5,0.5)}{2\pi[\sin(\pi/M)]^{2(t_j+1)}}$
Coherent M-QAM (exact)	$\frac{4\left(1-\frac{1}{\sqrt{M}}\right)}{\pi}\int_0^{\pi/2}\exp\left(\frac{3\gamma}{2(M-1)\sin^2\theta}\right)d\theta$ $-\frac{4\left(1-\frac{1}{\sqrt{M}}\right)^2}{\pi}\int_0^{\pi/4}\exp\left(\frac{3\gamma}{2(M-1)\sin^2\theta}\right)d\theta$	$\frac{2}{\pi}\Gamma(t_j+1)[2(M-1)/3]^{t_j+1}\left(1-\frac{1}{\sqrt{M}}\right)$ $\times\left[B_1(t_j+1.5,0.5)-\left(1-\frac{1}{\sqrt{M}}\right)B_{0.5}(t_j+1.5,0.5)\right]$

Note: $\Gamma(.)$ and $B_x(.,.)$ denote the Gamma function and the incomplete Beta function, respectively. Also, the coefficients $C_i=\frac{1}{i!}\sum_{l=0}^{L-1-i}\binom{2L-1}{l}$ and q = 1 for BDPSK while q = 0.5 for BFSK.

compared to the large SNR approximation method [1]-[5], the determination of the required coefficients can be rather cumbersome from the moment generating function (MGF) of γ in certain fading environments (e.g., Rice and κ-μ fading) and for diversity systems under the realistic assumptions of independent but non-identically distributed (i.n.d) fading statistics and/or correlated diversity paths. Our article aims to address this concern.

The contributions of this article are three-fold: (i) First, we investigate the efficacy of using higher order Maclaurin series expansion for the PDF of γ in Nakagami-m and Rice fading channels in conjunction with Mellin transforms of various CEPs of binary and M-ary modulation schemes, thereby generalizing the results in [2] on two fronts; (ii) Secondly, we propose to greatly simplify the task of finding the required coefficients for the new class of outage/ASER approximations proposed in [13] for maximal-ratio combining (MRC) or square-law combining (SLC) diversity receivers by using the simple formulas derived by Welch-Satterthwaite [16]-[17] and Moschopoulos [18]-[19] for approximating the sum of independent and correlated Gamma random variates by another Gamma random variable; (iii) Lastly, we derive new closed-form approximations for the ergodic channel capacity in generalized fading environments.

II. MACLAURIN SERIES APPROXIMATION OF THE PDF OF SNR WITH MELLIN TRANSFORM OF THE CEP

The Maclaurin series (i.e., power series) expansion of the PDF of SNR γ at origin is given by

$$f_\gamma(\gamma)=\sum_{j=0}^{N}a_j\gamma^{t_j}+R_N(\gamma), \quad (1)$$

where the coefficients $a_j=\frac{1}{j!}\frac{d^j}{dx^j}f_\gamma(x)\Big|_{x=0}$, $t_j=t+j$ and the remainder term $R_N(x)=\frac{x^{N+1}}{(N+1)!}\frac{d^{(N+1)}}{dx^{(N+1)}}f_\gamma(x)$. Thus it is not very difficult to determine the coefficients a_j and t for the single channel reception (no diversity) case. For example, the corresponding coefficients in Nakagami-m and Rice fading are summarized in (2) and (3) respectively, viz.,

$$\text{Nakagami-m: } a_j=\frac{(-1)^j m^{m+j}}{j!\Omega^{m+j}\Gamma(m)},\ t=m-1 \quad (2)$$

$$\text{Rice: } a_j=\frac{(-1)^j(1+K)^{j+1}e^{-K}}{\Omega^{j+1}j!},\ t=0 \quad (3)$$

where m denotes the Nakagami-m fading severity index, K corresponds to the Rice factor and Ω is the mean SNR.

Since the accuracy at high mean SNR regime is dominated by the behaviour of the PDF of fading SNR as $\gamma\to 0^+$ (or the behaviour of MGF of SNR $\phi_\gamma(s)$ as $s\to\infty$), [2] considered only the first-term approximation of (1) (i.e., N = 0) for deriving simple closed-form approximations for the ASER and outage probability metrics. Although it is anticipated that a larger choice of N in (1) may lead to an improved approximation over a wider range of mean SNR values, this has not been adequately studied or reported in the literature.

For instance, how large should N be to yield accurate ASER and/or outage probability predictions at the low mean SNR regime? Furthermore, [13] indirectly argues that two-terms approximation of (1) might be adequate to achieve highly accurate ASER predictions over a wide range of mean SNR values (since their modified asymptotic PDF of SNR matches the two-terms Maclaurin series expansion of (1) exactly as $\gamma \to 0^+$). Therefore in the following we will investigate and provide an affirmative answer to this question.

To develop an improved approximation for the asymptotic ASER, one might want to integrate the CEP (i.e., error probability of a specified digital modulation over an AWGN channel) over the PDF of SNR depicted in (1), yielding

$$\overline{P_s} = \int_0^\infty g(\gamma) f_\gamma(\gamma) d\gamma \doteq \sum_{j=0}^{N} a_j \int_0^\infty g(\gamma) \gamma^{t_j} d\gamma = \sum_{j=0}^{N} a_j \psi(t_j + 1), \quad (4)$$

where $\psi(t_j) = \int_0^\infty g(\gamma) \gamma^{t_j - 1} d\gamma$ denotes the Mellin transform of the CEP, $t_j = t + j$ while the coefficients a_j and t are real constants that depend on the fading distribution. It should be evident by now that the improved approximation for the asymptotic ASER is simply a weighted sum of the Mellin transform of CEP, and this quantity for a broad class of coherent, differentially coherent and noncoherent modulation schemes are summarized in Table I. The above unified expression also generalizes the result in [2] in two-ways (i.e., consideration of (N + 1)-terms Maclaurin series expansion of $f_\gamma(\gamma)$ and extension to other modulation schemes besides coherent BPSK). In Table II, we also summarize the parameters ρ and k where the exact CEPs of various digital modulations are further approximated as $P_s(\gamma) = \rho Q(\sqrt{k\gamma})$ to yield a more compact ASER approximation.

TABLE II: PARAMETERS ρ AND k FOR SOME DIGITAL MODULATIONS

$P_s(\gamma) = \rho Q(\sqrt{k\gamma})$	ρ	k
Coherent BPSK	1	2
Coherent BFSK	1	1
M-DEPSK ($M \geq 2$)	4	$2\sin^2(\pi / M)$
M-PSK ($M \geq 4$)	2	$2\sin^2(\pi / M)$
M-FSK ($M > 2$)	$M - 1$	1
Square M-QAM ($M \geq 4$)	$\frac{4(\sqrt{M} - 1)}{\sqrt{M}}$	$3/(M - 1)$
M-DPSK ($M \geq 2$)	2	$4\sin^2(\pi / (2M))$

Example 1: M-ary Phase Shift Keying

We consider the coherently detected M-PSK whose CEP can be tightly approximated as [10]

$$P_s(\gamma) = \rho Q(\sqrt{k\gamma}), \quad (5)$$

where $k = 2\sin^2(\pi / M)$, $\rho = 2$ if $M \geq 4$ and $\rho = 1$ if $M = 2$. Now utilizing the appropriate Mellin transform given in Table I, it is straight-forward to re-write (4) as

$$\overline{P_s} \doteq \rho \sum_{j=0}^{N} a_j \frac{2^{t_j} \Gamma(t_j + 1.5)}{(k)^{t_j + 1} (t_j + 1)\sqrt{\pi}}, \quad (6)$$

where the notation $\doteq$ denotes the asymptotic approximation. The asymptotic ASER approximations for M-PSK in Nakagami-m and Rice fading are obtained by substituting (2) and (3) into (6), respectively. Eq. (6) generalizes the results in [2] to higher order constellations. It is also important to highlight that the use of the Mellin transform entry of row 6 in Table I (instead of row 3) does not yield significantly better ASER approximation, and thus (6) is preferred in this case since the final expression is much more compact.

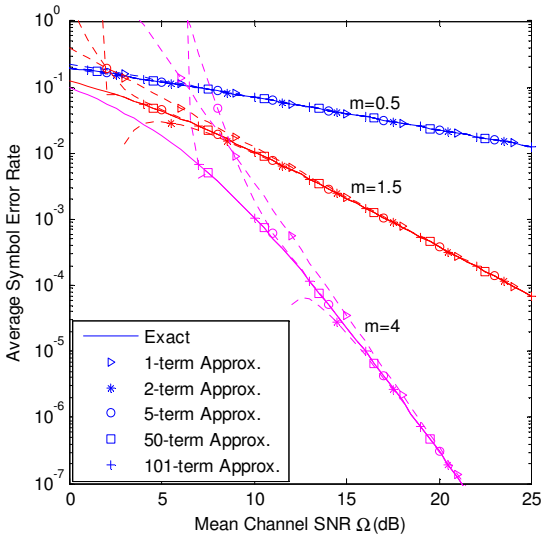


Fig. 1: ASER of BPSK over Nakagami-m channels with $m = 0.5$, 1.5 and 4.

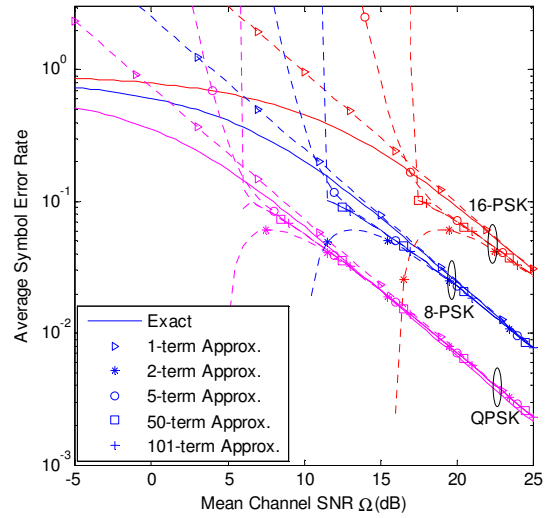


Fig. 2: ASER of M-PSK in Rice fading with Rice factor $K = 1$.

Fig. 1 investigates the efficacy of our improved asymptotic approximations for the ASER of BPSK (see (6)) in different Nakagami-m fading environments. It is important to note that the curve corresponding to the "exact" analysis was generated using [14, Eqs. (5.3) and (2.22)], viz.,

$$\overline{P_s} = \frac{1}{\pi} \int_0^{\frac{\pi}{2}} \left(1 + \frac{\Omega}{m \sin^2 \theta}\right)^{-m} d\theta. \quad (7)$$

It is evident that the single-term approximation is quite accurate over a wide range of mean SNRs for small values of fading severity index (m = 0.5) but its accuracy deteriorates rapidly as m increases. It is also interesting to note that choosing N as high as 101 terms still does not guarantee very good accuracy in the low mean SNR regime especially when the channel experiences less severe fading (i.e., Eq. (6) fails to yield accurate ASER predictions in this scenario). In such cases, an extremely large $N \to \infty$ will be required.

Fig. 2 depicts the performance of our asymptotic ASER approximations for different M-PSK signal constellations in a Rice fading environment. The exact performance curve was generated with the aid of [14, Eq. (5.78)], viz.,

$$\overline{P_s} = \frac{1}{\pi}\int_0^{(M-1)\pi/M} \phi_\gamma\left(\frac{\sin^2(\pi/M)}{\sin^2\theta}\right)d\theta, \quad (8)$$

along with the MGF of SNR [14, Eq. (5.11)]

$$\phi_\gamma(s) = \frac{1+K}{1+K+s\Omega}\exp\left(\frac{-Ks\Omega}{1+K+s\Omega}\right). \quad (9)$$

We observe that the asymptotic ASER approximations become less accurate as the size of the signal constellation increases. From our numerical results, we can safely conclude that alternative solutions are needed for higher order modulations especially when the channel experience less severe fading. Before concluding this section, we would also like to point out that (1) cannot be used for deriving the asymptotic approximations for the ergodic channel capacity since the Mellin transform for the logarithmic function does not exist (i.e., diverging series).

III. Modified Asymptotic PDF of SNR

In [13], the authors proposed a new approximation for asymptotic PDF of SNR in the form of

$$f_\gamma(\gamma) \doteq a\gamma^t e^{-b\gamma}, \quad (10)$$

by noting that $a_0\gamma^t + a_1\gamma^{t+1} = a_0\gamma^t[1+(a_1/a_0)\gamma] \approx a_0\gamma^t e^{\gamma(a_1/a_0)}$ as $\gamma \to 0^+$. Therefore, $a = a_0$ and $b = -a_1/a_0$.

In contrast, here we propose a much more direct approach for determining the optimum values for the coefficients a, t and b by comparing (10) to the exact PDF of SNR in different fading environments as $\gamma \to 0^+$. Some of these results are discussed in Examples 2 and 3 below.

A. Single Channel Reception (No Diversity)

Example 2: Nakagami-m Fading Channel

The PDF of SNR in a Nakagami-m channel is given by

$$f_\gamma(\gamma) = \frac{m^m}{\Omega^m\Gamma(m)}\gamma^{m-1}e^{-\frac{m\gamma}{\Omega}}, \ \gamma \geq 0. \quad (11)$$

Hence by comparing (10) and (11), we immediately obtain $a = \frac{m^m}{\Omega^m\Gamma(m)}$, $t = m-1$ and $b = m/\Omega$. It is also important to note that in this case, (10) is a proper PDF (i.e., Gamma density) and as a consequence, one can quickly invoke the well-known results for approximating the sum of independent and/or correlated Gamma variates with another Gamma random variable. This property is very useful for deriving tight approximations for the ASER and outage probability of coherent modulations with MRC diversity or noncoherent modulations with SLC diversity receivers. Additional details can be found in Section III.B.

Example 3: κ-μ and Rice Fading Channels

The PDF of SNR in a κ-μ fading channel is given by

$$f_\gamma(\gamma) = \left(\frac{1+\kappa}{\Omega}\right)^{\frac{\mu+1}{2}}\left(\frac{\gamma}{\kappa}\right)^{\frac{\mu-1}{2}}\mu e^{-\mu\kappa}e^{-\frac{\mu(1+\kappa)\gamma}{\Omega}}I_{\mu-1}\left(2\mu\sqrt{\frac{\kappa(1+\kappa)\gamma}{\Omega}}\right) \quad (12)$$

The above PDF can be expressed in the form of (10) by invoking the small argument approximation for the modified Bessel function $I_v(x) \approx (x/2)^v/\Gamma(v+1)$. Hence we obtain the following coefficients: $a = \left[\frac{\mu(1+\kappa)}{\Omega}\right]^\mu \frac{e^{-\mu\kappa}}{\Gamma(\mu)}$, $b = \frac{\mu(1+\kappa)}{\Omega}$ and $t = \mu - 1$. The Rice fading can be treated as a special case by substituting $\mu = 1$ and $\kappa = K$. Also letting $\mu = m$ and $\kappa = 0$, we get exactly the same coefficients as in Example 2, as anticipated.

Besides the new developments discussed above, we also feel that the explanation offered in [13] for the intuition behind the asymptotic approximation (10) appears to be incorrect (based on our numerical results in Section II). In fact, we believe that the closed-form ASER approximations derived using (10) offer significant improvement in the mean SNR range and accuracy because the exponential term $e^{-b\gamma} = \sum_{k=0}^{\infty}\frac{(-b\gamma)^k}{k!}$ allows one to conveniently capture and closely approximate a very large number of terms in the Maclaurin series expansion (1) (instead of tightly approximating its first two-terms only).

Example 4: Tight Approximations for ASER and Outage Probability Performance Metrics

In this example, we will consider a generic CEP expression as depicted in Table II for several different modulation/detection schemes. Hence a tight approximation for the desired ASER can be derived by finding the statistical expectation of (5) over the asymptotic PDF (10), viz.,

$$\overline{P_s} \doteq \rho a\int_0^\infty Q(\sqrt{k\gamma})\gamma^t e^{-b\gamma}d\gamma. \quad (13)$$

Eq. (13) can be computed in closed-form with the help of identity [15, Eq. (6.286.1)], viz.,

$$\overline{P_s} = \frac{\rho 2^t a\sqrt{k}\Gamma(t+1.5)}{\sqrt{\pi}(t+1)(2b+k)^{t+1.5}}\,{}_2F_1\left(1, t+1.5, t+2, \frac{2b}{2b+k}\right), \quad (14)$$

where ${}_2F_1(.,.;.;.)$ denotes the Gauss hypergeometric function. It is also interesting to note that the large-SNR approximation [2] can be readily deduced from (14) by setting b = 0 and recognizing that ${}_2F_1(.,.;.;0) = 1$. Also the outage probability can be calculated in closed-form as

$$P_{out} \doteq \int_0^{\gamma_T} ax^t e^{-bx}dx = \frac{a\gamma(t+1, b\gamma_T)}{b^{t+1}}, \quad (15)$$

with the aid of [15, Eq. (3.381.1)].

B. Maximal-Ratio and Square-Law Diversity Receivers

The outage probability and ASER analyses of diversity systems are generally more involved and/or cumbersome since the PDF of the combiner output SNR are typically not available in closed-form. In this section, we will investigate the efficacies of several distinct methods for approximating the coefficients a, b and t in (10) for MRC and/or SLC diversity receivers under different assumptions of fading statistics including the i.i.d, i.n.d and correlated diversity paths. These coefficients can be directly applied in (14) and (15) to obtain their corresponding ASER and outage probability performance predictions.

B.1 Welch-Satterthwaite Approximation

Suppose $\gamma_i \sim G(\alpha_i, \beta_i)$ follows the Gamma distribution with parameters α_i and β_i. Then its PDF is given by

$$f_{\gamma_i}(\gamma) = \frac{\gamma^{\alpha_i - 1} e^{-\gamma/\beta_i}}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}, \ \gamma \geq 0. \tag{16}$$

Using the Welch-Satterthwaite approximation [16]-[17], we can also approximate the sum of L i.n.d Gamma random variables $\gamma_C = \sum_{i=1}^{L} \gamma_i \sim G(\hat{\alpha}, \hat{\beta})$ by another Gamma random variable with parameters shown in (17), viz.,

$$\hat{\alpha} = \frac{\left(\sum_{i=1}^{L} \alpha_i \beta_i\right)^2}{\sum_{i=1}^{L} \beta_i^2 \alpha_i} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{L} \beta_i^2 \alpha_i}{\sum_{i=1}^{L} \alpha_i \beta_i}. \tag{17}$$

Hence the coefficients for asymptotic PDF of combiner output SNR can be deduced by comparing (10) and (16), viz.,

$$a = \hat{\beta}^{-\hat{\alpha}} / \Gamma(\hat{\alpha}), \ t = \hat{\alpha} - 1, \ b = 1/\hat{\beta}. \tag{18}$$

For the simple case of i.i.d fading SNRs in a Nakagami-m fading environment, (18) simplifies into

$$a = \frac{(m/\Omega)^{Lm}}{\Gamma(Lm)}, \ t = Lm - 1, \ b = m/\Omega. \tag{19}$$

It is also possible to extend the above analysis to arbitrarily correlated Gamma random variables. In this case, we set $\alpha_i = \alpha$ and replace β_i with the eigen values of a matrix which is product of a $L \times L$ diagonal matrix with entries β_i and a positive definite $L \times L$ correlation matrix.

B.2 Moschopoulos Approximation

Our second method is based on the first-term approximation of the infinite series formula for the PDF of a sum of L i.n.d Gamma random variates derived in [18], viz.,

$$f_{\gamma_C}(\gamma) \doteq \frac{\left[\prod_{i=1}^{L}\left(\frac{\beta_{(1)}}{\beta_i}\right)^{\alpha_i}\right] \gamma^{\left(\sum_{i=1}^{L}\alpha_i\right)-1} e^{-\gamma/\beta_{(1)}}}{\left(\beta_{(1)}\right)^{\sum_{i=1}^{L}\alpha_i} \Gamma\left(\sum_{i=1}^{L}\alpha_i\right)}, \tag{20}$$

where $\beta_{(1)} = \min(\beta_1, \beta_2, ..., \beta_L)$. In this case, we get

$$a = \frac{\prod_{i=1}^{L}\left(\frac{\beta_{(1)}}{\beta_i}\right)^{\alpha_i}}{\left(\beta_{(1)}\right)^{\sum_{i=1}^{L}\alpha_i} \Gamma\left(\sum_{i=1}^{L}\alpha_i\right)}, \ t = \left(\sum_{i=1}^{L}\alpha_i\right) - 1, \ b = \frac{1}{\beta_{(1)}}. \tag{21}$$

It is interesting to note that for the special case of i.i.d fading SNR statistics in Nakagami-m fading, (21) simplifies into (19). However, the corresponding coefficients obtained from these two approaches will be slightly different for the i.n.d fading statistics and correlated diversity cases.

B.3 Asymptotic MGFApproximation

Suppose that the asymptotic MGF of SNR for each diversity path can be expressed in the following form:

$$\phi_{\gamma_i}(s)\big|_{s\to\infty} = \frac{c_i}{s^{\mu_i}} + \frac{d_i}{s^{\mu_i+1}} + O\left(s^{-\mu_i-2}\right), \ i = 1, ..., L. \tag{22}$$

Then the asymptotic MGF of $\gamma_C = \sum_{i=1}^{L} \gamma_i$ is given by [13]

$$\phi_{\gamma_C}(s)\big|_{s\to\infty} = \frac{\prod_i c_i}{s^{\sum \mu_i}} + \frac{\sum_j d_j \prod_{i\neq j} c_i}{s^{1+\sum \mu_i}} + O\left(s^{-2-\sum \mu_i}\right), \tag{23}$$

while the coefficients for the asymptotic PDF of combiner output SNR (10) are summarized below, viz.,

$$a = \frac{\prod_i c_i}{\Gamma(\sum_i \mu_i)}, \ t = -1 + \sum_i \mu_i, \ b = \frac{-\sum_i d_i / c_i}{\sum_i \mu_i}. \tag{24}$$

On the other hand, if we have the knowledge of the coefficients (a_i, t_i, b_i) for each diversity path (e.g., see Examples 2 and 3), then the corresponding coefficients for the asymptotic combiner output SNR can be readily determined with the aid of (25), viz.,

$$a = \frac{\prod_i [a_i \Gamma(t_i + 1)]}{\Gamma\left(\sum_i (t_i + 1)\right)}, \ t = -1 + \sum_i (t_i + 1), \ b = \frac{\sum_i b_i (t_i + 1)}{\sum_i (t_i + 1)}. \tag{25}$$

Once again for the special case of i.i.d fading SNRs in a Nakagami-m channel, (25) reduces into (19).

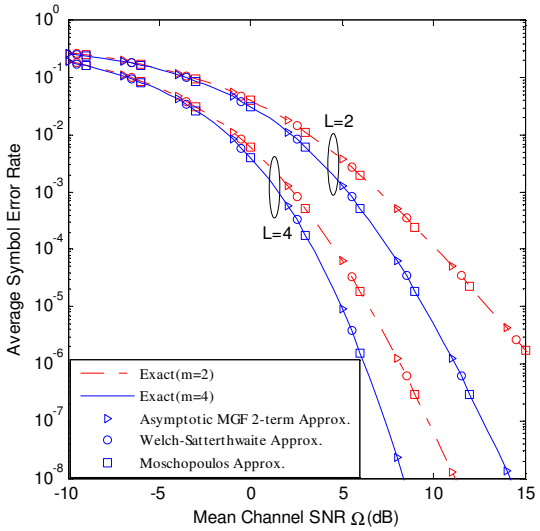


Fig. 3: ASER of coherent BPSK with L-branch MRC diversity receiver over i.i.d Nakagami fading channels (m = 2, 4).

In Fig. 3, we investigate the accuracy of various asymptotic ASER approximations for coherent BPSK equipped with MRC diversity over i.i.d Nakagami-m channels. The exact performance curve was generated using (26), viz.,

$$\bar{P}_s = \frac{1}{\pi}\int_0^{\frac{\pi}{2}}\prod_{i=1}^{L}\left(1+\frac{\Omega_i}{m_i \sin^2\theta}\right)^{-m_i} d\theta \,. \qquad (26)$$

It is evident that the asymptotic ASER approximation curves generated with the aid of (14) are extremely accurate even at very low mean SNR values. Moreover, we also observe that all of the three different asymptotic approximations yield identical results for the special case of i.i.d fading SNR statistics in Nakagami-m channels (i.e., since the values of the coefficients a, b and t are exactly the same for all these approximations). Nonetheless, we expect to see some variations in their accuracies when i.n.d fading statistics or correlated diversity paths are considered.

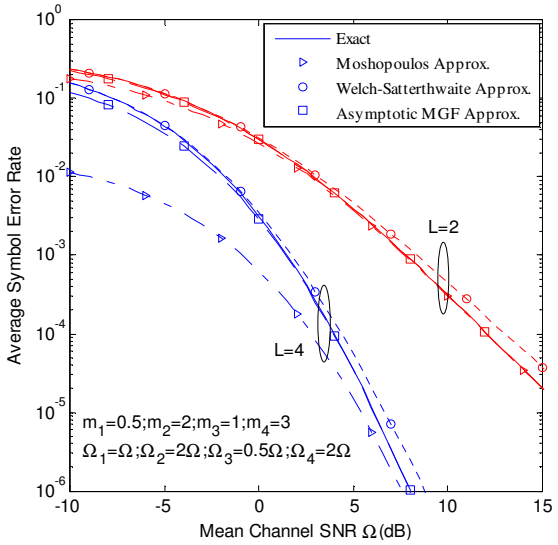


Fig. 4: ASER of coherent BPSK equipped with a MRC diversity receiver (L = 2 or 4) in an i.n.d Nakagami-m fading environment.

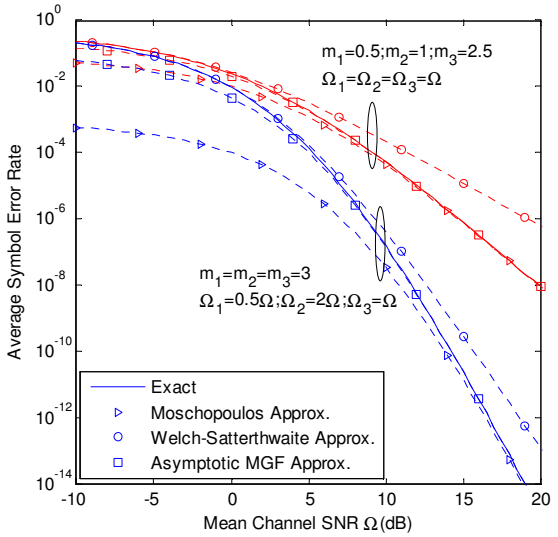


Fig. 5: ASER of coherent BPSK with MRC diversity (L = 3) over i.n.d Nakagami-m fading with non-identical fading severity parameters and/or unequal mean signal strengths.

In Fig. 4 and Fig. 5, we investigate the accuracies of asymptotic MGF approximation (25), Welch-Satterthwaite approximation (17)-(18) and the first-term approximation of Moschopoulos's infinite series formula (20)-(21) for approximating a sum of i.n.d Gamma random variables by another Gamma random variable. In this case, we found that the Moschopoulos's approximation is much tighter than the Welch-Satterthwaite approximation at high mean channel SNRs but it fails to yield reasonable estimates of the exact ASER at low mean channel SNRs. This observation becomes even more pronounced with the increasing diversity order L. On the other hand, the gap between the exact ASER and Welch-Satterthwaite approximation curves diminishes at very low mean SNR values (< 0 dB). Moreover, the asymptotic MGF approximation (25) is considerably tighter than the Moschopoulos approximation over a wide range of mean SNR values. Therefore, one might want to choose an appropriate approximation method (i.e., between (21) and (25)) depending on the SNR region of interest.

Fig. 6 shows a comparison between the various asymptotic ASER approximations for coherent M-PSK with a MRC diversity receiver in i.i.d Rice fading environment (K = 5). The exact ASER curve was generated by numerically evaluating (8) with the MGF of $\gamma_C = \sum_i \gamma_i$ given by $\phi_{\gamma_C}(s) = [\phi_\gamma(s)]^L$ and (9). In computing the corresponding asymptotic ASER based on the Welch-Satterthwaite and Moschopolous approximations, we first approximate the Rice factor K to Nakagami-m fading severity index m using the relationship [14, Eq. (2.26)] viz.,

$$m \simeq \frac{(1+K)^2}{1+2K},\ K \geq 0 \,. \qquad (27)$$

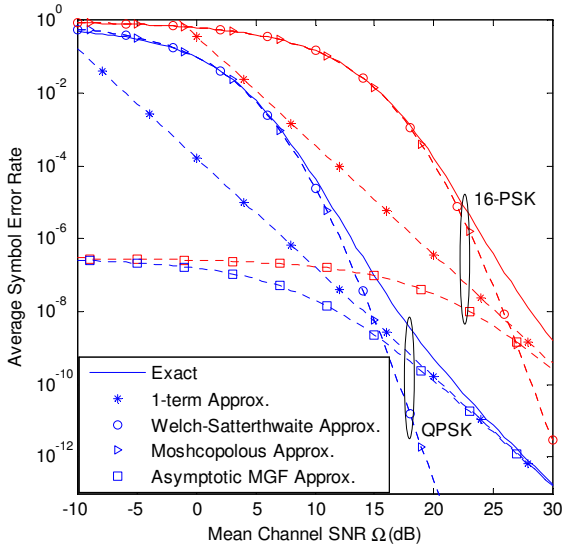


Fig. 6: ASER of coherent M-PSK (M = 4 or 16) with MRC diversity receiver (L = 3) over i.i.d Rice fading (Rice factor K = 5).

For the 1-term asymptotic approximation, we have utilized the asymptotic PDF of MRC/SLC diversity [5], viz.,

$$f_{\gamma_C}(\gamma) \doteq \left[\prod_{i=1}^{L} a_i g_i^{-(t_i+1)}\Gamma(t_i+1)\right]\frac{\gamma^{-1+\sum_{i=1}^{L}(t_i+1)}}{\Gamma\left(\sum_{i=1}^{L}(t_i+1)\right)}, \qquad (28)$$

where $g_i = \Omega_i / \bar{\Omega}$, $\bar{\Omega} = \frac{1}{L}\sum_{i=1}^{L}\Omega_i$ and Ω_i denotes the mean SNR of the i^{th} diversity path. It is apparent from Fig. 6 that the curves corresponding to the Welch-Satterthwaite and Moschopolous approximations overlap (i.e., this is not very surprising since their coefficients *a*, *b* and *t* will be identical for specific case of i.i.d fading statistics) but more importantly, they tightly approximate the exact ASER performance at low mean SNRs. In contrast, the one-term asymptotic approximation converges to the exact ASER performance only at high mean SNR values. Hence these two complementary asymptotic approximations could be exploited to obtain better predictions of the ASER over a wider SNR range. It is also important to highlight that the asymptotic MGF approximation (25) did not produce satisfactory ASER estimates in the Rice fading case, which is in stark contrast to its performance trend in Nakagami-m fading.

C. Ergodic Channel Capacity

The average channel capacity is given by

$$\bar{C}/B = \frac{1}{\ln 2}\int_0^\infty \ln(1+\gamma) f_\gamma(\gamma) d\gamma = \frac{1}{\ln 2}\int_0^\infty \frac{1-F_\gamma(\gamma)}{1+\gamma} d\gamma . \quad (29)$$

The asymptotic PDF in the form $f_\gamma(\gamma) \doteq a\gamma^t$ cannot be used to develop approximations for the ergodic capacity since the resulting integral does not converge. Perhaps for this reason, there is lack of significant contributions on the asymptotic ergodic capacity analysis in the literature. However, we will utilize the modified asymptotic PDF (10) to derive a simple closed-form expression for this metric. In this case, we need to evaluate the integral (30) in closed-form, viz.,

$$\bar{C}/B \doteq \frac{a}{\ln 2}\int_0^\infty \ln(1+\gamma)\gamma^t e^{-b\gamma} d\gamma = \frac{a}{\ln 2} I_{t+1}(b) . \quad (30)$$

For non-negative integer *t* (*t* = 0, 1, 2, …), the above integral can be evaluated in closed-form as

$$I_n(b) = \frac{1}{b^n}\int_0^\infty \frac{\Gamma(n, b\gamma)}{1+\gamma} d\gamma = \Gamma(n) e^b \sum_{k=0}^{n-1} \frac{\Gamma(-k, b)}{b^{n-k}} . \quad (31)$$

The case of real $t \geq 0$ can also be treated without much difficulty. In this case, we obtain a closed-form solution in terms of MacRobert's E-function, viz.,

$$I_t(b) = \int_0^\infty {}_2F_1(1,1;2;-\gamma)\gamma^t e^{-b\gamma} d\gamma = b^{-(t+1)} E(1,1,t+1:2:b) , \quad (32)$$

with the help of [15, eqs. (9.121.6) and (7.522.1)]. To the best of our knowledge, the above expression is new. It is also possible to express the MacRobert's *E*-function in terms of the more familiar Meijer *G*-function as

$$E(\alpha_1,\ldots,\alpha_p : \beta_1,\ldots,\beta_q : x) = G_{q+1,p}^{p,1}\left(x \middle| \begin{matrix} 1, \beta_1,\ldots,\beta_q \\ \alpha_1,\ldots,\alpha_p \end{matrix} \right) . \quad (33)$$

IV. Concluding Remarks

In this article, we have developed several new asymptotic closed-form approximations for the ergodic channel capacity, outage probability and the ASER of broad class of digital modulations in a wide range of fading environments. We have also shown that only a marginal improvement in accuracy can be attained by considering up to 100 terms in the Maclaurin series expansion of the PDF, especially in channels that experience less severe fading and/or for larger constellation sizes. We have also investigated the efficacies of several different asymptotic approximations for the PDF of a sum of Gamma random variables, which finds application in the analysis of MRC and/or SLC diversity systems.

References

[1] H. Abdel-Ghaffar and S. Pasupathy, "Asymptotical Performance of M-ary and Binary Signals over Multipath/Multichannel Rayleigh and Rician Fading," *IEEE Trans. Commun*., vol. 43, pp. 2721–2731, 1995.

[2] Z. Wang and G. Giannakis, "A Simple and General Parameterization Quantifying Performance in Fading Channels," *IEEE Trans. Commun*., vol. 51, no. 8, pp. 1389–1398, 2003.

[3] Y. Ma, Z. Wang and S. Pasupathy, "Asymptotic Performance of Hybrid-Selection/Maximum-Ratio Combining Over Fading Channels," *IEEE Trans. Commun.,* vol. 54, no.5, pp. 770-777, 2006.

[4] A. Nasri, R. Schober, and Y. Ma, "Unified Asymptotic Analysis of Linearly Modulated Signals in Fading, Non-Gaussian Noise, and Interference," *IEEE Trans. Commun*., vol. 56, pp. 980–990, 2008.

[5] A. Annamalai and M. Buehrer, Tutorial Notes on 'Space-Time Processing' presented at the MPRG Annual Symposium/Wireless Summer School, June 2005.

[6] A. J. Goldsmith and S.-G. Chua, "Variable-Rate, Variable-Power MQAM for Fading Channels," *IEEE Trans. Commun*., vol. 45, no. 10, pp. 1218-1230, Oct. 1997.

[7] A. Annamalai, E. Adebola and O. Olabiyi, "Further Results on the Dirac Delta Approximation and the Moment Generating Function Techniques for Error Probability Analysis in Fading Channels," *International Journal of Computer Networks & Communications*., vol.5, no.1, pp. 21-39, Jan. 2013.

[8] E. Adebola, O. Olabiyi and A. Annamalai, "On the Dirac Delta Approximation and the MGF Method for ASER Analyses of Digital Communications over Fading Channels," *IEEE Commun. Lett.*, vol. 17, no. 2, pp. 245-248, Feb. 2013.

[9] O. Olabiyi and A. Annamalai, "Unified Analysis of Two-Hop Cooperative Amplify-and-Forward Multi-Relay Networks," *International Journal of Computer Networks & Communications*, vol.4, no. 6, pp. 1-20, Nov. 2012.

[10] O. Olabiyi, and A. Annamalai, "Invertible Exponential-Type Approximations for the Gaussian Probability Integral $Q(x)$ with Applications," *IEEE Wireless Communication Letters*, vol. 1, pp. 544-547, Oct. 2012.

[11] O. Olabiyi and A. Annamalai, "New Exponential-Type Approximations for erfc(.) and erfc^p(.) Functions with Applications," *Proc. 8th IEEE International Conference on Wireless Communication and Mobile Computing*, Cyprus, Aug. 2012, pp. 1221-1226.

[12] C. Tellambura, Y. Dhungana and M. Soysa, "Uniform Approximations for Wireless Performance in Fading, Noise and Interference,"*Proc. IEEE ICC'12: Communications Theory Symposium*, pp. 2410-2415.

[13] Y. Dhungana and C. Tellambura, "New Simple Approximations for Error Probability and Outage in Fading," *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1760-1763, Nov. 2012.

[14] M. K. Simon and M.S. Alouini, *Digital Communication over Fading Channels*, New York: Wiley, 2nd Edition, 2005.

[15] I. Gradsteyn and I. Ryzhik, *Table of Integrals, Series and Products,* 7th edition, Academic Press, 2007.

[16] F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, vol. 2, 1946, pp. 110–114.

[17] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, vol. 34, 1947.

[18] P.G. Moschopoulos, "The Distribution of the Sum of Independent Gamma Random Variables, *Annals of the Institute of Statistical Mathematics*, vol. 37, Part A, pp. 541–544, 1985.

[19] M.S. Alouini, A. Abdi and M. Kaveh, "Sum of Gamma Variates and Performance of Wireless Communication Systems over Nakagami-m Fading Channels," *IEEE Trans. Vehicular Tech.*., vol. 50, no.6, pp. 1471-1480, 2001.

SESSION

SENSOR NETWORKS, SECURITY ISSUES, MOBILE SYSTEMS, MIMO, WIMAX, GPS, STANDARDS, AND OTHER RELATED ISSUES

Chair(s)

Prof. Hamid Arabnia
University of Georgia

Modified GPSR Protocol using mobility in Wireless Sensor Networks

Abdullilah Alotaibi , Abdulaziz Almazyad
Computer Engineering Department,
King Saud University
Email: Abdullilahawad@yahoo.com, Mazyad@KSU.EDU.SA

Abstract

Greedy Perimeter Stateless Routing (GPSR) provides routing support for wireless sensor network (WSN) environment. IP addresses don't designate the sensor nodes in WSN so, the sensor node is designated by its location. When the receiver is outside the sensing field boundary, GPSR suffers by energy inefficiency because it has to route through all the sensor nodes in the boundary until reach the receiver. The GPSR utilizes a greedy forwarding strategy and perimeter forwarding strategy to route messages to receiver. It uses a neighbourhood message that contains the identity of a sensor node and its location. However, instead of sending this message periodically and the network is congested, GPSR piggybacks the neighbourhood message on each message that is sent or forwarded by the sensor node. GPSR finds a sensor node that is nearer to the receiver than itself and forwards the message to that sensor node. This method fails sometime so, the GPSR presents another method which is called perimeter routing which utilizes the right-hand graph traversal rule.

In this paper, we present modification to GPSR routing when the sensor node has data to transmit. We exploit the movement of data collector to guarantee the message that carry the location of data collector reach properly without unnecessary transmissions to outside the sensing region. Consequently the lifetime is maximized. Simulation results prove that our modification to GPSR increases the lifetime of WSN.

Keywords: wireless sensor network; network lifetime; data collector; GPSR

I. INTRODUCTION

Wireless sensor networks usually consist of a number of sensor nodes, which are battery-powered, a short-range wireless communication and a low capacity processor. The sensor nodes are energy constraint so, they send their data using multi hop to the sink .When any sensor node has data to transmit ,if it is in the sink communication range it will send directly , otherwise it will send using multi-hop to deliver its data to sink. Multi hop transmission results an unbalanced energy load around the sink [1]. The routing protocol should consumes less energy and also it decreases the delay during transferring information. Various routing protocols are available for wireless sensor networks. One of these routing protocols is GPSR which has two methods to transmit data.

Some of the researchers try to modify the GPSR routing, the work in [2] makes a slight modification to GPSR where the message originating at a node ni, is destined to an extreme point outside the sensing field, it is delivered to the nearest boundary node nj , and the location of nj node is returned to ni node. Subsequent announcement or query messages from ni are destined to nj directly instead of an extreme point, to avoid traversing the external face.

In our work we classify data into two types: delay sensitive data and delay tolerant data as presented in my previous work [3]. The data collector moves around the sensing region in anti clockwise direction and announce its location periodically to the sensors.

After cluster head selection, each cluster head publishes its position vertically, and the data collector sends queries horizontally asking about cluster head locations, and so, the data collector updates its trajectory accordingly.

Data collector sends queries horizontally to tell about its location and replies are returned with cluster heads location. Queries and replies are sent using GPSR to extreme points outside the sensing field. When the node has data to transmit, if it has delay tolerant data, data are sent directly to data collector if it is near or sent to the cluster head that the node belongs to. If the node has delay sensitive data to transmit data are sent directly to data collector but the location of data collector must be known. We did modification to GPSR where the sensor node has delay sensitive data to transmit, it needs first the data collector location. To get the location it sends queries horizontally and the replies are returned as the latest data collector location is known instead of completing the transmission to outside the sensing field.

In this paper, a modified GPSR is being proposed in mobile WSN environment where the data collector is mobile. It is proved to support for delivering messages efficiently in WSN.

To the best of our knowledge , this is the first method that exploit the movement of data collector with clockwise direction, and combines data gathering with clustering algorithm and mobility of data collector to deliver data online without synchronization between sensor nodes and the data collector. It maximizes life time of network and reduces the delay and traffic .

The rest of this paper is organized as follows. In Section II we describe GPSR algorithm. In Section III, we present our modification to GPSR routing scheme. In section IV, we present the results and analysis, finally we conclude the paper.

II GPSR Algorithm

Greedy Perimeter Stateless Routing (GPSR)[4],[5] is a routing protocol for wireless sensor networks that uses the locations of routers, and destination to forward

packets. It consists of two methods to forward packets: greedy forwarding, which is used wherever possible, and perimeter forwarding, which is used when the greedy forwarding fail.

In case of greedy forwarding the decisions is based only on information about a router's immediate neighbors in the network topology. if a node knows its radio neighbors' positions, the locally optimal choice of next hop is the neighbor geographically closest to the packet's destination. When a message reaches a region where greedy forwarding is failed, the perimeter forwarding is used and so on until the message reaches the destination. For example, when we have X,Y,Z,A nodes where X is the sender and Y,Z are neighbors of X and A is a destination, the distance between neighbors of X to destination A is tested and the nearest is taken to be the next hop. This greedy forwarding process repeats, until the message reaches node A. If the method is failed another method will be used (perimeter forwarding). Each node in GPSR has a neighbourhood table of its own. Whenever a message needs to be sent, the GPSR tries to find a node that is closer to the destination than itself and forwards the message to that node. GPSR can use local topology information to find correct new routes quickly.

III Modification to GPSR

Our proposed routing scheme in [3] is based on the fact that the data collector moves along averages of the sensing region in anti clockwise direction. The sensing region is divided into four squares and each square have center point , these points represent the trajectory of data collector. After cluster head selection, each cluster head publishes its location to the data collector and data collector updates its trajectory.

We classify the data into two types: delay sensitive data and delay tolerant data as in our previous work. In case of delay sensitive data, , relaying nodes (around data collector) receive data from other sensor nodes and send them to data collector or to a relaying node that currently near the data collector. Delay tolerant data are sent to a cluster head that the node belongs to, and the cluster head waits for the data collector to comes and pick up data. The data collector sends its location periodically. When any sensor nodes has data to transmit, it depends on the data type. If data is delay tolerant it will be sent to cluster head and cluster head deliver data to data collector. In case of delay sensitive data the data will be sent directly to the data collector in any location. To send data, the data collector location should be known to the nodes so, the sensor node sends queries horizontally to know the data collector location and the replies are returned with the latest location of data collector. Queries and replies are sent using GPSR algorithm. In our modification to GPSR instead of sending queries to outside the sensing region, the replies are returned as the latest location of data collector is known.

We exploit the movement direction of data collector in anti clockwise direction as shown in Fig 1.

when a sensor node has delay sensitive data to transmit, it will send queries horizontally to east and west directions to extreme points outside the sensing region. The replies are returned at the same path carrying the latest location of the data collector.

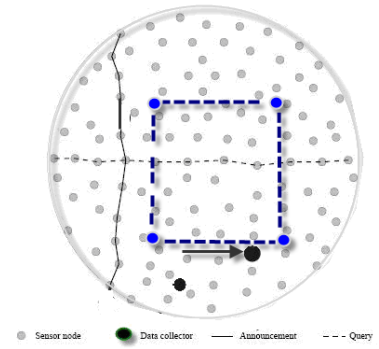


Fig 1: Direction of Data Collector in anti clockwise.

When the data collector changes its position, it will periodically send announcements vertically to north and south directions using GPSR with different time stamp. The data collector starts moving from the first location which is the center of the first square and moves toward the right of the sensing field in anti clockwise direction. When the data collector moves to right, it will send a new announcement with ascending order which means the time stamp for the data collector is sorted from earlier to latest as shown below in Fig 2.

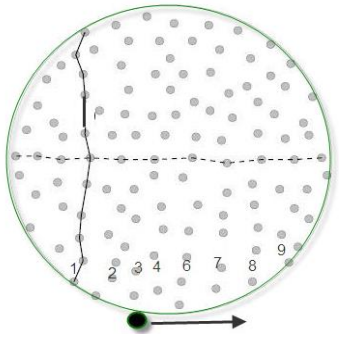


Fig 2: Numbering of Announcements.

The first timestamp is the earlier to take place, while the next is the current and so on. When a sensor node shown as a clear spot in the sensing field has delay sensitive data to transmit as shown in Fig 3; it will add two fields in its query message, the time stamp (timestamp) and data collector location(loc), and send the query message using GPSR to both directions east and west toward an extreme points outside the sensing field (-10, sendingnode.y) and (network dimension +10 , sendingnode.y). The sending node fills the two fields where loc is loaded with the location of data collector. At the same time the timestamp is loaded with any value like -1 in the initial because it is unrecognized, and then it sends the query message to both direction.

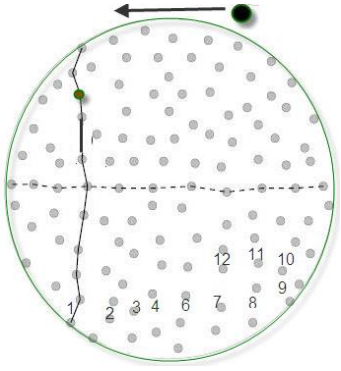


Fig 3: Sensor Node with Clear Spot has Sensitive Data.

When the sending node neighbor receives a query message either in both east, and west direction, it compares its own timestamp value with the timestamp value in the query message being received. If its time stamp value is greater than the transmitting node's time stamp value, the sending node neighbor will update the two fields with the latest values represented as a current location of data collector and the latest timestamp then resend the query message. This will be continued until the sensor node finds a time stamp greater than the query message timestamp. If so, it will update two fields with the latest information, and it will return a reply message carrying the latest information. Algorithm 1 describe our modification to GPSR routing

---

Algorithm 1: Modification to GPSR

Dc publishes vertically unique announcements using GPSR
Foreach sensor node ni wants to send data
Sendingnode ni create query message Q with two fields
Sendingnode ni.timestamp = -1
Sendingnode ni.location= empty
Sending node ni forward Q to neighbors
If neighbor ni.timestamp > Q.timestampe then
Q.timestamp = neighbor ni.timestamp
Q.location= neighbor ni.location
Neighbor ni Forward query message Q to another neighbor toward destination
Else
Q.timestamp = current timestamp
Q.location = current location
Return a reply message R with recent information
End
When a sending node ni receives two replies : ni send data to the latest DC location
using GPSR
End

IV. Results and Analysis

We use in our simulation the network sizes of 200,400,600,800 and 1000 sensor nodes that are randomly distributed in fields of 300*300,400*400,500*500,600*600,700*700 m^2 fields, respectively. we test 10 instances and take the average for each network size. We also use all the simulation parameters that were used in our work [3] and all the same assumptions.

we use the general energy consumption model that presented in [6] which can be as follows.

$ETr(r, b) = b \times (Eelec + Eamp \times r \hat{} \gamma)$ (1)

$ERc(b)=b \times Eelec$ (2)

where ETr (r, b) is the energy spent to send b bits over r m , ERc(b) is the energy spent to receive b bits, Eelec is the energy spent by the transmitter or receiver to send or receive one bit., Eamp is the energy spent by the transmission amplifier for one bit and γ is the path-loss exponent.

In this simulation we compare our proposed scheme in my previous work with two other schemes :a stationary scheme that has a stationary data collector and a mobile scheme that has a mobile data collector moves along boundary of the sensing field[2]. All experiments are done after our modification to GPSR.

Our scheme has the objective of maximizing the network lifetime and minimizing the delay by maximizing the packet delivery rate to the data collector.

We define the packet delivery rate in our work as the number of messages received successfully of data collector per round. Data collector changes its location periodically so, that the load is distributed evenly among all sensor nodes.

To determine a trajectory for mobile data collector along averages, we divide the sensing field into four equal-size squares; the trajectory of data collector is along averages of each square. Data collector moves to a cluster head that lies outside its communication range. Fig 4 shows that the lifetime of the network after our modification to GPSR. It is observed that the lifetime of the network after our modification to GPSR is improved with 35% over before modification to GPSR, where our modification is applied to our scheme (cluster) and mobile scheme.

Ultimately, our modification will decrease the number of transmissions that are sent by many sensor nodes in the network to get the latest location of data collector. Thus, sending unnecessary query message to a point outside the sensing field is cancelled, because the latest location of data collector is already available. This leads to minimize the overall energy consumption, and then maximize the lifetime of Wireless Sensor Networks

Fig 5 shows a comparison of the average energy consumed per bit between our scheme and the two other schemes for different number of nodes. Exploiting the movement of data collector produces a reduction in transmission over the network and also decreases the total consumed energy. Fig 6 and Fig 7 show the average number of sensitive messages and tolerant messages delivered successfully to data collector in a round.

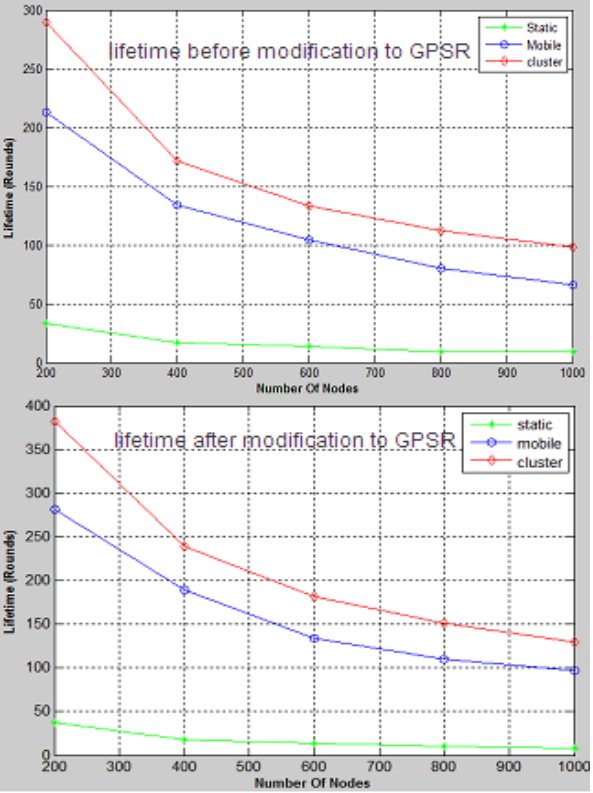


Fig 4: The Lifetime Comparison of Three Schemes Before and After modification to GPSR.

.

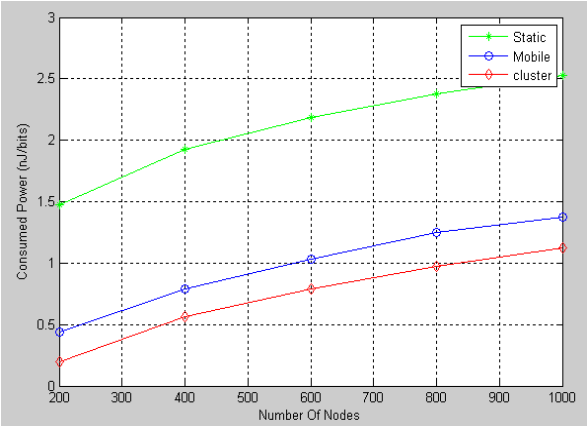


Fig 5: The Average Consumed Energy Comparison of Three Schemes after modification.

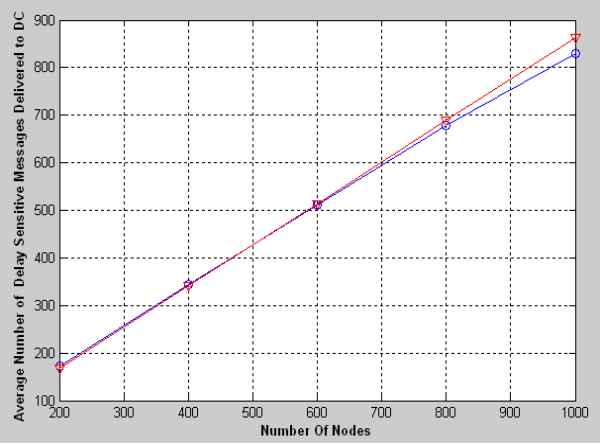


Fig 6: Average Number of Delay Sensitive Messages Delivered to DC in a Round after modification.

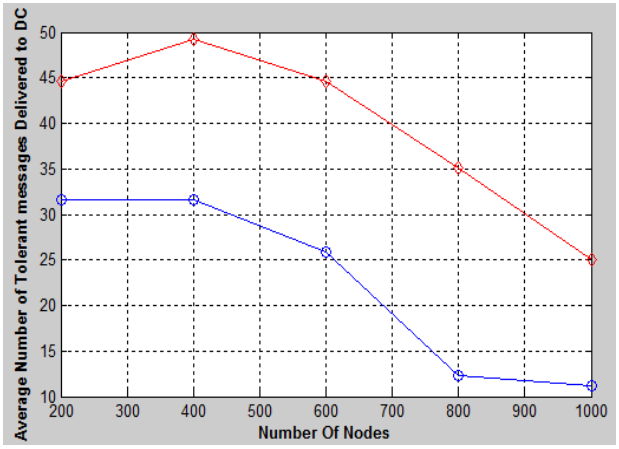


Fig 7: Average Number of Delay Tolerant Messages Delivered to DC in a Round.

V. Conclusion and future work

Our routing scheme with modified GPSR minimize the number of transmissions in the network which leads to maximize the lifetime of WSN and minimize the delay and the traffic in the network.

The results of the proposed routing indicates that the energy consumption is minimized.

We currently work with another scenario in which the sensor nodes send requests to the data collector to come and pick the delay tolerant data instead of using this trajectory so, we need to find a good route for the data collector to visit the sensor nodes which have sent data pick up requests.

References

[1] S. Gandham, M. Dawande, R. Prakash, and S. Venkatesan, "Energy efficient schemes for wireless sensor networks with multiple mobile base stations," in Proc. The IEEE Global Telecommunications Conference (GLOBECOM), December 2003.

[2] Waleed Alsalih, Hossam Hassanein, and Selim Akl "Routing to a mobile data collector on a predefined trajectory" in the IEEE ICC 2009 proceedings.

[3] Abdullilah Alotaibi , Abdulaziz Almazyad," Mobility of Data Collector Along Averages and Clustering Algorithm to Maximize the Lifetime in Wireless Sensor Networks", in ICWN - 12'The 2012 International Conference on Wireless Networks.

[4]B. Karp and H.Kung, "GPSR: greedy perimeter stateless routing for wireless networks," in Proc. The Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking, August 2000.

[5] L. Nithyanandan, G. Sivarajesh and P. Dananjayan, " Modified GPSR Protocol for Wireless Sensor Networks", in International Journal of Computer and Electrical Engineering, Vol. 2, No. 2, April, 2010 1793-8163

[6] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energyefficient communication protocol for wireless microsensor networks," in Proc. The 33rd Annual Hawaii International Conference on System Sciences, January 2000.

AODV Security Considerations

Asma Ahmed[1], **S. Razak**[2], **A. Hanan**[2], **Izzeldin Osman**[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 – *Ad hoc On-demand Distance Vector (AODV) is a reactive routing protocol in which the network generates routes at the start of communication. AODV has been developed specifically for Mobile Ad hoc Networks (MANETs). It obtains the routes purely on-demand which is makes it a very useful and desired algorithm for MANETs. AODV protocol is suffered from external and internal attacks which are disturb the protocol operation. This paper begins with review of AODV protocol and describes the mechanism of how AODV protocol is work. Then the security consideration in AODV is presented as well as the methods that were proposed to protect AODV from impersonation and modification attacks.*

Keywords: Ad hoc on-demand distance vector (AODV), malicious attacks, Routing, mobile ad hoc networks (MANETs).

1 Introduction

Mobile Ad Hoc Network (MANET) is autonomous and decentralized wireless systems. The open medium, rapidly changing topology and no centralized administration of MANET makes it more vulnerable to be attacked than the infrastructure networks [1]. Routing operations in MANET are even more likely to be disrupted where the unauthorized node can easily get access to the network and move around. In MANET, the correct transport of the packets depends on the honesty of the information given by the other nodes [3]. Several attacks can be initiated because weakness of the routing. The emission of false routing information by a node can create bogus entries in routing tables throughout MANET, which is decrease the efficient of the communication. In other case, by fooling the routing algorithm, a malicious node can control the traffic to and from entire parts of ad hoc network [4]. Ad hoc On-demand Distance Vector (AODV) [2] is one of the most popular routing protocols in MANET. AODV is a simple, efficient, and effective on-demand ad hoc routing protocol In AODV control packets carry important control information that governs the behavior of data transmission in MANET. Since the level of trust in a traditional network cannot be measured or enforced, enemy nodes or compromised nodes may participate directly in the route discovery and may intercept and filter the control packets to disrupt communication. Malicious nodes can easily cause redirection of network traffic and DoS attacks by simply altering these fields. The aim of this paper is to study how AODV is works and the security threats against its operation.

The rest of paper is organized as follows Section 2 provide an overview of AODV and explain how it works. Section 3 presents the characteristics that make AODV one of the most desirable protocols for MANET. Section 4 present the security threats in AODV and the solutions that are proposed to overcome the threats. The paper is concluded with plan for future work in Section 5.

.2 AODV Routing Protocol

AODV is a very simple, efficient, and effective routing protocol for MANET. This algorithm was motivated by the limited bandwidth that is available in the media that are used for wireless communications. Obtaining the routes purely on-demand makes AODV a very useful and desired algorithm for MANETs [2].

2.1 Routing Tables

Each mobile node in the network maintains a route table entry for each destination of interest in its route table. Each entry of this table contains the following information:

- Destination Node Address.
- Next hop of the source or intermediate node.
- Number of hops.
- Destination sequence number.
- Active neighbors for this route.
- Expiration time for this route table entry.

2.2 Control Packets

There are four messages used in AODV routing protocol.

These messages are used to control the process of route discovery and route maintenance.

1) Route Request Message (RREQ): When the source node want to connect with destination node and it has no route entry to the destination node, a control packet; named Route Request message (RREQ); was broadcasted by the source node. The request ID is incremented each time the source node sends a new RREQ. The pair (source address, and request ID) identifies a RREQ uniquely. As RREQ travels from node to node, it automatically sets up the reverse path from all these nodes back to the source. Each node that receives this packet records the address of the node from which it was received. This is called Reverse Path Setup (RPS).

2) Route Reply Message (RREP): If a node is the destination, or has a valid route to the destination, it unicasts a Route Reply message (RREP) back to the source. RREP message travels back to the source based on the reverse path that it records. As the RREP travels back to source, each node along this path sets a forward pointer to the node from where it is receiving the RREP and records the latest destination sequence number to the request destination. This is called Forward Path Setup (FPS).

3) Route Error Message (RERR): All nodes monitor their own neighborhood. When the route is broken or be invalid, a Route Error message (RERR) is generated to notify the other nodes that uses this route, that the route is became invalid. This is to avoid retransmitting by that route.

4) HELLO Message: Each node can get to know its neighbourhood by using local broadcasts, so-called HELLO messages. Nodes neighbours are all the nodes that it can directly communicate with. HELLO message is used to inform the neighbours that the link is still alive.

2.3 Sequence Numbers

The sequence number is an important feature of AODV to determine the freshness of routing information and guarantee loop-free routes. The destination sequence number for each destination node is stored in the routing table, and it is updated when the node receives message with a greater sequence number.

2.4 Route Discovery Process

When a source node wants to send a data packet to a destination node; first, it checks its routing table to determine if there is an available route to the destination node. If so, it uses this route to send the packets to destination node. In case where there is no route to destination node, route discovery process is initiated by broadcasting a RREQ message. If the node has already received a RREQ with the same source address and request ID, the new RREQ message will be discarded. The RREQ ID is increase by one every time the source node sends a RREQ message. Figure1 shows how RREQ message is propagating in MANET. In this figure, when the source node S wants to send a data packet to a destination node D, it has these steps:

- Node S sends RREQ to its neighbors; A, B.
- Node A sets up reverse path and forwards RREQ message to its neighbor D.
- Node B sets up reverse path and forwards RREQ message to its neighbor C.
- Node C sets up reverse path and forwards RREQ message to its neighbor D.
- When node D receive the RREQ from node C, it will discard it because it was already received it from node A.

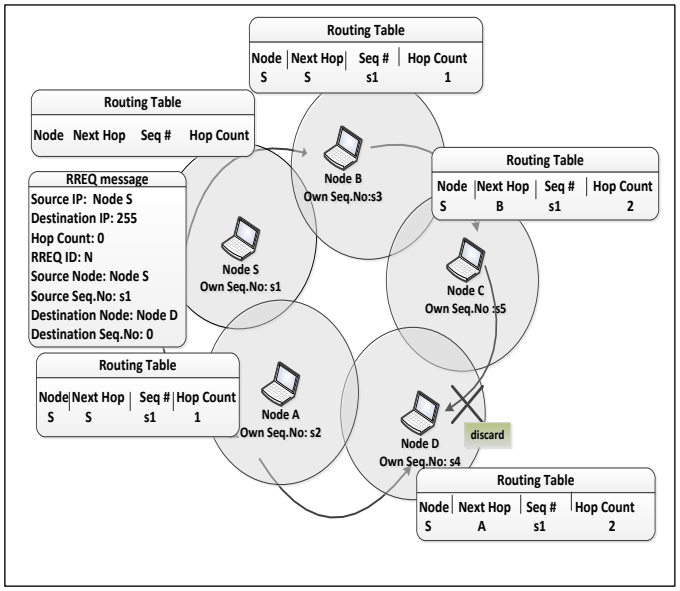


Figure1: RREQ message Propagation

If an intermediate node has a route entry for the desired destination in its routing table, it compares the destination sequence number in its routing table with that in the RREQ message. If the destination sequence number in its routing table is less than that in the RREQ, it rebroadcasts the RREQ to its neighbors. The exchange of route information will be repeated until a RREQ reaches at destination node or an intermediate

node that has a fresh enough route entry for the destination.

When the destination or intermediate node that has route to destination receives the RREQ, it sends a RREP to the source node and updates its routing table with accumulated hop count and the sequence number of the destination node. Afterwards the RREP message is unicasted to the source node. When the source node receives the RREP, then a route is established. Figure2 shows how RREP message is unicast in MANET. The explanation of the figure is as follows:

- Node D creates an RREP message and updates its routing tables with accumulated hop count (HC) and the sequence number.
- Then it unicast the RREP to node A.
- When node A receives RREP from node D, it updates its routing tables with accumulated HC and the sequence number of the node D.
- Then node A sets up forward path and forwards RREP to source node S.

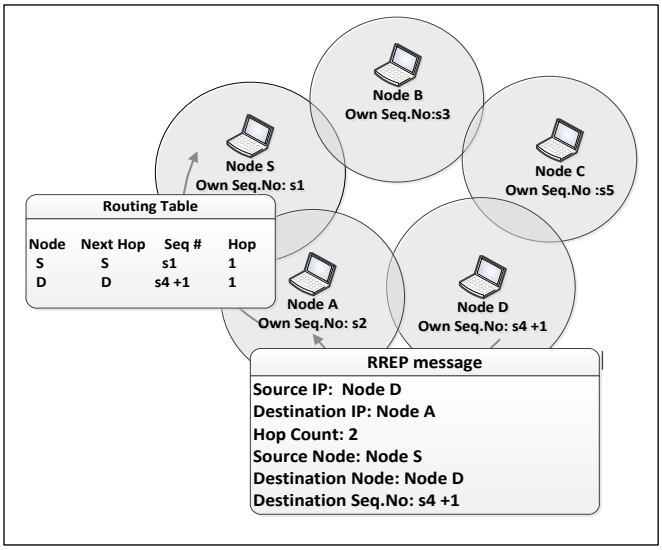


Figure2: RREP message Unicasting

In case a node receives multiple RREPs, the RREP have largest Dst-Seq is selected. If Dst-Seq were the same, then the smallest HC will selected. The HC is used to determine the shortest path and it is increased by 1 if a RREQ or RREP message is forwarded each hop. That means the intermediate node updates routing information and unicast new RREP only in these cases:

i) If the Dst-Seq is greater, or
ii) If the new Dst-Seq is same and HC is small.

Otherwise, it just skips the new RREP. This ensures that algorithm is loop-free and only the most effective route is used [2].

2.5 Link Breakage

Because a node in MANET can move at any time, link breakages can occur. If a node does not receive a HELLO message from one of its neighbours for specific amount of time called HELLO interval, then:

- The entry for that neighbour in the table will be set as invalid.
- The RERR message will be generated to inform other nodes of that link breakage.

During the route discovery process if any node identifies a link failure it generates Route Error message (RERR) and puts the invalidated address of that node into list, then it sends it to all other nodes which uses that link for their communication to other nodes. RERR messages inform all sources using a link when a failure occurs.

3 AODV Characteristics

AODV is one of the most popular routing protocols, which is a simple and efficient on-demand MANET routing protocol. The concepts of AODV that makes it desirable for MANETs with limited bandwidth include the following:

- Minimal space complexity: The algorithm makes sure that the nodes that are not in the active path do not maintain information about this route. After a node receives the RREQ and sets a reverse path in its routing table and propagates the RREQ to its neighbors, if it does not receive any RREP from its neighbors for this request, it deletes the routing info that it has recorded.
- Maximum utilization of the bandwidth: This can be considered the major achievement of the algorithm. As the protocol does not require periodic global advertisements, the demand on the available bandwidth is less, and a monotonically increased sequence number counter is maintained by each node in order to supersede any stale cached routes. All the intermediate nodes in an active path updates their routing tables also make sure of maximum utilization of the bandwidth. Since, these routing tables will be used repeatedly if that intermediate node receives any RREQ from another source for same destination. Also, any RREPs that are received by the nodes are compared with the RREP that was propagated last

using the destination sequence numbers and are discarded if they are not better than the already propagated RREPs.

- Simple: It is simple with each node behaving as a router, maintaining a simple routing table, and the source node initiating path discovery request, making the network self-starting.
- Most effective routing info: After propagating a RREP message, if a node receives RREP with smaller hop-count, it updates its routing info with this better path and propagates it.
- Most current routing info: The route info is obtained on demand. Also, after propagating an RREP, if a node receives RREP with greater destination sequence number, it updates its routing info with this latest path and propagates it.
- Loop-free routes: The algorithm maintains loop free routes by using the simple logic of nodes discarding the packets for same broadcast-id.
- Coping up with dynamic topology and broken links: When the nodes in the network move from their places and the topology is changed or the links in the active path are broken, the intermediate node that discovers this link breakage propagates an RERR message. And the source node re-initializes the path discovery if it still desires the route. This ensures quick response to broken links.
- Highly Scalable: The algorithm is highly scalable because of the minimum space complexity and broadcasts avoided.

4 AODV Security Considerations

4.1 Threats using Impersonation

Impersonate means the attacker assumes the identity of another node in the network, thus receiving the messages that are directed to the node that it fake [5][6][7]. In AODV route discovery, assume that node A establishes a route to another node B by sending a RREQ message towards it. Node B is supposed to reply the RREQ with RREP. However, any node who receives the RREQ is able to reply this RREQ. A malicious node can pretend to be the node B and reply a RREP, in order to redirect packets addressed to node B to itself. In the absence of any higher level authentication information, a malicious node can mislead node A into believing that it is communicating with node B. Thus, even though node A will finally receive multiple RREPs, the fake RREP having the shortest hop count will be accepted anyway.

Secure ad hoc routing protocols have been proposed as a technique to enhance the security in MANETs against impersonation attack. . Secure Ad hoc On-Demand Distance Vector Routing Protocol (SAODV) [8], which uses signed routing messages, is proposed to protect the routing messages of the original AODV protocol. SAODV is an extension of the AODV routing protocol that can be used to protect the route discovery mechanism providing security features like integrity, authentication and non-repudiation. SAODV used two mechanisms to authenticate the routing information that is digital signature [9] and hash chains [10].The survey in [18] overviewed the various secure routing protocols and pointed out their drawbacks and advantages. They also proposed a secure on-demand ad hoc network routing protocol (Ariadne) [19], which prevents the compromised nodes from tampering with the uncompromised routes, and the secure efficient ad hoc distance (SEAD) [20], which is a secure routing protocol, using efficient one-way hashing functions and not using asymmetric cryptographic operations. In [21], the authenticated routing for ad hoc networks (ARAN) is proposed by using public-key cryptographic mechanisms based on the AODV.
These methods can only guard against external attacks. However, the internal attacks mounted by the malicious or compromised hosts may still have a severe impact on the network performance, as well as on the connectivity among the nodes in the targeted MANET.

4.2 Threats using Modification

Modification usually performed by modifying the routing information aiming to compromise the integrity of routing computations. In AODV control packets carry important control information that governs the behavior of data transmission in MANET. Since the level of trust in a traditional network cannot be measured or enforced, enemy nodes or compromised nodes may participate directly in the route discovery and may intercept and filter the control packets to disrupt communication. Malicious nodes can easily cause redirection of network traffic and DoS attacks by simply altering these fields. In AODV the attacks can be classified as remote redirection attacks and DoS attacks as follows:

1) Remote Redirection with Modified Route Sequence Number**:** AODV uses the destination sequence number (Des-Seq) to determine the

freshness of routing information and guarantee loop-free routes. The destination sequence number is monotonically increasing number representing the freshness of routing request. The malicious node increase destination sequence number then it diverts traffic through itself by advertising a route to a node with a destination sequence number greater than the authentic value. Also it can decrease the distention sequence number on RREQ. The destination node receiving this RREQ will compare the latest received destination sequence number to the new one. If the new one has smaller value, this RREQ will be discarded. By making RREQs toward certain destination to be discarded, a denial-of-service attack is launched.

2) Redirection with Modified Hop Count: Attackers can also modify the hop count field to advocate the shortest route. When routing decisions cannot be made by other metrics, AODV uses the hop count field to determine a shortest path, by choosing route having the least hop count. In AODV, malicious nodes can attract route towards themselves by resetting the hop count field of the RREP to zero. Similarly, by setting the hop count field of the RREP to infinity, routes will tend to be created that do not include the malicious node. Once the malicious node has been able to insert itself between two communicating nodes it will be able to do anything with the packets passing between them. It can choose to drop packets to perform a DoS attack, or alternatively use its place on the route as a first step in man-in-the-middle attack.

Deng *et al.* [25] proposed an approach that requires the intermediate nodes to send a route reply (RREP) packet with the next hop information. When a source node receives the RREP packet from an intermediate node, it sends a "Further Request" packet to the next hop to verify that it has a route to the intermediate node and a route to the destination. As a response to this request, the intermediate node will send another RREP packet. When the next hop receives a "Further Request" packet, it sends a "Further Reply" packet that includes the verified result to the source node. Based on the information in the "Further Reply" packet, the source node judges the validity of the route. Again, the method in [26] requires the intermediate node to send the route confirmation request (CREQ) to the next hop node toward the destination, and then, the next hop node receives the CREQ and looks into its cache for a route to the destination. If it has such a route to the destination, then it sends a route confirmation reply (CREP) message to the source node with its route information. The source judges whether the path in RREP is valid by comparing the information with CREP. In these methods, the routing protocol has to be modified. These modifications may increase the routing overheads, which results in the performance degradation of the bandwidth-limited MANETs.

5 Discussion and Summary

This paper has discussed how the AODV routing protocol is works and also the features that make AODV the most desirable protocol for MANET environment has been explained. The security threats against AODV have been highlighted and the methods that have been proposed to overcome those threats have been discussed. Future works will be focused on construct a detection algorithm to handle the complex attack against AODV routing protocol in MANET.

References

[1] P.V.Jani, "Security within Ad Hoc Networks," Position Paper, PAMPAS Workshop, Sept. 16/17 2002.
[2] Charles E. Perkins. Ad Hoc Networking. Addision Wesley, 2001.
[3] H. Deng, W. Li, and D. P. Agrawal, "Routing security in ad hoc networks," IEEE Communications Magazine, vol. 40, no. 10, pp. 70-75, Oct. 2002.
[4] W. Lou and Y. Fang, A Survey of Wireless Security in Mobile Ad Hoc Networks: Challenges and Available Solutions. Ad Hoc Wireless Networks, edited by X. Chen, X. Huang and D. Du. Kluwer Academic Publishers, pp. 319-364, 2003.
[5] J. Y. choi. Security problems for ad hoc routing protocols survey paper. Technical report, Indian University at Bloomington, 2003.
[6]S. E. D Lim and X. Ee. Study of secure reactive routing protocols in mobile ad hoc networks. Technical report, National University Singapore, 2003.
[7] S. Gupte and M. Singhal. Secure routing in mobile wireless ad hoc networks. In Proceeding od Ad Hoc Networks. Volume 1, pages 151-174. Elsevier, 2003.
[8] Manel Guerrero Zapata, "Secure Ad hoc On Demand Distance Vector (SAODV) Routing", Technical University of Catalonia (UPC), Mobile Ad Hoc Networking Working Group, Internet Draft, 15 September 2005.

[9] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public key cryptosystems," Commun. ACM, vol. 21, no. 2, pp. 120–126, Feb. 1978.
[10] D. Eastlake, III and P. Jones, US Secure Hash Algorithm 1 (SHA1), Sep. 2001. IETF RFC 3174 (Informational).
[11] M.Abolhasan, T.Wysocki, E.Dutkiewicz, " A Review of Routing Protocols for Mobile Ad Hoc Networks," Telecommunication and Information Research Institute University of Wollongong, Australia, June, 2003.
[12] P. Yi et al., "A New Routing Attack in Mobile Ad Hoc Networks," Int'l. J. Info. Tech., vol. 11, no. 2, 2005.
[13] P. Michiardi and R. Molva. Ad hoc networks security. ST Journal of System Research, 2003.
[14] P. Papadimitratos and Z. J. Hass. "Secure routing for mobile ad hoc networks". In Proceedings of SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS), San Antonio, TX, January 2002.
[15] Levent, E. and Chavan, N. J. (2007). Elliptic Curve Cryptography based Threshold Cryptography (ECC-TC) Implementation for MANETs. IJCSNS International Journal of Computer Science and Network Security. 7(4), 48–61.
[16] Kong, J., Zerfos, P., Luo, H., Lu, S. and Zhang, L. (2001). Providing robust and ubiquitous security support for mobile ad-hoc networks. In International Conference on Network Protocols. Dept. of Comput. Sci., California Univ., Los Angeles, CA, USA, 251 – 260.
[17] Junaid Arshad and Mohammad Ajmal Azad, "Performance Evaluation of Secure on-Demand Routing Protocols for Mobile Ad-hoc Networks", 1-4244-0626-9/06 © 2006 IEEE.
[18] H. Yih-Chun and A. Perrig, "A survey of secure wireless ad hoc routing," IEEE Security Privacy, vol. 2, no. 3, pp. 28–39, May/Jun. 2004.
[19] H. Yih-Chun, A. Perrig, and D. Johnson, "Ariadne: A secure on-demand routing protocol for ad hoc networks," Wirel. Netw., vol. 11, no. 1/2, pp. 21–38, Jan. 2005.
[20] H. Yih-Chun, D. Johnson, and A. Perrig, "SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks," Ad Hoc Netw., vol. 1, no. 1, pp. 175–192, Jul. 2003.
[21] K. Sanzgiri, D. LaFlamme, B. Dahill, B. N. Levine, C. Shields, and E. M. Belding-Royer, "Authenticated routing for ad hoc networks," IEEE J. Sel. Areas Commun., vol. 23, no. 3, pp. 598–610, Mar. 2005.
[25] H. Deng, W. Li, and D. Agrawal, "Routing security in ad hoc networks," IEEE Commun. Mag., vol. 40, no. 10, pp. 70–75, Oct. 2002.
[26] S. Lee, B. Han, and M. Shin, "Robust routing in wireless ad hoc networks," in Proc. 31st ICPP Workshops, Aug. 2002, pp. 73–78.
[27] M.G.Zapata; "Secure On Demand Distance Vector (SAODV) Routing". INTERNET-DRAFT draft guerrero-manet-saodv- 06.txt, Sep. 2006.
[28] K. Biswas and Md. Liaqat Ali, "Security threats in Mobile Ad Hoc Network", Master Thesis, Blekinge Institute of Technology" Sweden, 22nd March 2007.
[29] Kim, J. and Tsudik, G. (2009). SRDP: Secure route discovery for dynamic source routing in MANETs. Ad Hoc Networks. 7(6), 1097 -1109. ISSN 15708705.
[30] J. Y. choi. Security problems for ad hoc routing protocols survey paper. Technical report, Indian University at Bloomington, 2003.
[31] J. Binkley and W. Trost. "Authenticated ad hoc routing at the link layer for mobile systems". Wireless Networks, 7(2): 139–145, 2001.
[32] Hu, Y. C., Perrig, A. and Johnson, D. B. (2003*). Packet leashes: a defense against wormhole attacks in wireless networks. vol. 3. 1976{1986 vol.3.*
[33] Hao Yang, Haiyun Luo. Fan Ye, Songwu Lu, and Lixia Zhang. "Security in mobile ad hoc networks: Challenges and solutions". IEEE Wireless Communications, February 2004.

A Framework for Multi-Mobile Orientation Models Using an Extended Kalman Filter and NFC

Trevor Zablocki
Computer Science Department
Central Michigan University
Mt. Pleasant, MI, U.S.A.

Roger Lee
Computer Science Department
Central Michigan University
Mt. Pleasant, MI, U.S.A.

Abstract - Mobile devices have become the primary method for person to person connection, and with new techniques of data creation and transfer, new methods of connecting people are constantly being developed. Many of these methods have incorporated location to enhance the user experience, but few of these use user location to its full potential and none use it to directly relate users to each other with significant precision. In this paper, a framework is discussed that could be implemented using various sensors and network data available on mobile devices to achieve a method for precisely relating two or more mobile devices based on distance and orientation.

Keywords: NFC, extended Kalman filter, wireless networks, Localization, orientation model

1.0 Introduction

Recently, there has been a surge in research and commercial use of mobile GPS and locational services [7]. This industry has experienced this sudden growth because of an increase in the availability of high precision sensors and GPS. Many GPS services can offer accuracy in measurements within the range of five percent [4]. These sensors and GPS capabilities can often be found in mobile phones and tablet computers. Although available, most of these services are seldom used in mainstream applications and are limited to simple gestures and movements [1]. Similarly, location is rarely accurately used as a multi-device information stream. It is important that location services overcome their current niche and become reliable and multi-device friendly.

For these reasons, the study of precise locational services in mobile devices have been highly theoretical with little application to real-world situations, often overcomplicating them. Modern systems' lack of speed, accuracy, and/or multi-device environments make them incapable of creating a real-world positioning model to compare several mobile devices. By combining several methods of mobile location tracking and the capabilities of Android mobile phones, such a model can be conceived. Through the use of sensors that measure a devices geolocation and real-space changes in position, we will show it is possible to create a real-time relational object that can be used to couple devices accurately.

In this paper, it is proposed that near field communication (NFC), mobile phone sensors, GPS, and network signatures can be used to create an extended Kalman filter (EKF) that can quickly and efficiently generate a multi-device orientation model. We describe each of these inputs and how they relate to each other to calculate usable positioning data.

In the following sections, we describe the current state of publicly available mobile devices and new technologies; and methodologies/applications of odometry. In section 3 we will discuss our approach and specific methods used in creating a relational model between mobile devices. In the last section, section 4, the paper is summarized and possible applications for this method are given.

2.0 Background

With the impressive accuracy that we have achieved with regards to global location-based technologies such as GPS, we find ourselves looking for more precision. One recent development in this movement is iGPS, or indoor Global Positioning System. This technique was developed to aid us in navigation through buildings and areas smaller than traditional GPS is capable of. Current systems achieving this are very complex, requiring the intricate mapping of entire buildings and plotting of several variables to this map; or in some cases systems of lasers to achieve the needed precision [8]. Other systems rely on wifi and cell phone tower connections to help generate these systems, which are also unreliable at generating precise positions. Most of these systems, aside from the rotating laser method, require a strong connection of some sort to an external system, including those developed at Duke University [6] and the University of Missouri [12]. In the method developed at Duke University, the phones' sensors are used in combination with an electronic map of an area to generate location [6]. This method is fairly accurate in areas without reliable GPS or phone signal, but the mapping of large areas in this way would be very hard to maintain and would become unreliable when these maps become out-of-date. Similarly, the method developed by Devin Smittle, Veselin Georgiev, Yi Shang, and Dan Wang creates reference points such as stride length using GPS [12]. By using GPS they allow this reference point, open to change at any time, to incur a high level of error. In the same way, there is no efficient and effective way of determining the relationship between several mobile devices. Indoor GPS systems are very important because they could make navigation of large buildings, such as a university or hospital, seamless.

Odometry is the use of data collected as movement is occurring to calculate the position of a device. The field of odometry may not be perfect, but through the use of correction factors, such as frequent calculations and measurements, it can be very precise and accurate. The extended Kalman filter can be used to reduce error in the odometry. This has been very efficiently used in mobile robotics, namely using a gyro sensor and sensors that collect information about its environment, to calibrate and account for any error that had been generated by estimations of location [2]. Although difficult, phones have many sensors that allow them to take in information about their environment. One of the most important environmental inputs is bluetooth connections. The bluetooth connection can be used to give an estimated distance between devices. Hossain AK and Soh Wee-Seng S. found that received power level is a good indicator of the distance between two bluetooth devices [5]. Sudden changes in this value could be used to help correct errors in odometry. In raw form this data is useless, but after being processed this data becomes very useful in positioning.

Mobile phones are improving in a variety of ways that make them increasingly capable of interacting with and responding to changes in their environment. The primary reason for this is the wide variety of sensors that now come standard in most mobile phones and their high level of precision. These sensors/services include: GPS, accelerometer, gyroscope, compass, microphone, wifi, bluetooth, barometer, and NFC [7]. These sensors/services create a very large reserve of information with ever increasing accuracy and precision that can be used to make estimations of the mobile phone/tablet's environment. These capabilities are frequently ignored because of the wide varieties of phones that exist with different components, ranges of accuracy, and operating systems. But these problems are diminishing as the components and capabilities of these devices become required by popular operating systems (Android and IOS). This increase in data availability in mobile phones and tablets makes them the perfect candidate to replace expensive systems that track environmental changes on a personal level.

3.0 Method

3.1 Basic Calculations

The raw data produced by a mobile devices sensors cannot be directly plugged into a simple extended Kalman filter, some basic calculations are needed to convert the raw data into information usable to the filter. On the most basic level, aside from the accelerometer and gyroscope data, the use of specific

orientation data is useful in making more accurate positioning calculations. This orientation data can be easily determined through the use of the accelerometer raw data and the raw data from the magnetic field sensor along with native Android functions. This function uses matrix theory to transform the product of the magnetic field sensor data and accelerometer sensor data into a three dimensional representation of the devices current orientation.

Another simple calculation that is made to help make the calculations more accurate is the calculation of the error propagation incurred when the data from the various sensors and services from the mobile device are combined. This error is simply calculated through the use of a differential equation of the calculations used to combine the data then later used to help account of possibly incurred error.

3.2 Position Estimation

The estimation of the position of the mobile devices involved is very important because without accurate estimations the model relating the devices becomes worthless. To achieve a high level of accuracy in our predictions an extended Kalman filter was chosen. It was also chosen because the level of noise in mobile device sensors is very high due to the unsteady movement of the operators.

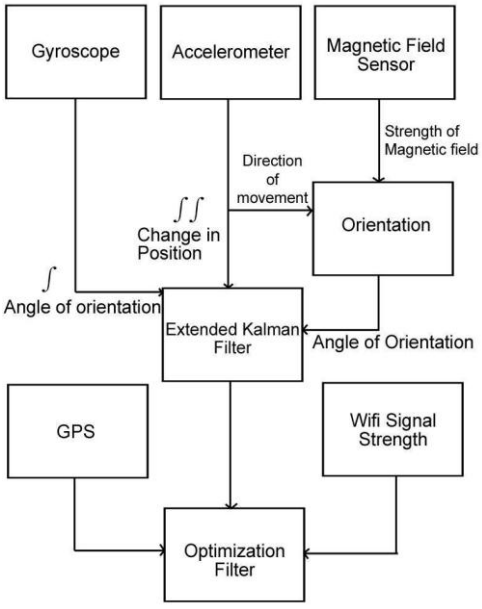


Figure 1. Position estimation data flow

3.2.1 Extended Kalman Filter

The extended Kalman filter was used to predict the position of mobile devices by predicting the position of the device with existing information and updating the model for later calculations. So this two piece process was used to account for error in each of the sensors and then update these methods of prediction for the next round of measurements from the sensors and other services. The first step, prediction, is fairly simple and involves the calculation of the orientation, speed, and change in position of the device. As the raw data is received and recorded, the error for each measurement is also calculated and accounted for.

The error model for the accelerometer

The calculations that use the accelerometer suffer from a very large rate of drift due to the fact that the raw data from this sensor is integrated twice to make it into useful positioning data. The position of the mobile device on the x, y, and z axis (represented by the variables x, y, and z) from the acceleration data, along with calculated bias drift vector B and velocity vector of the device v, is calculated in equation 1. Both of the vectors B and v are representations of the bias drift and velocity in the x, y, and z-axis directions respectively.

$$k = (k+1) - k \quad (1)$$
$$x(k+1) = x(k) + v_x(k) * k$$
$$y(k+1) = y(k) + v_y(k) * k$$
$$z(k+1) = z(k) + v_z(k) * k$$
$$v(k+1) = A(k) * k + v(k) + B(k)$$

where k is a given time; k+1 is the next time with recorded accelerometer data; k is the time between the two consecutive accelerometer recordings k and k+1.

The error model for the gyroscope

The calculations using the gyroscope do not face as much drift because the data only relies on a single integration. The position of the mobile device from the gyroscope data along with calculated drift for the gyroscope sensor is described in equation 2. A

similar equation was used by Zunaidi, Kato, Nomura and Matsui through the use of a gyroscope in a mobile robot [13].

$$\theta_x(k+1) = \theta(k) + \omega_x(k) * \Delta t + d_x(k) \quad (2)$$
$$\theta_{xe}(k+1) = x(k) + (\omega_x(k) + \delta\omega_x(k) * \Delta t + d_{xe}(k)$$
$$d_{xe}(k) = d_x(k) + \delta d_x(k)$$
$$x_e(k) = \theta_x(k) + \delta\theta_x(k)$$

where $\theta_{xe}(k)$ and $d_{xe}(k)$ are the estimations of the orientation and drift of the orientation around the x-axis respectively. $\omega_x(k)$ is the rate of rotation of the mobile device around the x-axis. This orientation can be calculated similarly around the y and z axis.

The error model for the orientation

The error/drift associated with the orientation as calculated by the accelerometer and magnetic field sensor data was calculated by equation 3 through the use of differential equations.
$f(a_1, a_2, a_3, m_1, m_2, m_3)$ is a pre-written function in the Android system to determine orienation.

$$e(k+1) = e(k) + a_1(f_{a_1}) + a_2(f_{a_2}) + a_3(f_{a_3}) + m_1(f_{m_1}) + m_2(f_{m_2}) + m_3(f_{m_3}) \quad (3)$$

where f is the pre-written android function to determine one of the three rotations around either the x, y, or z axis. x_1, x_2, and x_3 are the acceleration of the device in the x, y, and z directions with regards to the device. x_4, x_5, and x_6 are the magnetic field strengths along the x, y, and z axis with regards to magnetic north.

The error model for the wifi and GPS

The error associated with the GPS and wifi is given by the event listeners and are only used to reduce the size of the feasible area in which the mobile devices could lie in real space. So the error is simply what is returned with the GPS coordinates or wifi scan data.

Unlike the GPS data, the wireless connection strength (RSSI value) data had to be analyzed to calculate the distance of the user from the wireless router. This distance is given by equation 4 [10].

$$D = 10^{[(A-R)/k]} \quad (4)$$

where k (estimated for the conditions of the experiment as 3) is a constant determined by several factors including the environment and frequency that the wireless routers runs at and A is the connection strength at a distance of 0 meters from the router. R is the connection strength value in dBm. The value for A is calculated using linear regression and equation 4.

Implementation of the EKF

Using all of the previously stated methods for obtaining error for each method of obtaining orientation and change in position of the device, they can be combined to obtain better, more consistent estimations of position through the use of equation 5 and equation 6 [3].

State equations: (5)
$$x(k+1) = x(k)A(k) + s(k)$$

Orientation: (6)
$$f(k) = \delta\theta_o(k) - \delta\theta_g(k) + v(k)$$

where v(k)is the measurement noise for orientation, and s(k) is system noise. Both s(k)and v(k)can be understood as zero-mean Gaussian white noise sequences and A(k)act as system matrices [13].

3.2.2 Delayed optimization of position

Once the sensor data above is fused and used to generate a position and orientation of the device, the position can be further and more precisely chosen. Aside from the position of the device from the previous calculation, we can also obtain an error range for that value. The wifi connection strengths and GPS position measurements are still available for use. If the calculated position and error range of the device is treated as an objective function then the information from the GPS service and wifi connection strengths, along with error, can be used as constraints to create a smaller feasible set of positions for the device. This concept is illustrated in figure 2.

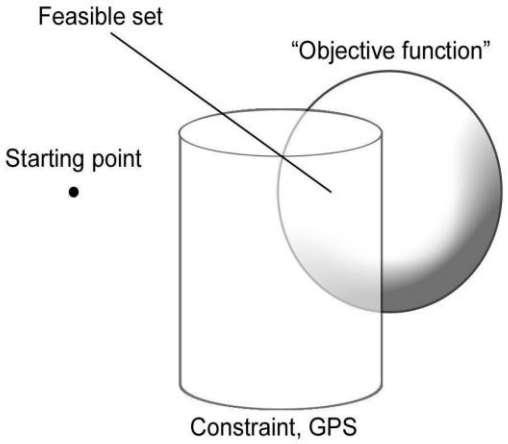


Figure 2. System optimization

Although only the position calculated by EKF and GPS are expressed in this figure the wireless signal strengths can be represented similarly as a constraint to the objective functions as the space between two cylinders, both centered at the starting point.

3.3 The Device Relational Model

After all of the data from the various mobile device sensors, services, and devices themselves are combined to generate a position for each of the mobile devices, the devices can then be related to each other with regards to their real-space positions. This is simply done by transmitting their coordinate information generated by the Kalman filter to the other over bluetooth. This information is then used to find the distance each device is from each other and their difference in orientation over each of the axis. This is simply done with the distance formula and subtraction of one devices orientation matrix from the others as is done in equation 7 and equation 8 respectively.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (7)$$

where x, y, and z are the coordinates in real space on the x, y, and z axis respectively.

$$o = [x_1, y_1, z_1]^T - [x_2, y_2, z_2]^T \quad (8)$$

where x is the pitch (rotation around the x axis), y is the roll (rotation around the y axis), z is the azimuth (rotation around the z axis), and o is the orientation of one device with respect to another.

4.0 Conclusion

We have presented a method for calculating and creating a multi mobile device relational model. This method differs from all techniques currently in use in both improved accuracy and real time capability. The approach is a feasible method of performing many location based, multi device activities such as mobile interactive gaming.

The NFC, EKF, and optimization based method that we have presented has many possible advantages over techniques currently in use. First, since this technique uses NFC to create a starting point for tracking the initial error of the positioning with respect to real space, all initial measurements can be taken to be zero since NFC only has an operating range of about 10 cm [9]. Second, by using several inputs to generate and update the position model, it becomes a lot more accurate and precise, in contrast to pure methods that only use GPS or wifi connections to generate a location [11]. Last, the method we presented has the advantage of not relying on any one source of data to generate a location; it can use many combinations of sensors and services to generate a location. Methods relying on one input cannot offer consistent accuracy or precision.

5.0 References

[1] Arase, Y., Ren, F., & Xie, X. (2010, September). User activity understanding from mobile phone sensors. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing* (pp. 391-392).

[2] Batlle, J. A., Font-Llagunes, J. M., & Barjau, A. (2010). Calibration for mobile robots with an invariant Jacobian. *Robotics and Autonomous Systems*, *58*(1), 10-15.

[3] Chui, C.K., and G. Chen. 1990. Kalman filtering– with real-time applications. Springer-Verlag second edition: ISBN 3–540–54013-X.

[4] Coutts, A. J., & Duffield, R. (2010). Validity and reliability of GPS devices for measuring movement demands of team sports. *Journal of Science and Medicine in Sport*, *13*(1), 133-135.

[5] Hossain AK and Soh Wee-Seng S., 2007, A comprehensive study of Bluetooth signal parameters for localization. In: IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, Greece, 1185-1189.

[6] I. Constandache, R. R. Choudhury, and I. Rhee. "Toward Mobile Phone Localization without War-Driving," presented at IEEE Infocom, San Diego, CA, 2010.

[7] Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE*,*48*(9), 140-150.

[8] Muelaner, J. E., Wang, Z., Martin, O., Jamshidi, J., & Maropoulos, P. G. (2011). Verification of the indoor GPS system, by comparison with calibrated coordinates and by angular reference. *Journal of Intelligent Manufacturing*, 1-9.

[9] Ortiz Jr, S. (2006). Is near-field communication close to success?. *Computer*,*39*(3), 18-20.

[10] Parameswaran, A. T., Husain, M. I., & Upadhyaya, S. (2009, September). Is rssi a reliable parameter in sensor localization algorithms: An experimental study. In *Field Failure Data Analysis Workshop (F2DA'09)*.

[11] Radu, V., Li, J., Kriara, L., Marina, M. K., & Mortier, R. (2012, June). Poster: a hybrid approach for indoor mobile phone localization. In *Proceedings of the 10th international conference on Mobile systems, applications, and services* (pp. 527-528). ACM.

[12] Smittle, D., Georgiev, V., Shang, Y., & Wang, D. (2010, July). Indoor localization on mobile phone platforms using adaptive dead reckoning. In *MU Summer Undergraduate Research and Creative Achievements Forum, Univ. of Missouri Columbia, MO*.

[13] Zunaidi, I., Kato, N., Nomura, Y., & Matsui, H. (2006). Positioning system for 4-wheel mobile robot: encoder, gyro and accelerometer data fusion with error model method. *CMU Journal*, *5*(1), 1-14.

Modeling Aspects of MIMO Communication Channels Based on Space-Time Block Codes

M. Andrei and V. Nicolau
Department of Electronics & Telecommunications
"Dunarea de Jos" University of Galati, 47 Domneasca St., 800008, Galati, Romania

Abstract - *In wireless communications, transmitting large amounts of data, at high speed and low error probability, is a major challenge. In this paper, modeling aspects of Multiple-Input Multiple-Output (MIMO) communication channels are studied, using binary phase-shift keying (BPSK) modulation and Alamouti space-time block code (STBC) with 2 transmitting and 2 receiving antennas. Alamouti code performance is simulated for different values of signal-to-noise ratio (SNR) and fading channel characteristics. Simulation results and performance analysis on image broadcasting using Alamouti STBC are presented for a grayscale rose image, which is more sensible to noise. The distribution analysis of error bits into the data stream and also into the error bytes are computed. The error bits are well distributed in accordance with the model of MIMO channel.*

Keywords: MIMO channels, space-time codes, wireless communications

1 Introduction

Digital wireless communications is a relative new but intensively studied domain, due to its major impact into many aspects of people's life. Transmitting large amounts of data through wireless communication channels, at high speed and low error probability, is a major challenge for the researchers in the field.

In general, communications through wireless channels are affected by time-varying characteristics of propagation environment. A signal sent from a transmitter traverses many different paths to its destination, in so-called multipath propagation, so that multiple versions of the transmitted signal reach the receiver [1]. Therefore, the received signal in communications through wireless radio channel cannot be modeled in a simple way as a copy of the transmitted signal corrupted by additive Gaussian noise [2].

On each path, a signal fading process is present, in which the signal experiences different attenuations, time delays, and phase shifts on one or more frequency components. The observed signal at the receiver is a sum of these multiple signals, being different from the original one transmitted through fading channel (FC). In addition, the number of paths and their characteristics can change in time, due to changing of relative positions of the transmitter, receiver, and also of the objects in the environment [3]. To improve data transmission through FC, coded modulation techniques were proposed [4].

There are moments of time, when the signal observed by a receiver is not sufficient to recover the actually transmitted signal, and this is an important problem in wireless communications. A solution is to transmit more replicas of the signal, using a technique called transmit diversity, which can be provided using temporal, frequency, polarization, and spatial resources [5].

Spatial diversity is obtained by deploying multiple antennas at both the transmitter and receiver. In this case, the wireless communication channel is Multiple-Input Multiple-Output (MIMO) type. Each antenna element in a MIMO system operates on the same frequency and therefore does not require extra bandwidth [6]. In addition, the total power through all antenna elements is less than or equal to that of a single antenna system.

Transmit diversity has been studied extensively as a method of combating detrimental effects in wireless fading channels due to its feasibility of having multiple antennas at base station [3], and relative simplicity of implementation [7]. Combined effects of transmitter diversity and channel coding were also studied [8].

Space-time codes (STC) are used in wireless communication systems with multiple transmitting antennas, to improve the reliability of data transmission, especially at high data rate [9]. Data streams are divided into multiple, redundant copies, which are transmitted to the receiver, assuring reliable decoding and full recovery of the data [10].

Space–time trellis coding (STTC) techniques extend trellis encoded modulation in spatial dimensions and combine coding techniques appropriate to multiple transmit antennas with complex decoding algorithms at the receiver [11]. These codes provide both diversity gain which can be maximized, and encoding gain depending on the complexity of the code (the number of states in the trellis) without loss of spectral efficiency.

Space-time block codes (STBC) reduce the decoder complexity of the receiver, using linear decoder based on orthogonal construction of the code matrix. However, STBC

codes lack of encoding gain [12]. The most practical STC are designed for two to four transmit antennas, which perform extremely well especially in slow fading environments. Alamouti space-time codes represent simple methods to develop space-time diversity, by using 2 transmitting antennas and N_R receiving antennas [7].

In this paper, modeling aspects of MIMO communication channels are studied, using binary phase-shift keying (BPSK) modulation and Alamouti STBC with 2 transmitting and 2 receiving antennas. Alamouti code performance is simulated for different values of signal-to-noise ratio (SNR) and fading channel characteristics. Also, simulation results on image broadcasting using Alamouti STBC are presented.

The paper is organized as follows. Section 2 describes mathematical models for MIMO communication systems and STC techniques. In section 3, Alamouti space-time code is described. Simulation results are presented in section 4 and conclusions are pointed out in section 5.

2 Mathematical models

In MIMO wireless communication system, multiple antennas at both transmitter and receiver are used. This allows multiple space-independent channels to be created. In addition, towards creating spatial diversity, the antenna arrays can be used to divert energy towards desired coordinates or to create parallel multiple channels used to stream (spatial multiplexing in transmission).

Consider a MIMO communication system with N_T transmitting and N_R receiving antennas, as shown in Fig. 1. In this paper, only propagation channels with flat fading and without memory are considered.

The signals submitted by the N_T antennas during a symbol period are denoted x_i, i = 1...N_T, where i index represents the signal emitted by the i antenna. These signals form the column vector **x** of size [N_T, 1]:

$$\mathbf{x} = \left[x_1, x_2 \ \ldots \ x_{N_T}\right]^T . \qquad (1)$$

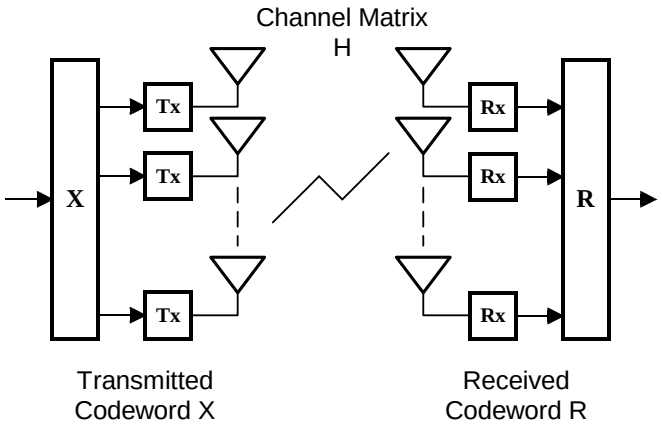


Figure 1. MIMO wireless communication system

The signals received by the N_R antennas during a symbol period form the column vector **r** of size [N_R, 1]:

$$\mathbf{r} = \left[r_1, r_2 \ \ldots \ r_{N_R}\right]^T . \qquad (2)$$

Similarly, the column vector of Gaussian noise is:

$$\mathbf{n} = \left[\eta_1, \eta_2 \ \ldots \ \eta_{N_R}\right]^T . \qquad (3)$$

The MIMO channel, considered without memory and with flat fading, is modeled by the channel matrix **H**, of size [N_R, N_T], which is also called the transfer function of MIMO channel. It contains the channel fading coefficients, h_{jk}, between the broadcasting antenna k and the receiving antenna j. At every *t* moment of time, the channel matrix is:

$$\mathbf{H}_t = \begin{bmatrix} h^t_{1,1} & h^t_{1,2} & \cdots & h^t_{1,N_T} \\ h^t_{2,1} & h^t_{2,2} & \cdots & h^t_{2,N_T} \\ \vdots & \vdots & \ddots & \vdots \\ h^t_{N_R,1} & h^t_{N_R,2} & \cdots & h^t_{N_R,N_T} \end{bmatrix}. \qquad (4)$$

The coefficients of **H** matrix can be deterministic or random. There are several statistical models for representing random coefficients. For a Rayleigh fading channel, which is considered in this paper, the fading coefficients are complex Gaussian random variables. They contain independent random variables for real and complex part, with identical distributions with zero mean.

The MIMO channel linear input-output relationship is:

$$\mathbf{r} = \mathbf{H}\cdot\mathbf{x} + \mathbf{n}. \qquad (5)$$

In general, channel matrix **H** varies in time. The MIMO channel behavior can be characterized based on the changing rate of channel fading coefficients, which defines the channel coherence time, denoted t_c.

If the channel matrix **H** varies slowly in time, being constant during the transmission of an entire frame with L symbols, but changing from frame to frame, then the channel is quasi-static or slow fading. In this case, the channel parameters vary more slowly than those of the base-band signal, and the channel coherence time is bigger than the time of frame transmission, T_F: $T_F = L\cdot T < t_c$, where L is the number of symbols in the frame, and T is the transmission time of a symbol.

If the channel matrix **H** remains constant during the symbol transmission, but varies from symbol to symbol during the frame transmission, then the channel is fast fading. In this case, the coherence time is: $T < t_c < L\cdot T$.

One part of the transmitted signals goes through propagation channels and it is received subsequent to the N_R receiving antennas. At every *t* moment of time, the received signal by the receiving antenna j, denoted $r_{t,j}$, is a linear combination of all signals, fading and channel-added noise:

$$r_j^t = \sum_{i=1}^{N_T} h_{ji}^t \cdot x_i^t + \eta_j^t \,, \qquad (6)$$

where η_j^t is Gaussian noise, with zero mean and σ^2 diversity.

Taking into account the transmission of a frame with L symbols, which are transmitted successively during the frame time, denoted T_F, the column vector **x** transforms into space-time codeword matrix **X** of size $[N_T, L]$:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^L \\ x_2^1 & x_2^2 & \dots & x_2^L \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_T}^1 & x_{N_T}^2 & \dots & x_{N_T}^L \end{bmatrix}. \qquad (7)$$

Each j column represents the signals submitted by the N_T transmitting antennas during of j symbol period, j = 1…L.

Similarly, column vectors **r** and **n** transform into matrices of size $[N_R, L]$:

$$\mathbf{R} = \begin{bmatrix} r_1^1 & r_1^2 & \dots & r_1^L \\ r_2^1 & r_2^2 & \dots & r_2^L \\ \vdots & \vdots & \ddots & \vdots \\ r_{N_R}^1 & r_{N_R}^2 & \dots & r_{N_R}^L \end{bmatrix}. \qquad (8)$$

$$\mathbf{N} = \begin{bmatrix} \eta_1^1 & \eta_1^2 & \dots & \eta_1^L \\ \eta_2^1 & \eta_2^2 & \dots & \eta_2^L \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{N_R}^1 & \eta_{N_R}^2 & \dots & \eta_{N_R}^L \end{bmatrix}. \qquad (9)$$

In this case, the input-output model for MIMO slow fading channel is:

$$\mathbf{R} = \mathbf{H} \cdot \mathbf{X} + \mathbf{N}. \qquad (10)$$

For fast fading channel, the model is:

$$\mathbf{R} = [\mathbf{H_1} \cdot \mathbf{x_1}, \mathbf{H_2} \cdot \mathbf{x_2}, \dots, \mathbf{H_L} \cdot \mathbf{x_L}] + \mathbf{N}. \qquad (11)$$

Space-time codes are used to improve the reliability of data transmission, especially at high data rate, by increasing spatial diversity and minimizing the likelihood of error.

In space-time coding, for every symbol in the data sequence, each transmitting antenna transmits a different version of the same input, generated by the space-time encoder. A MIMO communication system with STC is illustrated in Fig. 2.

During one symbol, the space–time encoder generates N_T modulated complex symbols, which form the column vector **x**, as input to the MIMO wireless communication channel.

At the receiver, the N_R received signals, which form the column vector **y** as output from MIMO channel, are used by the space-time decoder to obtain the original symbol.

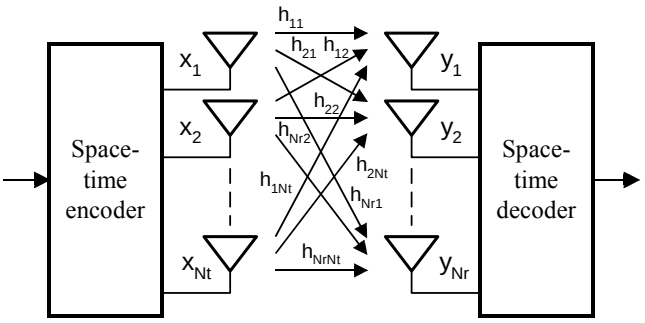


Figure 2. MIMO communication system using space-time coding

It is assumed that the receiver decoder estimates the maximum plausible sequence of information transmitted. At the reception, squared Euclidean distance is used between presumably received sequence and the received one:

$$\sum_t \sum_{j=1}^{n_R} \left| y_t^j - \sum_{i=1}^{n_T} h_{j,i}^t x_t^i \right|^2 . \qquad (12)$$

The space–time codes provide the best possible tradeoff between constellation size, data rate, diversity advantage, and trellis complexity. When the number of transmit antennas is fixed, the decoding complexity of space–time trellis coding (measured by the number of trellis states in the decoder) increases exponentially as a function of both the diversity level and the transmission rate.

3 Alamouti space-time codes

Trying to solve the problem of decoding complexity at the receiver, Alamouti had discovered a remarkable scheme for transmission, using only two transmitting antennas. Based on orthogonal construction of the code matrix, Alamouti codes reduce the decoder complexity of the receiver, using linear decoder. But, low encoding complexity is obtained, in exchange for a reduced encoding gain, which is the main drawback of these codes.

Space-time block codes generalize the transmission scheme discovered by Alamouti to an arbitrary number of transmitting antennas and are able to achieve the full diversity promised by the transmitting and receiving antennas. These codes retain the property of having a very simple maximum likelihood decoding algorithm based only on linear processing at the receiver.

The structure of STBC encoder, which generates codeword matrix **X**, is illustrated in Fig. 3.

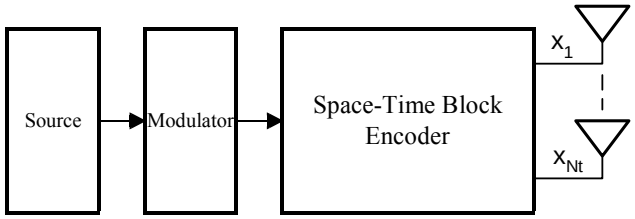


Figure 3. Space-time block encoder

At each *t* moment of time, the source generates an *m*-block of binary information, which is first modulated using a constellation C with 2^m points. The modulator produces a modulated block of K real or complex symbols from C. Next, the space-time block encoder generates the codeword matrix **X** of size [N_T, L], which is transmitted through N_T antennas, in L time slots equals to L symbol periods. The **X** matrix elements are linear combinations of the K modulated symbols and their conjugates.

In STBC, the number of symbols that encoder receives as inputs to each encoding operation is K. The number of transmission periods used to forward the coded symbols through each antenna is L. In other words, L symbols are transmitted by each antenna for each input block of K symbols. In this case the coding rate is:

$$R = \frac{K}{L}. \tag{13}$$

A *j* column of **X** matrix represents broadcasted symbols by all N_T antennas during *j* time slot, while *i* line contains the L sequentially broadcasted symbols by *i* antenna:

$$\mathbf{x}_i = [x_{i1}\ x_{i2}\ \ldots\ x_{iL}],\ i = 1\ldots N_T. \tag{14}$$

The codeword matrix **X** is built with lines being orthogonal each other, resulting:

$$\langle x_i, x_k \rangle = \sum_{j=1}^{L} x_{ij} \cdot x_{kj}^* = 0,\ i \neq k,\ \ i,k \in \{1,\ldots N_T\}, \tag{15}$$

where < , > is the inner product for vectors with complex elements.

As a result, the **X** matrix has the property:

$$\mathbf{X} \cdot \mathbf{X}^H = c \cdot \left(|x_1|^2 + |x_2|^2 + \ldots + |x_K|^2 \right) \cdot \mathbf{I}_{N_T}, \tag{16}$$

where $\mathbf{X}^H$ is complex conjugate transpose of **X**, *c* is a constant and $\mathbf{I}_{NT}$ is the unit matrix of size [N_T, N_T].

In this paper, Alamouti space-time code with $N_T = N_R = 2$ is used. The Alamouti encoder structure is shown in Fig. 4. For example, in the case of Quadrature phase-shift keying (QPSK) modulation, the modulator gets at the input two bits of information and produces two complex symbols [x_1 x_2], with $x_i \in C = \{1, j, -1, -j\}$. If Binary phase-shift keying (BPSK) modulation is used, then constellation C has real points and $x_i \in C = \{1, -1\}$.

Encoder outputs are transmitted during two consecutive symbol periods through the two transmitting antennas. For each encoding operation, the encoder takes the group of two modulated complex symbols, and generates the codeword matrix X:

$$\mathbf{X} = \begin{array}{c|cc} & t & t+T \\ \hline T_{x1} & x_1 & -x_2^* \\ T_{x2} & x_2 & x_1^* \end{array}. \tag{17}$$

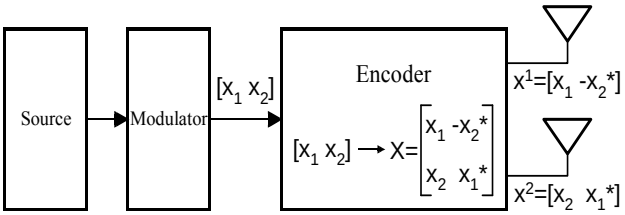


Figure 4. Alamouti encoder with PSK modulation

During the first period, the two signals x1 and x2 are transmitted simultaneously through the first and second antenna, respectively T_{x1} and T_{x2}. In the second period, T_{x1} antenna broadcasts $-x_2^*$ signal, and T_{x2} antenna broadcasts x_1^* signal. The inner product of the two lines is:

$$< \mathbf{x}^1, \mathbf{x}^2 > = x_1 x_2^* - x_2^* x_1 = 0. \tag{18}$$

The **X** matrix has the same property as all STBC:

$$\mathbf{X} \cdot \mathbf{X}^H = \left(|x_1|^2 + |x_2|^2 \right) \cdot \mathbf{I}_2. \tag{19}$$

The signals received by *j* antenna at t and t + T time slots are $r_{j,1}$ and $r_{j,2}$, respectively:

$$\begin{cases} r_{j,1} = h_{j,1} \cdot x_1 + h_{j,2} \cdot x_2 + \eta_{j,1} \\ r_{j,2} = -h_{j,1} \cdot x_2^* + h_{j,2} \cdot x_1^* + \eta_{j,2} \end{cases}. \tag{20}$$

The matrix form of (21) is:

$$\mathbf{r}_j = \begin{bmatrix} r_{j,1} & r_{j,2} \end{bmatrix} = \begin{bmatrix} h_{j,1} & h_{j,2} \end{bmatrix} \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} + \begin{bmatrix} \eta_{j,1} & \eta_{j,2} \end{bmatrix}. \tag{21}$$

The decoder uses a maximum plausibility algorithm, selecting the most likely look symbols $\hat{x}_1$ and $\hat{x}_2$. Considering a source of information without memory, x_2 and x_1 modulated symbols are independent to each other. Hence, it is possible separate decoding of the two symbols:

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} \arg\min\limits_{\hat{x}_1 \in S} \left(\sum\limits_{j=1}^{N_R} \left| \tilde{r}_{j,1} - \left(|h_{j,1}|^2 + |h_{j,2}|^2 \cdot \hat{x}_1 \right) \right|^2 \right) \\ \arg\min\limits_{\hat{x}_2 \in S} \left(\sum\limits_{j=1}^{N_R} \left| \tilde{r}_{j,2} - \left(|h_{j,1}|^2 + |h_{j,2}|^2 \cdot \hat{x}_2 \right) \right|^2 \right) \end{bmatrix}. \tag{22}$$

4 Simulation results

In this section, Alamouti code performance on large data transmission is studied, and some simulation results on image broadcasting using Alamouti STBC are presented.

For simulations, a MIMO channel is considered, which is affected by Rayleigh slow fading. Two solutions were chosen, with $N_T = 2$ and $N_R = 1$, denoted solution S_2, and $N_T = 2$ and $N_R = 2$, denoted solution S_3, respectively. They are compared with SISO channel, denoted S_1. The channel coherence time is: $t_c \geq T_F = 2 \cdot T$.

Different coherence time values were chosen in simulations: $t_c = N_F \cdot T_F$, where N_F is the number of frames in which the fading coefficients are constant, N_F = 1, 10 and 100 respectively. The decoder uses maximum plausibility algorithm to estimate broadcasted symbols, based on separate decoding presented in (22).

To simulate Alamouti code performance, a large random data sequence is used. It contains 10^6 bits of 0 and 1 values with equal probability. The data set is modulated using BPSK modulation, resulting constellation C with real points and C = {1, -1}. The modulated data sequence is divided into $5 \cdot 10^5$ blocks of 2 bits and the Alamouti encoder forms $5 \cdot 10^5$ different **X** matrices of size [2, 2].

The bit error rate (BER) is the performance criterion. The BER curves are obtained depending on the signal-to-noise ratio (SNR), which is a chosen vector in the range [2, 27] dB. For every SNR value, the entire data sequence is transmitted through MIMO channel and the error vector between received and original data sequences is computed. The simulation results are illustrated in Fig. 5.

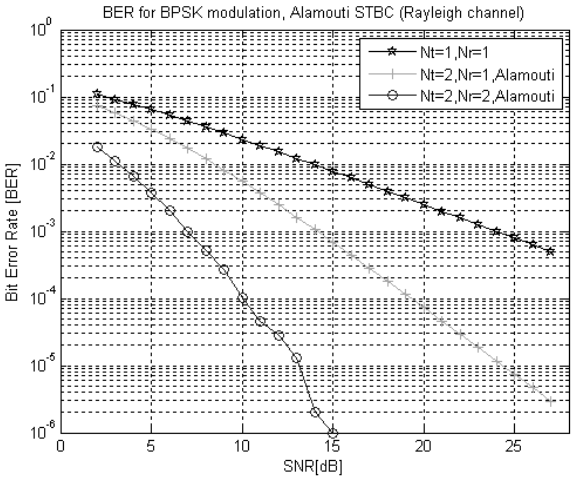


Figure 5. BER performance of Alamouti code for a random data sequence of 10^6 bits

The computed BER of Alamouti STBC with N_T=2 and N_R=2 (solution S3) is represented with "o" magenta mark, solution S2, with Alamouti code with N_T=2 and N_R=1 is drawn with "+" green mark, and the simple communication channel (solution S1) is drawn with "*" blue mark.

It can be observed that, for Alamouti STBC with N_T=2 and N_R=2, the BER curve decreases much faster than other two solutions. BER numerical values for two different values of SNR used for image transmission next are represented in Table I.

TABLE I. BER numerical values for three solutions

SNR [dB]	**BER*1000**		
	S1 Solution	***S2 Solution***	***S3 Solution***
5	64.1827	32.8577	3.7270
10	23.2687	5.5282	0.1100

The image transmission, through MIMO channel using BPSK modulation and Alamouti STBC with N_T=2 and N_R=2, is simulated on a grayscale rose image, which is more sensible to noise. The uncompressed image has 512*512 pixels with 8 bit grayscale, being illustrated in Fig. 6.

Figure 6. Original image with 512*512 pixels and 8 bit grayscale

After modulation, the data stream has 2^{21} bits, and it is divided into 2^{20} symbol blocks. The image transmission is repeated for two SNR values: 5 and 10 [dB], respectively.

The received image for SNR = 5 [dB], illustrated in Fig. 7, is compared with original one, and an error image is computed and represented in Fig. 8.

Figure 7. The received image, after Alamouti decoding

The bit error number (BEN) is computed by comparing the original data sequence and the estimated one determined by the Alamouti decoder. The numerical values of BEN and the corresponding computed BER (BER_C) are represented in Table II, for the two cases of SNR.

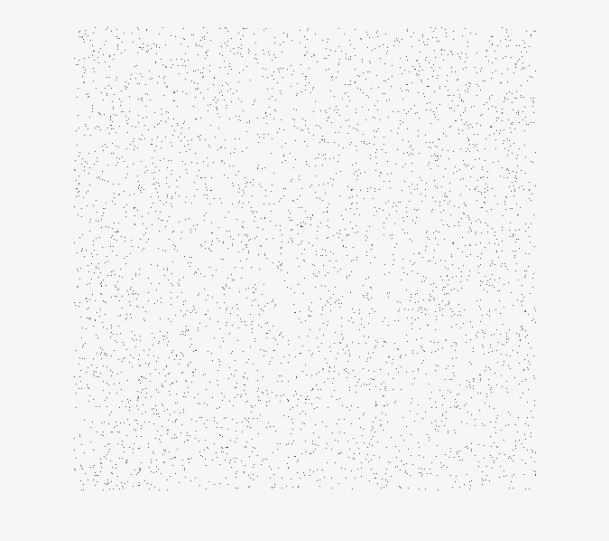


Figure 8. The error image

TABLE II. BEN and BER_C numerical values for S3 solution

SNR [dB]	Data set = 2.097.152 bits, S3 Solution		
	BEN	***BER_C*1000***	***BER*1000***
5	7710	3.6764	3.7270
10	238	0.1135	0.1100

BEN drastically decreases in the second case, with SNR = 10 [dB]. In addition, the corresponding BER_C values are comparable in both situations with real BER values, determined from Alamouti code performance simulation.

The distribution analysis of error bits into the data stream shows that the error bits are well distributed in accordance with the model of MIMO channel. The error bytes in the received image contain 1 or more error bits, most of them being affected by only one error bit. The number of error bytes affected by 1 or more error bits is shown in Table III. The number of error bytes which are affected by 2 or more error bits decreases exponentially.

TABLE III. Number of error bytes affected by error bits

SNR [dB]	The number of error bytes with 1 or more error bits			
	1 error bit/ byte	***2 error bits / byte***	***3 error bits / byte***	***4 error bits / byte***
5	7291	203	3	1
10	234	2	-	-

5 Conclusions

A MIMO channel model with BPSK modulation and Alamouti space-time block code with 2 transmitting and 2 receiving antennas was studied. The bit error rate was determined by simulations for different values of SNR. Also, some simulation results on image broadcasting using Alamouti STBC are presented. The decoder estimates well the broadcasted symbols, even for small values of SNR.

6 References

[1] G. J. Pottie, "System design issues in personal communications", IEEE Personal Communications. Mag., vol. 2, no. 5, pp. 50–67, 1995.

[2] H. Jafarkhani, Space-Time Coding: Theory and Practice, Cambridge, 2005.

[3] G.J. Foschini Jr., and M.J. Gans, "On limits of wireless comm. in a fading environment when using multiple antennas",Wireless Personal Comm, vol.6, pp.311-335, 1998.

[4] C.-E. W. Sundberg and N. Seshadri, "Coded modulation for fading channels: An overview", European Trans. Telecommun. Related Technol., pp. 309–324, 1993.

[5] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels", In Proc. IEEE VTC'96, pp. 136–140, 1996.

[6] N. Balaban and J. Salz, "Dual diversity combining and equalization in digital cellular mobile radio", IEEE Trans. Veh. Technol., vol. 40, pp. 342–354, 1991.

[7] S.M. Alamouti, "A simple transmit diversity technique for wireless communications", IEEE Journal on Selected Areas in Comm. 16 (8): 1451–1458, 1998.

[8] A.Hiroike, F.Adachi, N.Nakajima, "Combined effects of phase sweeping transmitter diversity and channel coding", IEEE Trans. Veh. Technol., vol.41, pp.170–176, 1992.

[9] V. Tarokh, N. Seshadri and A. R. Calderbank, "Space–time codes for high data rate wireless communication: Performance analysis and code construction", IEEE Transactions on Information Theory, 44 (2): 744–765, 1998.

[10] G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communications", IEEE GLOBECOM'96, pp. 1809–1814, 1996.

[11] S.M. Alamouti, V. Tarokh and P. Poon, "Trellis coded modulation and transmit diversity: Design criteria and performance evaluation", Proc. IEEE ICUPC 98, pp. 703–707, 1998.

[12] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space–time block codes from orthogonal designs", IEEE Trans. on Information Theory, vol. 45 (5): 1456–1467, 1999.

Study of the Feasibility of Radio over Fiber Technology for WiMAX Systems

Nisar Khan
College of Engineering
King Saud University
Riyadh, Saudi Arabia
nmohmand@ksu.edu.sa

Ahmad Yahya M. Faloos
College of Engineering
King Saud University
Riyadh, Saudi Arabia
floos@ksu.edu.sa

Muhammad Zeb Khan
College of Engineering
King Saud University
Riyadh, Saudi Arabia
zebkhan@ksu.edu.sa

***Abstract*— To meet the explosive demands of high-capacity and broadband wireless access, modern cell based wireless networks have trends, i.e., continuous increase in the number of cells and utilization of higher frequency bands. It leads to a large amount of base stations (BSs) to be deployed; therefore, cost-effective BS development is a key to success in the market. In order to reduce the system cost, radio over fiber (RoF) technology has been proposed since it provides functionally simple BSs that are interconnected to a central control station (CS) via an optical fiber. The well known advantages of optical fiber as a transmission medium such as low loss, light weight, large bandwidth characteristics, small size and low cable cost make it the ideal and most flexible solution for efficiently transporting radio signals to remotely located antenna sites in a wireless network. In addition to its transmission properties, the insensitivity of fiber optic cables to electromagnetic radiation is a key benefit in their implementation as the backbone of a wireless network. This paper will provide an overview of RoF technology, followed by the description of suitable architectures for the deployment of WiMAX networks employing RoF systems. Main issues and challenges in the deployment of WiMAX employing RoF technology will be discussed in detail after reviewing some experimental and theoretical work. Furthermore, a simulation model for IEEE 802.16e is studied and simulation results are shown.**

Keywords- WiMAX; Radio over Fiber Technology; Broadband Wireless

I. INTRODUCTION

For the future provision of broadband, interactive and multimedia services over wireless media, current trends in cellular networks - both mobile and fixed - are 1) to reduce cell size to accommodate more users and 2) to operate in the microwave/millimeter wave (mm-wave) frequency bands to avoid spectral congestion in lower frequency bands. It demands a large number of base stations (BSs) to cover a service area, and cost-effective BS is a key to success in the market. This requirement has led to the development of system architecture where functions such as signal routing/processing, handover and frequency allocation are carried out at a central control station (CS), rather than at the BS. Furthermore, such a centralized configuration allows sensitive equipment to be located in safer environment and enables the cost of expensive components to be shared among several BSs. An attractive alternative for linking a CS with BSs in such a radio network is via an optical fiber network, since fiber has low loss, is immune to EMI and has broad bandwidth. The transmission of radio signals over fiber, with simple optical-to-electrical conversion, followed by radiation at remote antennas, which are connected to a central CS, has been proposed as a method of minimizing costs. The reduction in cost can be brought about in two ways. Firstly, the remote antenna BS or radio distribution point needs to perform only simple functions, and it is small in size and low in cost. Secondly, the resources provided by the CS can be shared among many antenna BSs. This technique of modulating the radio frequency (RF) subcarrier onto an optical carrier for distribution over a fiber network is known as *"radio over fiber"* (RoF) technology.

On the other hand, to meet the explosive demands of high-capacity and broadband wireless access, millimeter-wave (mm-wave) radio links (26 - 100 GHz) are being considered to overcome bandwidth congestion in microwave bands such as 2.4 or 5 GHz for application in broadband micro/picocellular systems, fixed wireless access and WLANs. The larger RF propagation losses at these bands reduce the cell size covered by a single BS and allow an increased frequency reuse factor to improve the spectrum utilization efficiency. Recently, considerable attention has been paid in order to merge RoF technologies with mm-wave band signal distribution. The system has a great potential to support cost-effective and high capacity wireless access. The distribution of radio signals to and from BSs can be either mm-wave modulated optical signals (RF-over-fiber) or lower frequency subcarriers (IF-over-fiber). Signal distribution as RF-over-fiber has the advantage of a simplified BS design but is susceptible to fiber chromatic dispersion that severely limits the transmission distance. In contrast, the effect of fiber chromatic dispersion on the distribution of intermediate-frequency (IF) signals is much less pronounced, although antenna BSs implemented for RoF system incorporating IF-over-fiber transport require additional electronic hardware such as a mm-wave frequency local oscillator (LO) for frequency up- and down conversion. These research activities fueled by rapid developments in both photonic and mm-wave technologies suggest simple BSs based on RoF technologies will be available in the near future. However, while great efforts have been made in the physical layer, little attention has been paid to upper layer architecture.

Specifically, centralized architecture of RoF networks implies the possibility that resource management issues in conventional wireless networks could be efficiently addressed.

II. RADIO OVER FIBER TECHNOLOHY

Radio-over-Fiber (RoF) technology entails the use of optical fiber links to distribute RF signals from a central location (head-end) to Remote Antenna Units (RAUs). In narrowband communication systems and WLANs, RF signal processing functions such as frequency up-conversion, carrier modulation, and multiplexing, are performed at the BS or the RAP, and immediately fed into the antenna. RoF makes it possible to centralize the RF signal processing functions in one shared location (head-end), and then to use optical fiber, which offers low signal loss (0.3 dB/km for 1550 nm, and 0.5 dB/km for 1310 nm wavelengths) to distribute the RF signals to the RAUs, as shown in Figure 1. By so doing, RAUs are simplified significantly, as they only need to perform optoelectronic conversion and amplification functions. The centralization of RF signal processing functions enables equipment sharing, dynamic allocation of resources, and simplified system operation and maintenance. These benefits can translate into major system installation and operational savings, especially in wide-coverage broadband wireless communication systems, where a high density of BS/RAPs is necessary as discussed above.

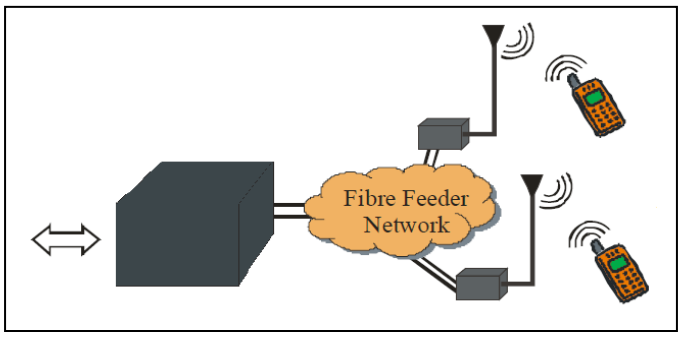


Figure 1: The Radio over Fiber System Concept

One of the pioneer RoF system implementations is depicted in Figure 2. Such a system may be used to distribute GSM signals, for example. The RF signal is used to directly modulate the laser diode in the central site (head-end). The resulting intensity modulated optical signal is then transported over the length of the fiber to the RAU. At the RAU, the transmitted RF signal is recovered by direct detection in the PIN photo detector. The signal is then amplified and radiated by the antenna. The uplink signal from the Mobile Unit (MU) is transported from the RAU to the head-end in the same way. This method of transporting RF signals over the fiber is called Intensity Modulation with Direct Detection (IM-DD), and is the simplest form of the RoF link.

III. BENEFITS OF ROF TECHNOLOGY

Some of the advantages and benefits of the RoF technology compared with electronic signal distribution are given below.

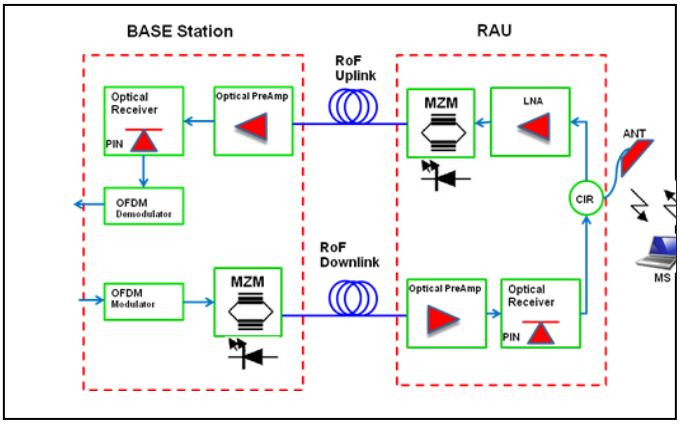


Figure 2: RoF - Basic Structure of the System

A. *Low Attenuation Loss*

Electrical distribution of high frequency microwave signals either in free space or through transmission lines is problematic and costly. In free space, losses due to absorption and reflection increase with frequency. In transmission lines, impedance rises with frequency as well. Therefore, distributing high frequency radio signals electrically over long distances requires expensive regenerating equipment. An alternative solution is to use optical fibers, which offer much lower losses. Commercially available standard Single Mode Fibers (SMFs) made from glass (silica) have attenuation losses below 0.2 dB/km and 0.5 dB/km in the 1.5 nm and the 1.3 nm windows, respectively. These losses are much lower than those encountered in free space propagation and copper wire transmission of high frequency microwaves. Therefore, by transmitting microwaves in the optical form, transmission distances are increased several folds and the required transmission powers reduced greatly.

B. *Large Bandwidth*

Optical fibers offer enormous bandwidth. There are three main transmission windows, which offer low attenuation, namely the 850*nm*, 1310*nm* and 1550*nm* wavelengths. For a single SMF optical fiber, the combined bandwidth of the three windows is in the excess of 50*THz*. However, today's state-of-the-art commercial systems utilize only a fraction of this capacity (1.6 THz). But developments to exploit more optical capacity per single fiber are still continuing. The main driving factors towards unlocking more and more bandwidth out of the optical fiber include the availability of low dispersion (or dispersion shifted) fiber, the Erbium Doped Fiber Amplifier (EDFA) for the 1550*nm* window, and the use of advanced multiplex techniques namely Optical Time Division Multiplexing (OTDM) in combination with Dense Wavelength Division Multiplex (DWDM) techniques.

C. *Immunity to Radio Frequency Interference*

Immunity to electromagnetic interference is a very attractive property of optical fiber communications, especially for microwave transmission. This is so because signals are transmitted in the form of light through the fiber. Because of this immunity, fiber cables are preferred even for short connections at mm-waves.

D. Easy Installation and Maintenance

In RoF systems, complex and expensive equipment is kept at the head-end, thereby making the RAUs simpler. For instance, most RoF techniques eliminate the need for a LO and related equipment at the RAU. In such cases a photo-detector, an RF amplifier, and an antenna make up the RAU. Modulation and switching equipment is kept in the head-end and is shared by several RAUs. This arrangement leads to smaller and lighter RAUs, effectively reducing system installation and maintenance costs. Easy installation and low maintenance costs of RAUs are very important requirements for mm-wave systems, because of the large numbers of the required RAUs.

IV. LIMITATIONS OF ROF TECHNOLOGY

Since RoF involves analogue modulation, and detection of light, it is fundamentally an analogue transmission system. Therefore, signal impairments such as noise and distortion, which are important in analogue communication systems, are important in RoF systems as well. These impairments tend to limit the Noise Figure (NF) and Dynamic Range (DR) of the RoF links. DR is a very important parameter for mobile (cellular) communication systems such as GSM because the power received at the BS from the MUs varies widely. That is, the RF power received from a MU which is close to the BS can be much higher than the RF power received from a MU which is several kilometers away, but within the same cell.

The noise sources in analogue optical fiber links include the laser's Relative Intensity Noise (RIN), the laser's phase noise, the photodiode's shot noise, the amplifier's thermal noise, and the fiber's dispersion. In Single Mode Fiber (SMF) based RoF, systems, chromatic dispersion may limit the fibre link lengths and may also cause phase de-correlation leading to increased RF carrier phase noise. In Multi-Mode Fiber based RoF systems, modal dispersion severely limits the available link bandwidth and distance. It must be stated that although the RoF transmission system itself is analogue, the radio system being distributed need not be analogue as well, but it may be digital (e.g. WLAN, UMTS), using comprehensive multi-level signal modulation formats such as xQAM, or Orthogonal Frequency Division Multiplexing (OFDM).

V. COST EFFECTIVENESS OF ROF TECHNOLOGY

Conventional BS drives the antennas over lossy electrical cable which necessitates the location of the BS very close to the antennas. This can create problems with acquiring suitable sites for coverage extension. It also increases the capital and operational expenses due to site purchasing or leasing, new BS installation and maintenance. Utilizing the idea of RoF and BS hostelling, one BS can control several RAUs and new BS is not required for coverage extension. The additional antennas can be served from the existing BS close to the cell tower. This dramatically reduces the requirements for cell site footprint and the cost of site acquisition. The electrical cables that drive the antenna are responsible for large amount power loss of the BS. The loss in these cables and their associated connectors can range from a typical value of 3db to as much as 10dB in extreme cases which means 50% to 90% of the radio transceiver's output power is dissipated in cable transmission. All this extra power required to drive the electrical feeder cables means that higher output power amplifiers must be deployed. These high power amplifiers are more expensive and have poor operating efficiencies of around 10%, further compounding the problem of high energy consumption by BS. By feeding the RAUs with optical fiber, transmission to the antenna location can be made virtually almost loss-free except some small amount of loss in the short electrical cable connections between the RAUs and the antennas.

In the conventional BS, the power dissipated as heat by the low-efficiency amplifiers requires the BS enclosure to have sophisticated metal enclosures with climate control facilities such as air conditioning, which also increases the expenses. RoF offers large reduction in the amount of thermal energy dissipated by the system. This means that the RAU can be designed without the need for any expensive climate control facilities at the remote site. In addition, the BS hostel can be installed in the more benign environmental conditions of an indoor facility. From the above discussion, it is clear that RoF technology has lots of possibilities to reduce the capital and operational expenses. In order to check the feasibility of transmission of IEEE 802.16a based WiMAX data through optical fiber link, we have done the simulation study.

VI. APPLICATIONS OF ROF TECHNOLOGY

Some of the applications of RoF technology include satellite communications, mobile radio communications, broadband access radio, Multipoint Video Distribution Services (MVDS), Mobile Broadband System (MBS), vehicle communications and control, and wireless LANs over optical networks. The main application areas are briefly discussed below.

A. Cellular Networks

The field of mobile networks is an important application area of RoF technology. The ever-rising number of mobile subscribers coupled with the increasing demand for broadband services have kept sustained pressure on mobile networks to offer increased capacity. Therefore, mobile traffic (GSM or UMTS) can be relayed cost effectively between the SCs and the BSs by exploiting the benefits of SMF technology. Other RoF functionalities such as dynamic capacity allocation offer significant operational benefits to cellular networks.

B. Wireless LANS

As portable devices and computers become more and more powerful as well as widespread, the demand for mobile broadband access to LANs will also be on the increase. This will lead once again, to higher carrier frequencies in the bid to meet the demand for capacity. For instance current wireless LANs operate at the 2.4 GHz ISM bands and offer the maximum capacity of 11 Mbps per carrier (IEEE 802.11b). Next generation broadband wireless LANs are primed to offer up to 54 Mbps per carrier, and will require higher carrier

frequencies in the 5 GHz band. Higher carrier frequencies in turn lead to micro- and pico-cells, and all the difficulties associated with coverage discussed above arise. A cost effective way around this problem is to deploy RoF technology. This greatly simplifies the remote transponders and also leads to efficient base station design.

C. Vehicle Communication and Control

This is another potential application area of RoF technology. Frequencies between 63-64 GHz and 76-77 GHz have already been allocated for this service within Europe. The objective is to provide continuous mobile communication coverage on major roads for the purpose of Intelligent Transport Systems (ITS) such as Road-to-Vehicle Communication (RVC) and Inter-Vehicle Communication (IVC). ITS systems aim to provide traffic information, improve transportation efficiency, reduce burden on drivers, and contribute to the improvement of the environment. In order to achieve the required (extended) coverage of the road network, numerous base stations are required. These can be made simple and of low cost by feeding them through RoF systems, thereby making the complete system cost effective and manageable.

VII. ROF BAASED WiMAX SYSTEM

In this section, some possible RoF deployment scenarios for WiMAX data transmission are proposed as a means for capital and operational expenses reduction. IEEE 802.16a standard based end-to-end physical layer model is simulated including intensity modulated direct detection RoF technology.

Due to the ever-increasing demand of wireless communication and mobility, various wireless communication systems have been developed and deployed. Worldwide Interoperability for Microwave Access (WiMAX) system is now closely examined by many companies for the last mile wireless connectivity to provide flexible broadband services to end users. The technology is based on the IEEE 802.16 and 802.16e standards. According to the WiMAX standard, the cell coverage can typically extend to 5km in the air, with higher data rate and more selectable channel bandwidth than 3G system. Radio-over- fiber (RoF) nowadays is a hot topic for integrating optical technologies with wireless systems. RoF deploys optical fiber, which has low loss and high bandwidth, to distribute radio frequency (RF) signals from central station (CS) or base station (BS) to remote antenna units (RAUs). For some applications, such as inside a long tunnel with many bends, the deployment of the wireless WiMAX is greatly hindered. Because of this, using RoF to carry the WiMAX signal is a good solution.

A. RoF Deployment Scenarios

Rapidly increasing demand for broadband services like high speed internet access and mobile multimedia forcing towards smaller radio cell size. Smaller cells imply that more antennas are needed to cover a certain area. Such an area may include the rooms in a residential home, a hospital, an office building, an airport lounge, or a conference site, etc. When it needs so many antenna sites, it becomes economically attractive to locate the microwave signal generation and modulation at a central BS from where the radio signals will be transmitted to the RAUs using RoF. The antenna units have to do the simple optical-to-electrical conversion, and to emit and receive the wireless signal. Centralizing the sophisticated signal handling process can bring many advantages in operating, maintaining and upgrading wireless networks. In WiMAX service provisioning, several approaches can be taken to utilize the benefits of RoF. Two particular deployment scenarios are given in Fig. 3 and Fig. 4.

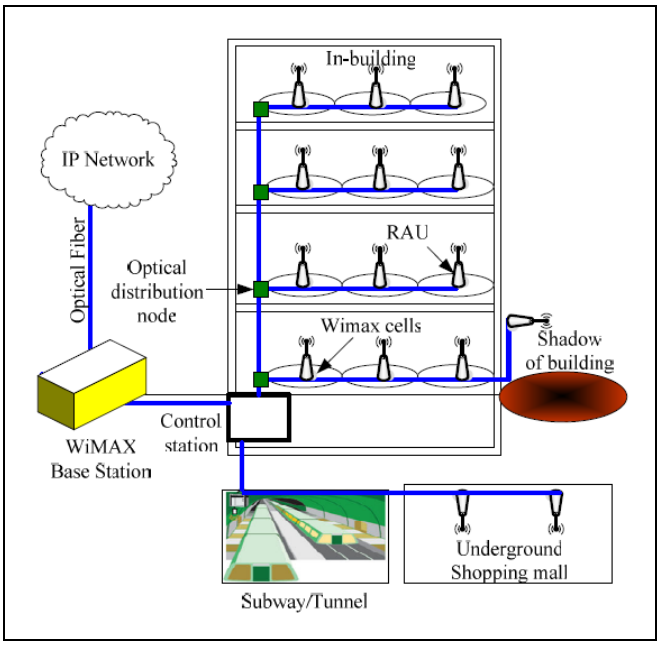


Figure 3: In Building Deployment Scenario of WiMAX over RoF

Fig. 3 shows indoor WiMAX cells, inside the residential buildings, offices, underground subways, tunnels, or shadowed areas, are served by the distributed antenna systems, where the RAUs are fed by WiMAX over fiber links from the WiMAX BS via a control station. BS hostelling can be another cost reducing deployment of RoF as shown in Fig. 4, where multiple macro cells are covered by a central BS and RoF links are used to feed the antenn in each cell. This type of deployment scenario results in lower capital and operational cost for the service providers.

B. Experimental Study Related to WiMAX RoF

In this section two different experiments related to check the feasibility of RoF for WiMAX are described in detail. Some of the good experimental results are summarized at the end of each sub-section.

a) Experiment for TDD Switching Architecture of WiMAX

Using time-division-duplex (TDD) is favored by a majority of implementations in wireless systems because of its advantages of providing flexibility in choosing uplink (UL)-to-downlink (DL) data rate ratios and having less complex transceiver design. However, it is worth to mention that the TDD system limits the transmission distance of systems.

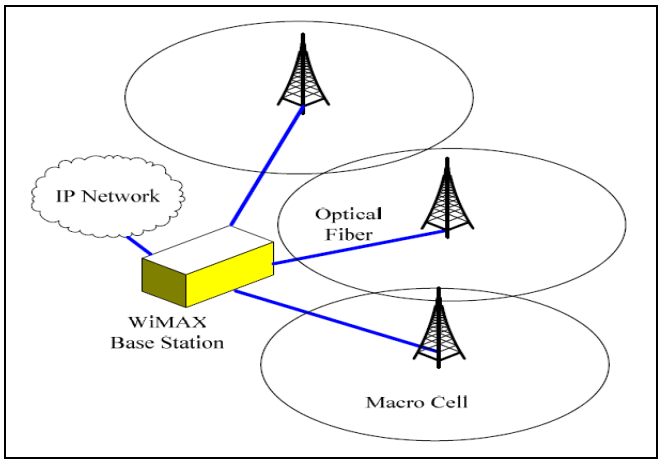
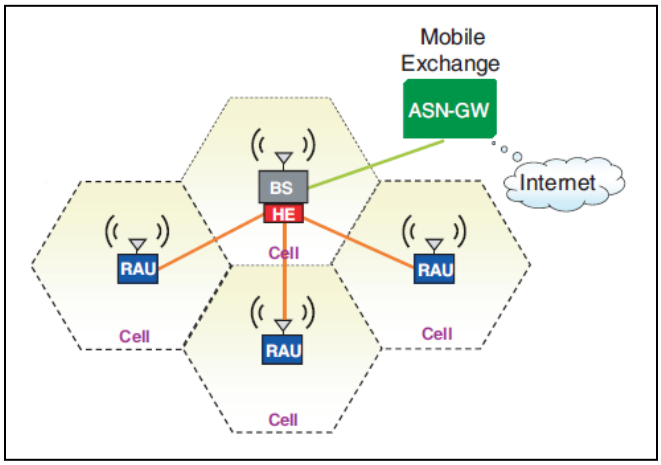


Figure 4: Base station Hosteling

The standard WiMAX signal generated from a commercial base station (BS) is applied to the RoF system. It is observed



that the maximum effective transmission length of the WiMAX RoF system is mainly limited by the synchronization in the TDD mode and not by the signal-to-noise ratio (SNR). TDD switch (SW) architecture is needed at the RAU in the WiMAX RoF link for switching between the UL and DL signals. In order to emit higher power for the RAU, leakage power from the DL may cause damage to the electronic components, such as RF amplifier, in the UL. A robust TDD switch architecture with self-detected switching in each RAUs is studied for the WiMAX RoF system.

The WiMAX signal access network using RoF is shown in Figure 5. In this RoF link, a head-end (HE), which consists of an optical-to-electrical (O/E) and an electrical-to-optical (E/O) modules, is used to connect to the BS. A remote antenna unit (RAU) will be used in each picocell.

Figure 5: WiMAX RoF Access Network

In order to realize the WiMAX RoF system, an experiment is performed. Here, Figure 6(a) shows the proposed WiMAX RoF link connecting a HE and a RAU. The HE and RAU consist of a pair of E/O and O/E converters for conversing electrical and optical signals. To characterize and analyze solely the performance of the WiMAX RoF system and to remove the atmosphere multipath fading effects of the signal, the antenna (ANT) in the RAU and mobile station (MS) are purposely removed. For the reported WiMAX-over-fiber system, the conventional antenna connecting to the base station via the electrical RF cable has been replaced by a pair of O/E-E/O converter and optical fiber. The detection of multipath fading signals in the conventional wireless antenna is the same as that by using the RAU. Since multipath fading issue has been considered in standard wireless WiMAX system, the multipath fading issue is not the main interest in this report. And this is the reason that the setup is simplified by focusing on the performance analysis owing to the optical fiber solely.

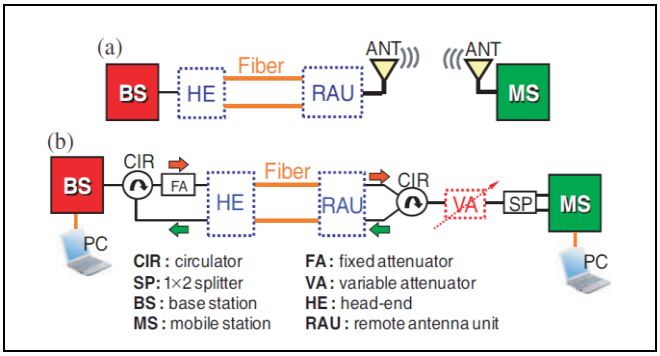


Figure 6: (a) Typical WiMAX RoF Setup
(b) Experimental Setup of Proposed WiMAX RoF Architecture

Hence, the RAU and MS are directly connected using high frequency electrical cables via a RF circulator (CIR), RF variable attenuator (VA) and a RF splitter (SP). Figure 6(b) According to WiMAX standard , for the TDD-based operation, there are two time gaps of transmit/receive transition gap (TTG) and receive/transmit transition gap (RTG) between DL-and-UL and UL- and-DL, respectively, as shown in Figure 7.

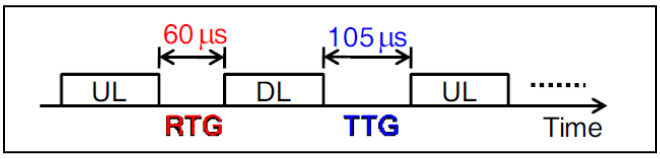


Figure 7: Time Diagram of TDD based WiMAX

The maximum time gaps of TTG and RTG are 105 and 60 µs, respectively, in standard WiMAX system. Initially, the WiMAX signal access was for UL traffic. After ending gap time of RTG, the BS begins to transmit the DL signal. The signal switching can be achieved by using 1 X 2 RF switch (SW) in BS and control by MAC within the gap time of RTG, as shown in Figure 8(a). However, when using the WiMAX access in RoF link, as seen in Figure 8(b), the TDD switch design of the distributed RAU must complete the DL/UL signal switching within the gap times of TTG and RTG. Besides, the maximum WiMAX output power emitted from BS was 35 dBm. Thus, the higher launched power to the ANT at RAU is desirable in order to increase the emitted RF signal power in WiMAX RoF system. Furthermore, due to the intrinsic power isolation of RF circulator, the leakage power from the DL may cause damage to the UL components, such as the low noise amplifier (LNA). Thus, the RF switch design

in RAU must take into account the TDD signal operating and high leakage power.

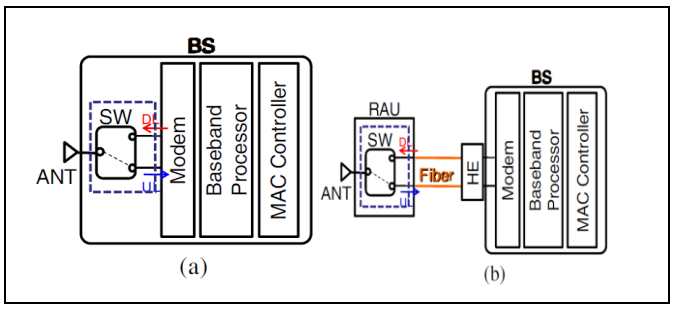


Figure 8: (a) WiMAX TDD switching for DL and UL traffic of BS. (b) TDD switching of RAU in WiMAX RoF architecture

Figure 9 shows the proposed TDD switch in the RAU, which is used to solve the limited power isolation issue of typical high-speed RF device and the TDD operating (used in the experiment). Hence, higher RF power can be launched into the ANT in order to enhance the SNR, while preventing the leakage power may cause damage to the electrical components. The proposed RAU consists of a 1X2 switch (SW1), a 1X1 switch (SW2), a power amplifier (PA), a low noise amplifier (LNA), a delay element (DE), a detector (DT), a control circuit (CC), and a pair transceiver (E/O and O/E converter), as shown in Figure 9. Based on the WiMAX standard, the maximum WiMAX power can be amplified to 35 dBm. To avoid the leakage power of DL signal into the LNA in the UL, two switches: SW1 and SW2 are used to block the leaked DL power. In addition, the proposed TDD switch also needs to consider the signal transmission completely under the gap times of TTG and RTG in fiber link. Moreover, the proposed TDD SW design with self-detection not only avoids the higher leakage power, but also can synchronize the DL and UL data traffic. For IEEE 802.16e WiMAX, the maximum time gaps of TTG and RTG are 105 and 60 µs, respectively. The TTG frame is 105 µs, which equates to approximately 9km roundtrip over standard single mode fiber (SMF). This is the theoretical maximum transmission distance for the WiMAX RoF governed by the WiMAX protocol for waiting the acknowledgement signal. The maximum fiber length may be reduced when the switching or electrical to optical conversion delays are included.

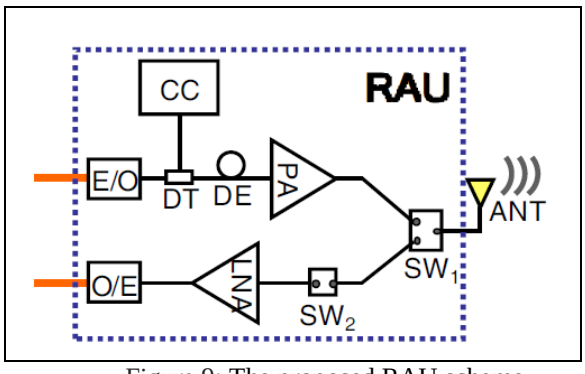


Figure 9: The proposed RAU scheme.

SW1: 1X2 switch; SW2: 1X1 switch; PA: power amplifier; LNA: low noise amplifier; DE: delay element; DT: detector; CC: control circuit; ANT: antenna.

VIII. WIMAX PHYSICAL LAYER SIMULATION MODEL

Fig. 10 below depicts the Physical layer of 802.16a and a classical IMDD optical link model for transmitting the signal to RAU. The laser diode is modulated by the RF signal in the downlink path. The resulting intensity modulated optical signal is then transmitted through the single mode fiber towards a RAU. At the RAU end, the received optical signal is converted to RF signal by direct detection through a PIN photodetector. The signal is then amplified and radiated by the antenna. The Uplink signal is transmitted from the RAU to the BS in a similar way.

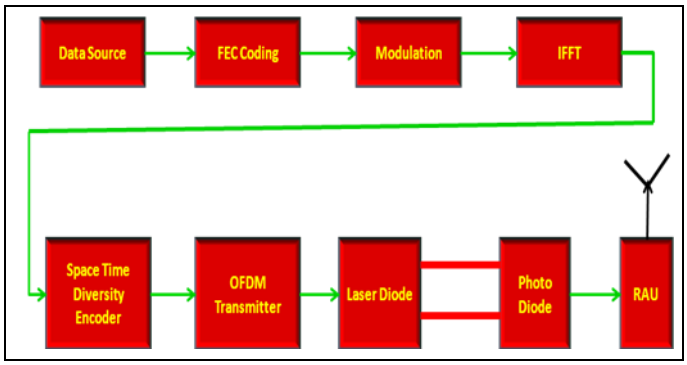


Figure 10: IEEE 802.61a Model with RoF using MATLAB

A. *Laser and Photodiode Mode*

The laser is usually a significant source of noise and distortion in an ROF link, and laser diode normally exhibits nonlinear behavior. When it is driven well above its threshold current, its input/output relationship can be modeled by a Volterra series of order 3 [6]. However, if the signal current dynamic range is within the linear region of the laser diode, it obviously will show linear response. In our present simulation we assume ideal linear characteristic of the laser diode. Output optical power versus current can be given as:

$$P_{opt} = \left(\frac{hf}{e}\right)\eta_L\left(I(t) - I_{th}\right) \quad (1)$$

where $I(t)$ is input current of the microwave signal including the dc bias, I_{th} is diode threshold current, h is Planck constant, f is frequency in hertz, e is charge of an electron, and η_L is laser quantum efficiency [1]. The detection of transmitted light waves is performed primarily by the photo-detector. In most cases the received optical signal is quite weak and thus electronic amplification circuitry is used, following the photodiode, to ensure that an optimized power signal-to-noise (SNR) is achieved. The PIN photodiode and receiver total noise are calculated and superimposed over the ideal photodiode signal current. To evaluate the effect of noise

added during the amplification process, a mathematical model explained in [1] has been used in our simulation. The noise in photodiode includes quantum shot noise i_{sh} , dark current noise i_{dk} , and the thermal noise i_{th} . The total current generated by the photodiode when optical power falls on it is expressed by

$$i_{total} = I_p + \sqrt{\langle i_{noise} \rangle^2} \qquad (2)$$

Where I_p and mean squared noise is given by:

$$I_p = \frac{\eta P_{opt} e}{hf}$$

$$\langle i_{noise}{}^2 \rangle = \langle i_{sh}{}^2 \rangle + \langle i_{dk}{}^2 \rangle + \langle i_{th}{}^2 \rangle$$

And

$$\langle i_{sh}{}^2 \rangle = 2eI_p B$$

$$\langle i_{dk}{}^2 \rangle = 2eI_d B$$

$$\langle i_{th}{}^2 \rangle = \frac{4kTB}{R}$$

Where I_{dk} = 25nA, is assumed to be dark current obtained from the DSC10H PIN photodiode datasheet of Semiconductor. Inc. *B* is the photodiode 3dB bandwidth, *B K* is Boltzmann's constant, *T* is the absolute temperature (ºK), and *R* is the photodiode load resistor assumed to be 50 ohm for ultra wideband receiver.

B. Simulation results

In order to study the feasibility of transmission of WiMAX signals through single mode fiber by IMDD, the simulation was carried out using MATLAB. The model consisted of IEEE 802.16a end-to-end physical layer. More specifically, it modeled the OFDM-based physical layer for downlink, supporting all of the mandatory coding and modulation options. The laser and photodiode are modeled using (1) and (2), respectively.

Fig. 11 show the simulation results for bit error rate vs different SNR values for BPSK without RoF and with RoF. It can be noticed from the figure that the introduction of fiber introduces a high bit error rate because the amplifier is not used at the receiving end. But the introduction of RoF will enhance the coverage area without the need for additional Base Stations, thus reducing the cost of overall deployment. Similar results are obtained for 16QAM with 2/3 coding as shown in Figure 12. These results can be interpreted as same as for the BPSK.

IX. CONCLUSION

Objective of this study was to investigate RoF technology for the transmission of WiMAX signals to the RAUs and hence to suggest feasible RoF deployment scenarios to reduce the capital and operational expenses of the service providers. We studied the performances and limitations of standard WiMAX signal optimized for wireless communication to the commercial RoF system. Results show that the effective RoF transmission fiber length is limited to 8km SMF transmission due to the TDD framing in the connection using standard WiMAX signal. The studeid results imply that if the total length of the WiMAX RoF is 8 km, the distance between the MS and RAU should be very close. Furthermore, the simulation results obtained proved the feasibility of RoF for WiMAX system.

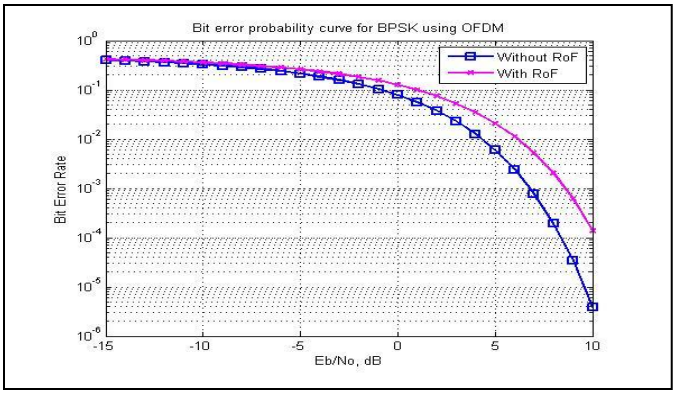


Figure 11: Bit Error Comparison of RoF and non-RoF WiMAX using BPSK

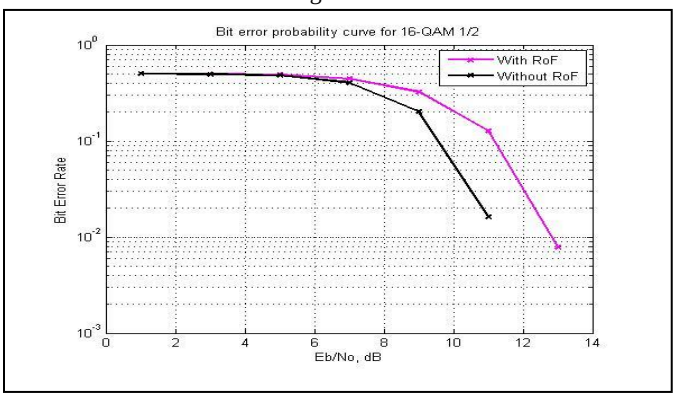


Figure 15: Bit Error Comparison of RoF and non-RoF WiMAX using 16QAM

REFERENCES

[1] Mohammad Shaifur Rahman, Jung Hyun Lee, Youngil Park, and Ki-Doo Kim, *"Radio over Fiber as a Cost Effective Technology for Transmission of WiMAX Signals"*, World Academy of Science, Engineering and Technology, Vol 56, 2009.

[2] C.-H. Yeh, C.-W. Chow, Y.-L. Liu, "Understanding Standard OFDM WiMAX Signal Access in Radio over Fiber System", *Progress In Electromagnetics Research C, Vol. 10, 201{214, 2009*.

[3] Bruce Chow, Ming-Li Yee, Michael Sauer, and Anthony Ng'Oma, *"Radio-over-Fiber Distributed Antenna System for WiMAX Bullet Train Field Trial"*, IEEE Mobile WiMAX Symposium, 2009.

[4] Martin Maier and Navid Ghazisaidi, *"The Audacity of Fiber-Wireless (FiWi) Networks"*, Third International Conference on Access Networks, AccessNets 2008, Las Vegas, NV, USA, October 15-17, 2008.

[5] Marcos D. Katz, Frank H.P. Fitzek, *"WiMAX Evolution Emerging Technologies and Applications"* Times by Sunrise Setting Ltd, Torquay, UK., pp 387-402, ISBN 9780470696804 (H/B), 2009.

City-wide Coordinated Ramp Metering Based on Wireless Sensor Networks

Saad A. Alyousef, Abdullah Al-Dhelaan, and Samir Elmougy
Dept. of Computer Science, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
saalyousef, dhelaan, mougy@ksu.edu.sa

Abstract - *Wireless sensor networks (WSNs) are considered to be one of the useful technologies with extensive range of applications. Ramp metering is considered as an effective way to solve traffic congestion. In this paper, we evaluate some of the current solutions and study their effectiveness. Moreover, we develop a coordinated ramp metering technique that covers a larger area of highways within a city. This system composed of a set of ramps, ramp metering controllers, and a traffic control center components. In order to test and evaluate the proposed algorithm used in the system, a simulation is used on a network of road interchanges (named A, B, C, and D). The simulated results show that for interchange B, the time spent of traffic has been reduced by 34.4% compared to the normal case. 38% traffic flow improvement for Interchange C, and 41.6% improvement for interchange D.*

Keywords: Wireless Sensor Network (WSN) - Ramp Metering - Intelligent Traffic System (ITS) - Congestion control.

1 Introduction

Wireless sensor networks (WSNs) is considered to be one of the useful technologies with extensive range of applications serving many areas such as environment monitoring, medical and health systems, military, home applications, etc. [1]. Vehicular Ad Hoc Network (VANET) is a new technology integrating ad hoc network, wireless LAN (WLAN) and cellular technology to achieve intelligent inter-vehicle communications. This technology also has many applications that serve different fields. One of the major applications is of VANETs is Intelligent Transportation Systems that includes various application such as monitoring of roads traffic, traffic flow control, etc. [2].

A wireless sensor network is generally consisted of an enormous number of sensor nodes. These sensor nodes have the capabilities of sensing a particular event, they can communicate with each other by transmitting data, and they can even perform simple data processing. There are many different types of sensing depending on the application. For instance, there are temperature sensors that can read the current temperature that usually used for forecasting applications. There are also other types of sensors that can detect different conditions such as humidity, noise, lighting condition, vehicular movement, etc. [3].

Nowadays, cars are considered one of the most useful transportation. Due to long distances between cities and within cities themselves, and with the low level of awareness of traffic regulations (rules), many car accidents could be happened. Generally, about 1.3 million people die in road crashes yearly, on average 3,287 deaths every day. Furthermore, there are about 20-50 million are injured or disabled [4]. As consequences of cars accidents, occurring of traffic congestion causes waste of travelling time and money. As a solution for traffic congestion, many research methods were proposed such as speed limits, route guidance, ramp metering, reversible lanes, etc. [5].

Ramp metering can be defined as traffic lights that used into ramps to control the frequency with which vehicles enter the flow of traffic on the freeway [6]. Ramp metering is considered as an effective way that used in freeway traffic system to solve traffic congestion [7]. Many researchers have discussed this issue and proposed different algorithms. Ramp metering algorithms can be classified into two main categories: (1) fixed time metering, and (2) traffic-responsive metering [8]. The second category is considered more effective and efficient in term of congestion problem. Ramp metering also can be categorized into local and coordinated types (strategies). Local ramp metering strategies work only on one ramp individually and unable to harness integrated consideration over the freeway as a whole. Because of the limitation of local ramp metering strategies, researchers proposed various coordinated techniques.

In this paper, we evaluate some of the current solutions and study their effectiveness and disadvantages to design a better application. Moreover, we develop a coordinated ramp metering technique covering a larger area of highways within a city in which wireless sensors is used to collect data about traffic flow (i.e. counting vehicles) and then send it to traffic control center to process and to take the appropriate actions.

The rest of this paper is organized as follows: In Section 2, some of relevant previous work is introduced. Section 3 presents the proposed system architecture and details about

each component. The network architecture of sensor nodes and data transmission are provided in Section 4. Section 5 presents the proposed algorithms including its requirements and the responsible component of each algorithm. Simulation of a coordinated ramp metering is presented in Section 6 to demonstrate the effectiveness of the proposed algorithm by comparing it with the case of no system is applied. Conclusion and future work are discussed in Section 7.

2 Related Work

Traffic congestion has become a serious problem nowadays due to the rapid increase in the number of vehicles. Traffic congestion happens when too many vehicles attempt to use a common transportation infrastructure and exceed its capacity. Traffic congestion leads to a degraded use of the available infrastructure. As consequences of traffic congestion, it reduces safety, increases environmental pollution, and results in extra delays. For these reasons, many researchers proposed different methods to help reducing traffic congestion such as speed limits, route guidance, ramp metering, and reversible lanes. Ramp metering is considered as the most direct and efficient solution to control traffic congestion. Ramp metering provides many advantages regarding traffic infrastructure. It increases the throughput of freeways because it reduces congestion. Moreover, it improves traffic safety due to reduction and safer merging (when getting into a freeway) [8].

Ramp metering algorithms can be classified into two categories: (1) fixed time metering, and (2) traffic-responsive metering [8] [9]. In the first category, ramps are blocked for a specified time of day (e.g., during rush hour). The period of blocking is constant and determined based on historical demands of a given road. The second type is more efficient in solving traffic congestion since it takes into account measurements based on calculations of real time data. Typical algorithms under this category include ALINEA [10], demand-capacity, occupancy algorithms [11], and linear quadratic regulation (LQR) [12]. Although these algorithms achieve a magnificent success in solving the problem of ramp metering, they are purely local. In other words, they work on a single ramp and deal with each ramp independently so that they are unable to attach integrated consideration over all ramps on freeway as a whole. Figure 1 [10] illustrates the local strategies process, where O_{out} (O_{in}): measured occupancy rate downstream (upstream), Q_{out} (Q_{in}): measured traffic volume downstream (upstream), r: on-ramp traffic volume, and δ: downstream bottleneck capacity.

Researchers proposed various coordinated techniques because of the limitation of local ramp metering strategies. Coordinated algorithms can deal with numerous ramps on the entire mainline. One of the first algorithms of this category was the SCOOT [13] proposed by Hunt, et al. There are also some useful ideas with efficient algorithms: the simple Helper ramp algorithm, the Bottleneck algorithm and SWARM [14], [15] [16]. The coordinated ramp metering problem is shown in Figure 2, where a freeway lane is divided into sections, and each section contains at most one off-ramp and one on-ramp.

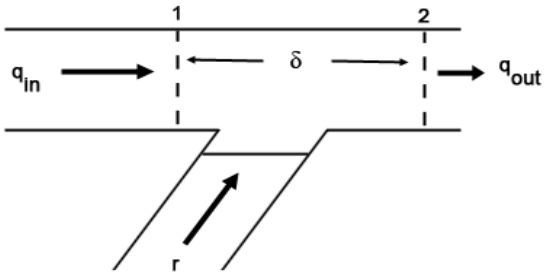


Fig. 1. Traffic flow process of local strategies

Ramp metering algorithms categories with some common algorithms examples are demonstrated in Figure 3 [8] [9].

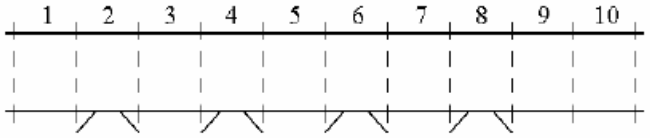


Fig. 2. A freeway lane

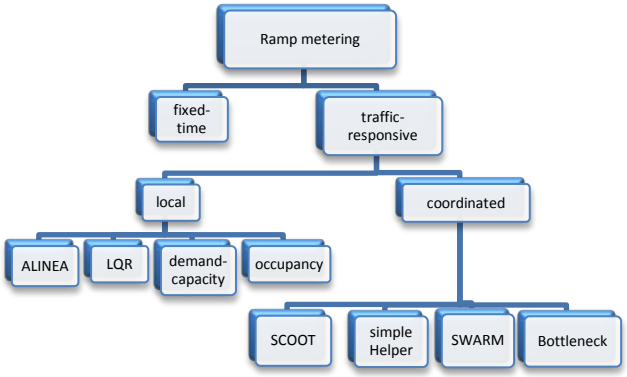


Fig. 3. Ramp metering algorithms categories and common examples of these categories

Although the techniques mentioned above provide a good result based on the used simulations in term of coordinating with multiple ramps, the ramp metering problem from WSNs perspective was not covered. Usually researchers who used those techniques were assuming that a freeway lane is divided into sections with at most one off-ramp and on-ramp on each section. To work with this assumption, the traffic flow should be computed which require to complete the traffic flow existence of sensor nodes at the beginning of each section. In other words, a lot of sensor nodes are needed. They did not devise how these sensors will communicate and only considered one type of ramps that simply turns right.

There are more complicated ramps that are needed to be considered such as ramps that are located on road interchanges where a set of ramps leading to other highways which connects all highways forming a network of roads infrastructure. In our

work, we consider the limitations of the previous work mentioned above. First, we propose a system architecture that has a higher level coordination which makes the coordination much easier and a specific job for each system components. Secondly, show how to collect real time data about traffic flow. We use sensor nodes that are able to accomplish this job whereas all previous work mentioned above had never been discussed. Finally, we discuss how these wireless sensors are communicated to send its collected data.

3 The Proposed System Architecture

In this work, we develop a coordinated ramp metering technique covering a larger area of freeways within a city. Wireless magnetic sensors is used to collect data about traffic flow (i.e. counting number of cars) [17] [18] and then data is sent to traffic control center to process these data and to take the appropriate action. Chinrungrueng and Kaewkamnerd [19] have developed a new traffic data collecting device based on magneto-resistive technology that can detect metal objects. A magnetic field is generated when a vehicle run on that sensor device. The next step is to process the generated signal by the device. Finally, the output of this process could be vehicle count, speed, occupancy time, and classification, depending on the applied process. The device also is able to communicate with base station via radio frequency. It has been proven that the accuracy of counting vehicles using magnetic sensors is very high [20] [21].

A set of these sensors can be placed in a specific location on the road. Consequently, real-time counting of vehicles is obtained. We can take the benefits from these devices to be installed on freeway ramps and on the freeway itself near to the ramp in order to reduce any possible traffic congestion. The collected data from different ramps is then sent to the traffic control center to be processed and to reply back to take an appropriate action. Each on-ramp have a traffic light that has two lights (green to pass, red to stop) that controlled by ramp metering controller.

There are three main components of the proposed system: (1) A set of ramps, equipped with sensor nodes to monitor traffic flow. (2) Ramp Metering Controllers, one ramp controller for each ramp to control ramp metering lights of its ramp. (3) Traffic Control Center (TCC) to receive data from ramp metering controllers, process these data and to reply back to the corresponding Ramp metering controller ordering it to take the appropriate action. Figure 4 illustrates the main architecture with its components of the proposed system, where RC_i is a ramp metering controller installed into a road interchange R_i , where $i \in \{1, 2, 3, \ldots, N\}$

3.1 Ramps

A set of ramps spreads all over a city are equipped with some sensors nodes to monitor traffic flow. There are different ramps infrastructures based on the nature of location, type of intersected roads, and whether there is a need to place a ramp, etc. When there is an intersection between two highways, typically eight ramps are required to fully serve all vehicles using these highways coming from different ways to take the other road more easily. Figure 5 illustrates a possible infrastructure of road ramps of two intersected highways (Cloverleaf Interchange, see Figure 6 for other types).

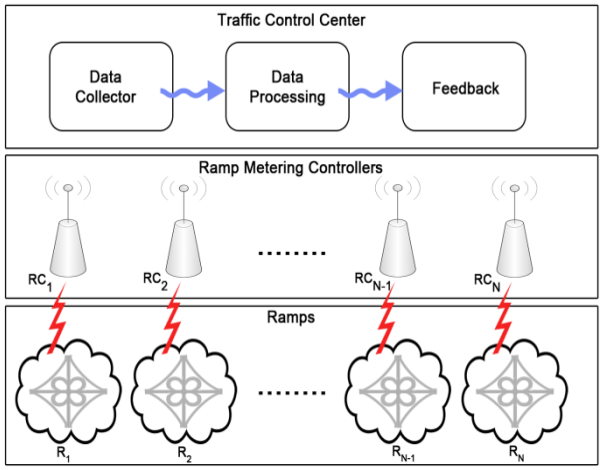


Fig. 4. The main architecture of the proposed system

Fig. 5. A common ramps infrastructure of two intersected highways (Cloverleaf interchange)

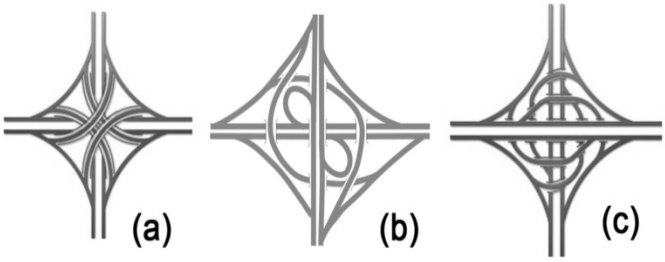


Fig. 6. Other types of road interchange: (a) Stack interchange (b) Cloverstack interchange (c) Turbine interchange

Figure 7 illustrates the required components for road interchange of type Cloverleaf. A set of sensors nodes are needed on the entrance of each ramp and before the merge area between off-ramp and highway in order to fully monitor traffic flow of all of highways and ramps (Figure 7 shows the location of each sensor). In this type of road interchange, there are eight ramps; each ramp has a traffic metering light at the end of it. Finally, a Ramp Metering Controller (sink) is needed. All sensor nodes can communicate directly to this

component to send what they have collected as shown in Figure 7. The power for sensor nodes can be supplied by either a direct electricity source or by solar cells attached to all sensor nodes

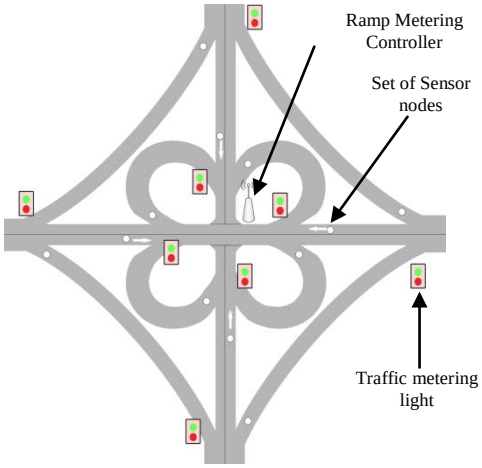


Fig. 7. Components of each road interchange

The roads generally are divided into lanes. Which means that specify precisely the location of each sensor node is important. Each vehicle can move on any road lane. In order to detect all vehicles more precisely, the best location for those sensor nodes is to deploy them on the middle of each lane with assumption that no vehicle will move between two lanes (i.e. moving on the cat's eyes). As mentioned early, sensor nodes are needed before the merge area between highroad and off-ramp, and on the beginning of each on-ramp. This number of needed sensor nodes depends on the number of available lanes. For instance, consider a road interchange with three lanes for highway, and two lanes for all ramps. In this situation, three sensor nodes are needed before each merge area, and two sensor nodes for each on-ramp as illustrated in Figure 8.

3.2 Ramp Metering Controllers

For each road interchange, a ramp metering controller is needed which is responsible for the following functions:

- Collecting data about traffic flow from all sensors nodes installed on that road interchange.
- Sending the collected data to the TCC to be processed.
- Controlling the ramp metering lights of the road interchange based on the decision that has been received from TCC.

Also, there are four road directions for each road interchange. In order to collect data correctly, Ramp Metering Controller must be able to distinguish between these directions and sensor nodes that affect them. As a solution, sensor nodes are categorized into groups and each group will have a unique identifier. Each group is assigned to the affected direction.

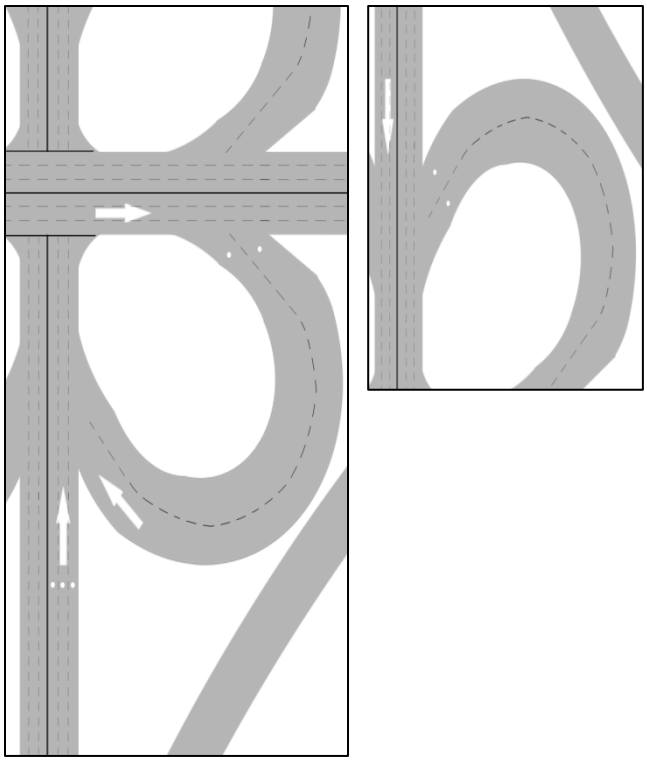


Fig. 8. The needed number of sensor nodes before the merge area (left) and on on-ramp entrance (right)

It has been observed that a vehicle can follow a specific direction using three ways: vehicles currently using the specified direction, vehicles turned right from on-ramp, and vehicles passed from looped on-ramp. For this reason, each group includes three sets of sensor nodes. All sensor nodes influence the current traffic data of the specified direction by an increase in vehicles except the ones installed in the looped on-ramps. These sensor nodes affect the previous direction by a decrease in traffic and it increases the new followed direction traffic. For instance, if vehicle was in direction d_1 and turned to d_2 using the looped on-ramp, then, the traffic of d_1 is decreased by one. And d_2 traffic is increased by one. Figure 9 illustrates the groups of sensor nodes for each direction and their locations on the road interchange.

3.3 Traffic Control Center (TCC)

This component works as a coordinator between all Ramp Metering Controllers. It has three main subsystems:

- Data Collector: responsible for collecting data from various Ramp Metering Controllers.
- Data Processing: processing the sent data based on the road interchanges that affect each other in term of traffic control.
- Feedback: the processed data are received by feedback subsystem and then reply back to the corresponding Ramp Metering Controller with the instructions to be applied by Ramp Metering Controller.

4 Network Architecture

Each road interchange has a set of sensors and one Ramp Metering Controller as mentioned in system architecture. Ramp Metering Controller is considered as a sink of the network that works as a base station of all other sensors in the network. To send data to the sink, each sensor node use a single-hop long-distance transmission, which leads to the single-hop network architecture, as shown in Figure 10.

The sink broadcasts queries every time interval t to the sensor nodes while all sensor nodes sense the traffic flow and sends the sensed data to the sink after receiving the query. The sink also works as a gateway to TCC through the Internet. After it collects data from the sensor nodes, it performs simple processing on the collected data, and then sends the processed data via the Internet to the TCC.

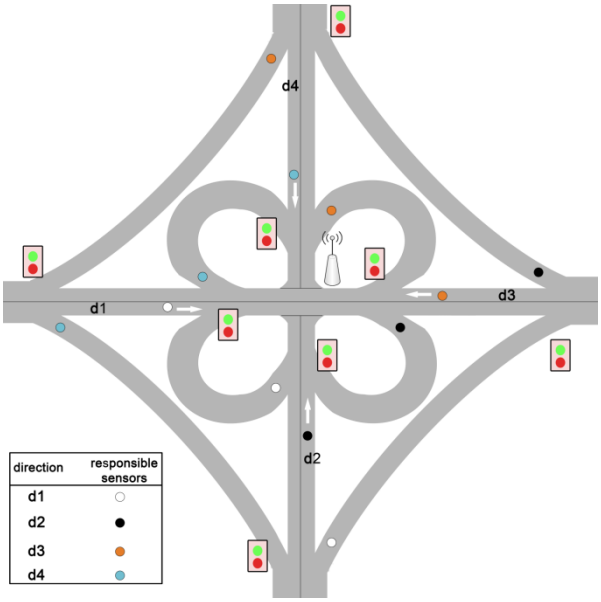


Fig. 9. Directions of a road interchange and sensor nodes responsible for each direction

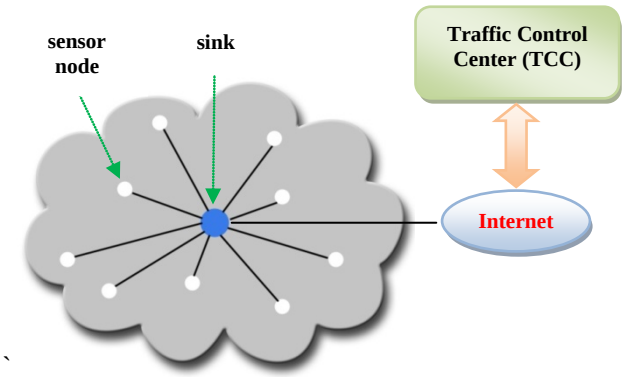


Fig. 10. Single-hop communication to the base station

5 The Proposed Algorithm

5.1 Requirements

Before the system can work, the relations between the different roads interchanges should be known. This can be done by constructing a directed graph $G = (V, E)$, where V is a set of vertices, E is a set of edges that connects two different vertices. Each vertex represents a road interchange and each edge represents a road. The number of vertices in G is equivalent to $|V|$, and the number of edges is equivalent to $|E|$, where $|V|$ is the number of roads interchanges that the system have installed on them, $|E|$ is the number of highways that links two road interchanges from $|V|$. The following requirements are also needed:

- The capacity of each road merge area within the network also should be known.
- Ramps with ramp metering controllers may lead to a longer queue of vehicles which may cause traffic congestion. So, fairness between ramps and highway should be considered. Sensors located on the entrance of each ramp can detect any possible congested ramps to report Ramp Metering Controller.
- The Traffic Control Center must be able to coordinate between different Ramp Metering Controllers in order to reduce any possible traffic congestion.
- The feedback sent from Traffic Control Center to Ramp Metering Controllers should have enough information so that the Ramp Metering Controller has the knowledge about which ramp metering light to control.
- Each sensor node must have a unique identifier on road interchange level in order to specify the roads affected by the installed sensor node.

5.2 Computing Ramp Metering Period

Consider $r = \{1, 2, 3, 4, .., n\}$, where r is the interchange number, and n is the number of interchanges. For each interchange, Equation (1) should be computed for each road direction, where max is a constant, S_r is the sensor node collected data, and C_r is the merge area capacity.

$$I_r = \frac{max \times S_r}{C_r} \qquad (1)$$

In order to find out the time period for blocking the ramp, Equation 2 is computed for each road interchange relation, where a and b are two neighboring, $D_{a,b}$ is the distance between a and b in kilometers, and avg_{speed} is the average speed of vehicles to get from the interchange a to b. Equation 2 represents the time period in seconds to block a specific ramp.

$$F_a = \begin{cases} I_a + I_b - \dfrac{D_{a,b}}{avg_{speed}} \times 60 \;\; , \;\; I_b \geq \dfrac{D_{a,b}}{avg_{speed}} \times 60 \\ I_a \;\; , \;\; I_b < \dfrac{D_{a,b}}{avg_{speed}} \times 60 \end{cases}$$
$$\text{where } a, b \in r,\ a \neq b \qquad (2)$$

Equations (1) and (2) are computed periodically every time at interval t (frequency interval) where $t \geq 2 \times \max$.

5.3 Road Interchange Algorithm (Local)

Algorithm 1 shows the algorithm of Ramp Metering Controller that is used to broadcast a query to all sensors every time interval t (frequency interval), receive all data collected by them, send these data to Traffic Control Center, and finally wait for the feedback to be executed. In this algorithm, S is the sensor nodes, Sink_ID is the IP of this controller, and TCC is Traffic Control Center IP. The purpose of procedure Append is that to append the id of Ramp Metering Controller with its data so that TCC can distinguish between each Ramp Metering Controller data. The procedure Execute is responsible for blocking traffic lights for the given time period.

ALGORITHM 1. RAMP METERING CONTROLLER ALGORITHM

Algorithm 1: Ramp Metering Controller

```
1  REPEAT
2      Wait for time interval t
3      Broadcast_Query(S)
4      DataSet := Receive_Result(S)
5      //dataset has 4 sets of data one for //each
6      direction
7      Append(Sink_ID, DataSet)
8      Send(DataSet, TCC)
9      Action := Wait_For_Insructions(TCC)
10     FOR EACH Ramp in Action
11         Execute(Ramp, Ramp.period)//block all
12         ramps leading to this road for computed
13         period
14     END
15 UNTIL ∞
```

The task of each sensor node is simply counting all vehicles passing over it and whenever it receives a request from Ramp Metering Controller, it will reply with the collected data and start over again as demonstrated in Algorithm 2.

5.4 Coordinated Roads Interchanges Algorithm (Global)

This algorithm is applied by the Traffic Control Center. Firstly, it collects data from all of Ramp Metering Controllers. Secondly, each Ramp Metering Controller data is processed to reduce a possible congestion by applying Equation (1) and Equation (2). Finally, the results of data processing are replied back to each Ramp Metering Controllers. Ramp Metering Controller blocks its ramps that lead to a specific road for the time period that is received by Traffic Control Center. Algorithm 3 illustrates how Traffic Control Center processes the data.

ALGORITHM 2. SENSOR NODE ALGORITHM

Algorithm 2: Sensor Node

```
1  Car_Count := 0
2  REPEAT
3        IF an Object passed over
4          Car_Count := Car_Count+1
5        IF(Query_Request()) THEN
6          Reply(Car_Count, Sink_ID)
7          GOTO 1
8        END IF
9      UNTIL ∞
```

ALGORITHM 3. TRAFFIC CONTROL CENTER ALGORITHM

Algorithm 3: Traffic Control Center

```
1  REPEAT
2      AllData := Receive_Requests()
3      FOR EACH interchange IN AllData
4        Sink_ID := getHeader(interchange)
5        FOR EACH sensorData IN interchange //4
6        directions
7          I_interchange = max × S_interchange/C_interchange
8        END
9      END
10     FOR EACH interchange IN AllData
11      Sink_ID := getHeader(interchange)
12      Neighbors :=FindNeighbors(Sink_ID)
13      FOR EACH N IN Neighbors
14        IF I_N ≥ D_interchange,N /avg_speed × 60
15            F_interchange = I_interchange + I_N
16            − D_interchange,N /avg_speed × 60
17        ElSE
18            F_interchange = I_interchange
19        END IF
20        Result.insert(F_interchange ) //insert result into
21        Result
22      END
23      Send(Sink_ID,Result)
24      Result.clear() //empty result
25    END
26 UNTIL ∞
```

6 Simulation

In order to test the proposed algorithm, a simulation was made on a network of highways. Since the available traffic simulators capabilities are insufficient for this study, a traffic simulator has been implemented that is able to count vehicles

in each road interchange. In addition, the collected data can be gathered by TCC and therefore our proposed algorithms could be applied (see Chapter V for further details). The developed application accepts several parameters of road properties such as road distance, current traffic, the capacity of each road merge area, etc.

This highway network consists of four interchanges (namely, A, B, C, and D) that connect four highways (H1, H2, H3, and H4). The distance of highways H1, H2, H3, and H4 are 20 km, 25 km, 15 km, and 30 km, respectively. For the purposes of this study, we assume that the simulation is a clockwise direction. Each Interchange will receive a traffic flow from two roads and merged into one road. Interchange B will receive 5000 vehicles from H1 and 3000 vehicles from H2 merged into H2. For interchange C, 2500 vehicles are located in H2 and 2500 in H3 will be merged into H3. Finally 4000 vehicles are coming from H3 and 4600 from H4 will be merged into H4. The highway network is illustrated in Figure 11. The used parameters of the proposed equations are: $\max = 10$, $t = 20$, $\text{avg}_{\text{speed}} = 100$ km.

We have figured out that the chosen values gives a better result in term of reducing traffic congestion since giving higher values for max and t will results traffic congestion because of the long period of blocking ramps.

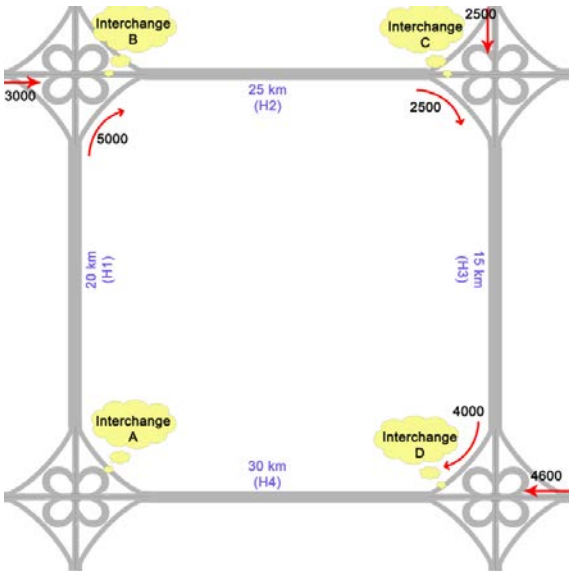


Fig. 11. The simulation of a highway network

6.1 Normal Case

In normal case, the highway network configurations previously discussed have been applied without coordination. It has been noticed that when traffic congestion is happen, the traffic flow falls sharply in interchanges *B*, *C*, and *D*. Afterward, traffic flow undulated until the end.

It has been observed that the traffic congestion suddenly happens when the number of vehicles exceeds the capacity of the merge area. For instance when looking at Interchange B in Figure 12 on the 30th second, the traffic flow reached the highest value which is about 900 then has decreased suddenly to 350 vehicles because it has exceeded the merge area capacity. The same thing applies to Interchanges *C* and *D*. The total time spent to pass all vehicles through Interchanges *B*, *C*, and *D* is 622 seconds, 984 seconds, and 1706 seconds, respectively. Information details of Interchanges *B*, *C*, and *D* are illustrated in Figures 12, 13, and 14.

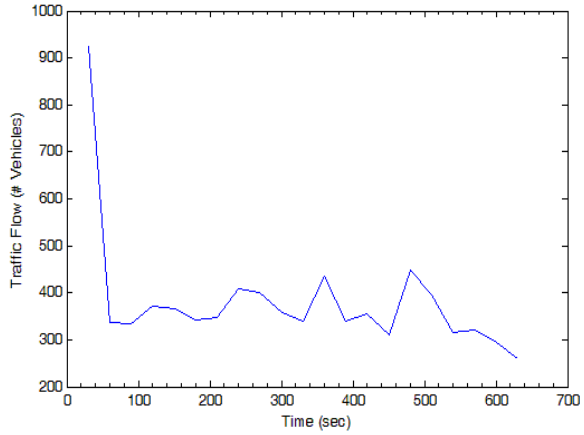


Fig. 12. Total time spent for interchange B in normal case

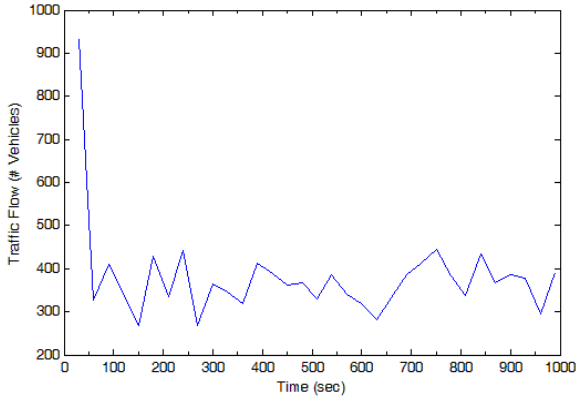


Fig. 13. Total time spent for interchange C in normal case.

6.2 Coordinated Case

The coordinated algorithm has been applied on the same highway network configurations. In the beginning, the coordinated algorithm tries to keep the traffic flow higher than 1000 for Interchanges B, C, and D. After that, all interchanges kept maintaining the high traffic flow. However, traffic congestion in all road interchanges started appearing. In interchanges B and C (Figures 15 and 16), the traffic flow started to fall at the 100th second until it reached to 500 vehicles at the 150th second. For interchange D in Figure 17 started to fall after approximately the 200th second. Afterward, the traffic flow flattened out with steady increase and decrease.

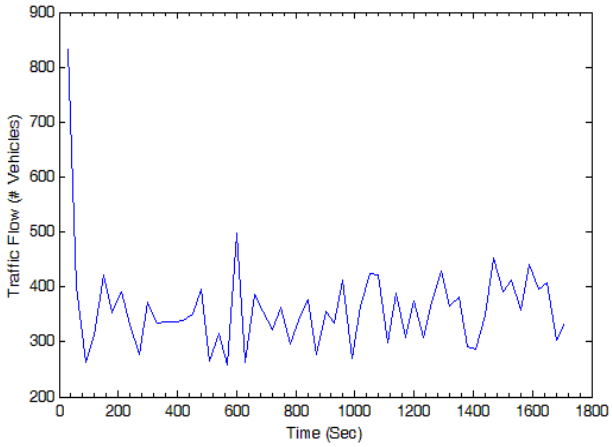


Fig. 14. Total time spent for interchange D in normal case

Although the traffic congestion happened in the coordinated case, but when comparing it with the traffic congestion in case of normal case, it can be observed that the congestion is reduced. Even after the traffic congestion happened, the traffic flow was much higher than the normal case. The coordinated control strategy helped in increasing the traffic flow. The resulting total time spent is 408 seconds for interchange *B*, 610 seconds for Interchange *C*, and 997 seconds for Interchange *D* (Figures 15, 16, and 17 respectively).

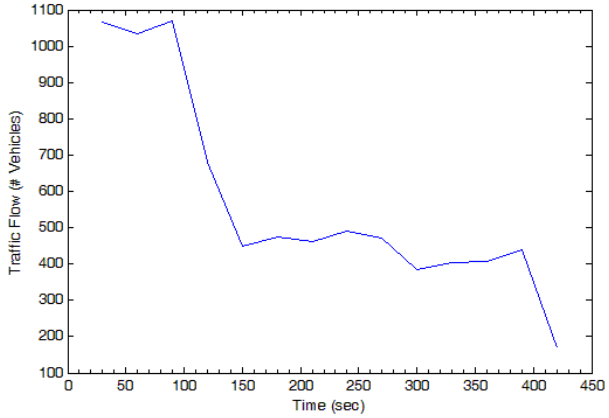


Fig. 15. Total time spent for interchange B using coordinated algorithm

6.3 Results

It has been observed that there is an improvement in traffic flow when applying the coordinated algorithm compared with the normal case. We believe that there is an improvement because the coordinated algorithm reduces the pressure in the merge area which leads to higher vehicles passing. Consequently, the time required to pass all vehicles is reduced.

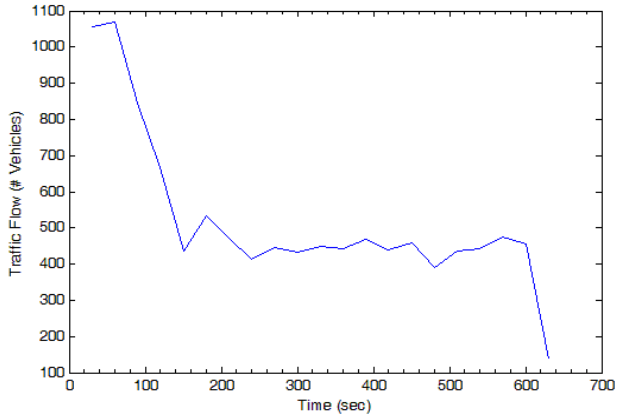


Fig. 16. Total time spent for interchange C using coordinated algorithm

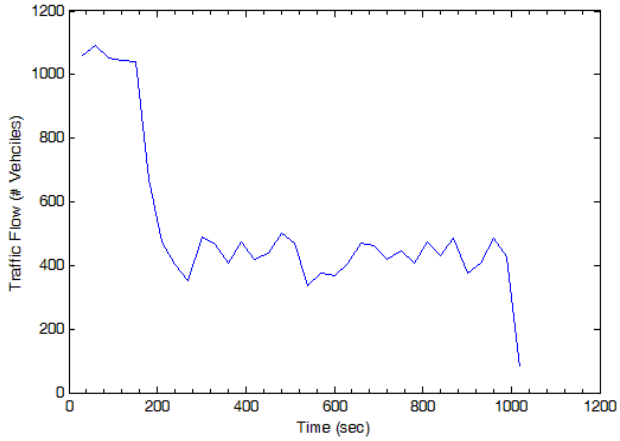


Fig. 17. Total time spent for interchange D using coordinated algorithm

For interchange *B*, the time spent of traffic has been reduced by 34.4% compared to the normal case. 38% traffic flow improvement for Interchange *C*, and 41.6% improvement for interchange *D*. A Comparison between the total time spent for each interchange of normal state and with coordinated algorithm is illustrated in Figure 18 and flow rate improvements are shown in Table 1.

TABLE 1. Flow rate improvement when applying ramp metering

	Road Interchanges		
	B	**C**	**D**
Normal case (no-control)	622	984	1706
Coordinated algorithm	408	610	997
Improvement %	**34.4%**	**38%**	**41.6%**

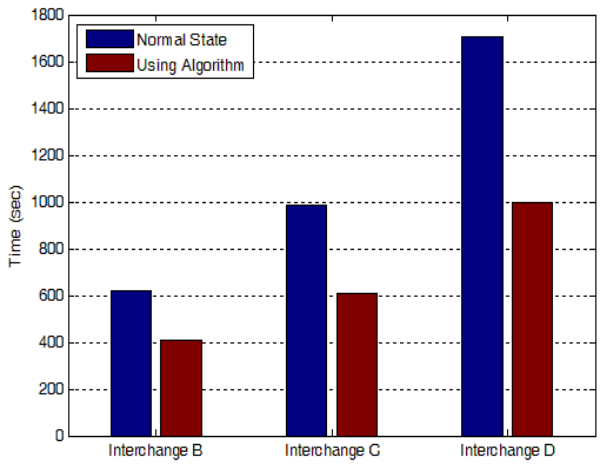


Fig. 18. Comparison between the total time spent for each interchange of normal state and with coordinated algorithm

7 Conclusion and Future Work

Traffic congestion is one of the most problems that cause an increase in travelling time, fuel waste, increase air pollution and the chance of cars accidents. As a solution for traffic congestion, many researchers have proposed different methods including speed limits, route guidance, ramp metering, reversible lanes, etc. Ramp metering can be defined as traffic lights that used into ramps that control the rate of vehicles flow in order to reduce a possible congestion on the freeway. In this work, we have proposed a coordinated ramp metering technique covering a larger area of highways within a city. Wireless sensor nodes have been used to collect data about traffic flow and to be sent to traffic control center for further processing. A simulation has been applied on the proposed algorithm. According to the simulation, the algorithm achieved a good result compared to no-control case (i.e. without using a system) and has reduced travelling time by approximately 38% on average. This system could be combined in future to other intelligent traffic systems such as dynamic traffic lights, variable speed limits, incident detection, etc. for wider traffic networks. More investigations are required to show and evaluate how to integrate these systems together.

8 References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," Elsevier Computer Networks, vol. 38, no. 4, pp. 393-422, 2002.

[2] F. Li and Y. Wang, "Routing in Vehicular Ad Hoc Networks: Aa Survey," IEEE Vehicular Technology Magazine, vol. 2, no. 2, pp. 12-22, June 2007.

[3] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, "Next century challenges: scalable coordination in sensor networks," in MobiCom'99, Washington, 1999, pp. 263-270.

[4] (2012) Association for Safe International Road Travel. [Online]. http://www.asirt.org/KnowBeforeYouGo/RoadSafetyFacts/RoadCrashStatistics/tabid/213/Default.aspx

[5] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: an overview," IEEE Transaction on Intelligent Transportation Systems, vol. 3, no. 4, pp. 271-281, 2002.

[6] (2012) Washington State, Department of Transportation. [Online]. http://www.wsdot.wa.gov/Traffic/Congestion/rampmeters

[7] J. Robinson and M. Doctor, "Ramp metering status in north America, final report," Office of Traffic Operations, Federal Highway Administration, U.S. Department of Transportation, Washington, DC, 1989.

[8] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," Proc. IEEE, vol. 91, no. 12, pp. 2043-2067, Dec. 2003.

[9] I. Papamichail and M. Papageorgiou, "Traffic-Responsive Linked Ramp-Metering Control," IEEE, vol. 9, no. 1, pp. 111-121, March 2008.

[10] M. Papageorgiou, H. Hadj-Salem, and J. M. Blosseville, "ALINEA: a local freeback control law for on-ramp metering; a real-life study," in Third International Conference on Road Traffic Control, London, 1990, pp. 194-198.

[11] R. Wilshire, R. Black, R. Grochoske, and J. Higonbotham, "ITE traffic control systems handbook," Institute of Transportation Engineers, Washington, DC, 1985.

[12] N. B. Golstein and K. S. P. Kumar, "A decentralized control strategy for freeway regulation," Tran. Research Part B, vol. 16, pp. 279-290, 1982.

[13] P. B. Hunt, D. L. Robertson, and R. D. Bretherton, "The SCOOT on-line traffic signal optimization technique," Traffic Eng. Control, vol. 23, pp. 190-192, 1982.

[14] M. Zhang et al., "Evaluation of On-ramp Control Algorithms," Institute of Transportation Studies, One Shields Avenue, University of California, Davis, CA 95616, 2001.

[15] L. Jacobsen, K. Henry, and O. Mahyar, Real-Time Metering Algorithm for Centralized Control. Washington State, United States: Washington State Transportation Center, 1989.

[16] C-H. Wei, "Applying an Artificial Neural Network Model to Freeway Ramp Metering Control," Transportation Planning Journal, vol. 25, no. 3, pp. 335-355, 1996.

[17] A. Sharma, "Applications of Wireless Sensor Network in Intelligent Traffic System: A Review," in 3rd International Conference on Electronics Computer Technology (ICECT), Kanyakumari, 2011, pp. 53-57.

[18] A. Haoui, R. Kavaler, and P. Varaiya, "Wireless magnetic sensors for traffic surveillance," Elsevier Transportation Research Part C: Emerging Technologies, vol. 16, no. 3, pp. 294–306, 2008.

[19] J. Chinrungrueng and S. Kaewkamnerd, "Wireless Magnetic Sensor Network for Collecting Vehicle Data," in IEEE SENSORS 2009 Conference, Christchurch, 2009, pp. 1792-1795.

[20] S. Y. Cheung et al., "Traffic Measurement and Vehicle Classification with a Single Magnetic Sensor," California Partners for Advanced Transit and Highways (PATH), Institute of Transportation Studies (UCB), UC Berkeley, 2004.

[21] A. Goel, S. Ray, and N. Chandra, "Intelligent Traffic Light System to Prioritized Emergency Purpose Vehicles based on Wireless Sensor Network," International Journal of Computer Applications, vol. 40, no. 12, pp. 36-39, February 2012.

[22] J. Zheng and A. Jamalipour, Eds., Wireless Sensor Networks: A Networking Perspective. New Jersey, Hoboken: John Wiley & Sons, Inc., 2009.

Mobile-Based Location Estimation Using Single Base Station

S. Al-Bawri[1] and A. Zidouri[2]
[1& 2] Electrical Engineering Department, College of Engineering Sciences,
King Fahd University of Petroleum & Minerals, Dhahran, KSA

Abstract - *The location of subscribers has received huge attention in application for the wireless services. It is well-known that the satellite-based positioning system (GPS- Global Positioning System) is a sufficient and reliable technique for coordinating services. Many techniques using base-stations depend on some wireless parameters such as signal strength, reflection factors, ducting are now being deployed for estimating the locations of the subscribers. This paper discusses some scenarios of wireless mobile location estimation by utilizing only single base-station (BS). A site of Al-Dhahran city in KSA will form the base map for the proposed scenarios and proper solutions will be suggested accordingly. The simulation results performance of the approaches used will be compared with the FCC standard readings.*

Keywords: Non linear least square (NL-LS), Triangulation calculation, AoA, ToA, Location Estimation, Position localization.

1 Introduction

In wireless communications, acquiring the location for emergency calls allows a coordinated response in states where users are corrupted, cannot respond or speak, or do not know their positions. On October 2001, a decision was made for emergency calls that use cellular phones by the Federal Communications Commission (FCC). They announced that all calls must have a localized accuracy of 67% within an area of 125 m^2 in these cases [1]. In addition, service providers are attracted to wireless applications involving position-localization techniques.

Position-localization capabilities lead the way to a new dimension of relatively unrealized information applications which can be provided to the consumer in addition to the standard telephony services. Moreover, improved public services such as airway booking management, traffic mapping, and real time vehicles services are made possible with position-location systems. To make this all possible, the wireless service providers must have knowledge of subscriber's locations. In order to locate subscriber position, at least three BSs are required to acquire satisfactory precision even in the most complex positioning algorithms.

This paper will perform an algorithm that utilizes only a single BS to locate mobile station (MS) in cellular networks using its own antenna array. Different scenarios will be considered on a site of Al-Dhahran city in KSA which is a regular region. Simulation results are compatible with those of U.S. FCC regulations.

2 Approach and methods

There are several methods that were proposed to detect and identify the mobile location. The approaches differ from one another depending on the number of BS's required. Subscriber location can be obtained from various signal parameters e.g. time delay, signal strength, the direction of main beam …etc. A number of locations estimation techniques are provided as follows:

2.1 Time of Arrival (ToA)

Precise position in wireless networks has gained huge interest over the past decades. Without utilizing satellite-based positioning systems (e.g., by Global Positioning System), many wireless positioning techniques have been presented by making use of only its own radio measurements while transmitting. One of these methods is (Time of Arrival) or time difference of arrivals. The basic principle of coordinating a MS depends on the triangular geometrics from a set of constant reference points such as the angular measurements case as shown in figure 1.

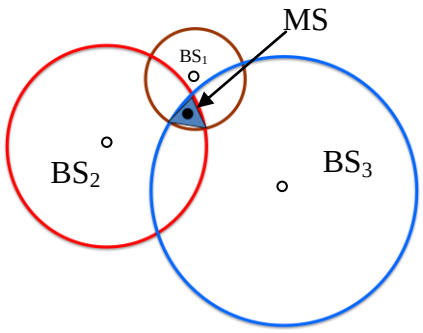


Figure 1. Three BS's in Time of Arrival Positioning

It is needed at least three known paths for measuring distances. However, in a more sophisticated wireless networks positioning scenario these distances are not directly recognized, but must be approximated from a larger set of recorded data; typically, those data can be

found in the mainstream exchanges between the base station and the mobile station.

$$\hat{x}_{MS} = \frac{x_1 + x_2 + x_3}{3} \quad (1)$$

$$\hat{y}_{MS} = \frac{y_1 + y_2 + y_3}{3} \quad (2)$$

where $\hat{x}_{MS}$ and $\hat{y}_{MS}$ are the estimed location of the MS.

2.2 Angle of Arrival (AoA)

This technique identifies the MS location by measuring angle of arrival (AOA) to determine a signal from an MS at various BSs through their antennas as shown in figure 2. Obstacles around the MS and BS will affect the measured AOA. The antennas will depend on a reflected signal if there is no LOS signal component which could not be coming from the direction of the MS. Even when there is LOS component, multipath components will still overlap with the measured angle [2].

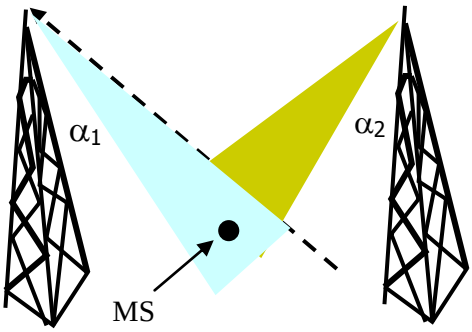


Figure 2. AOA (Angle of Arrival).

2.3 Single Base Station Location Technique

The modern location estimation implemented for microcellular wireless networks utilizes just only one BS. In a microcellular environment, the dominant propagation components contain diffraction, reflections, and scatterings within covered area, in addition to the free space path loss. These multi-path components MPCs are received and processed at the BS, which are then classified according to their time, angle and power. On the other hand, the MPCs arriving at the BS have large angular spread. By utilizing advantage of these features, the subscriber's location will be identified through some of the following demands and operations:

2.3.1 Requirements

The proposed technique uses the antenna array of a single-BS to locate unmodified MSs. Using a single-BS offer many advantages.

- Synchronized MS with other BSs is not necessary.
- The problem of several BSs coverage is no longer needed.
- Signalling, (back haul), for internet service requirement is reduced [3].

The algorithm requires some information about the environment around the MS. Moreover, the selected positioning algorithm requires another prerequisite which is the fundamental function φ(α) to identify whether the subscriber is LOS or NLOS.

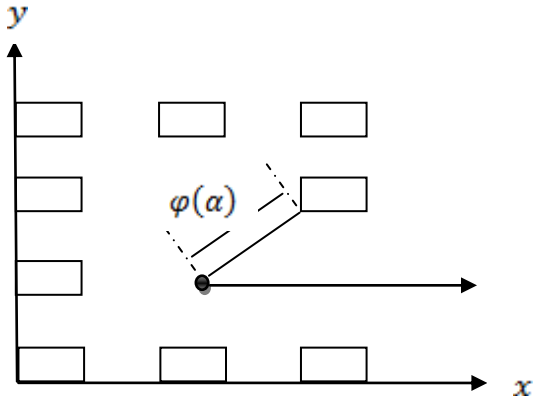


Figure 3. A definition of fundamental function

2.3.2 Description of LoS and NLoS

- *Estimation of LoS Conditions:*

The first MPC received at the BS will have an absolute distance computed as

$$d_i = \sqrt{(c\tau_i)^2 - (h_{BS} - h_{MS})^2} \qquad (i = 1, \dots, N) \quad (3)$$

where h_{BS} and h_{MS} are the heights of the BS and MS respectively, while τ_1 is the absolute value of propagation delay of the first MPC reaching the BS. The distance d_i, the $\varphi(\alpha)$ and the AoA of the first impinging MPC α_1 are used to decide whether the MS is in LoS condition or not. If $\varphi(\alpha_1) \geq d_1$ then LoS is assumed and the position of the MS is simply estimated as [3][4]

$$\begin{cases} \hat{x}_{MS} = x_{BS} + d_1 . \cos(\alpha_1) \\ \hat{y}_{MS} = y_{BS} + d_1 . \sin(\alpha_1) \end{cases} \quad (4)$$

- *Minimization of Algorithm in NLoS Conditions:*

The MS is considered to be NLoS if the previous test $\varphi(\alpha_1) \geq d_1$ is not satisfied. For instance, by minimizing the cost function, the MS position is determined from the received MPCs [3][5]. First through the scatterers, the coordinates are evaluated by the algorithm. Each MPC ray related to these scatterers are points where it has several reflections and/or diffractions before receiving it at the BS. Therefore, the coordinates are obtained as

$$\begin{cases} x_{Si} = x_{BS} + \varphi(\alpha_i) . \cos(\alpha_i) \\ y_{Si} = y_{BS} + \varphi(\alpha_i) . \sin(\alpha_i) \end{cases} \quad i = (1, \dots, N) \quad (5)$$

After identifying the scatterer locations, the algorithm computes the vector τ_{Ri}which represents the propagation delays between the MS and the obstacles.

$$\tau_{Ri} = \tau_i - \frac{\varphi(\alpha_1)}{c} \qquad (i = 1, \dots, N) \tag{6}$$

this delay will introduce a cost function :

$$F(x,y) = \sum_{i=1}^{N} f_i^2(x,y) \tag{7}$$

with

$$f_i(x,y) = c\tau_{Ri} \sqrt{(x - x_{Si})^2 - (y - y_{Si})^2} \tag{8}$$

Since the MS cannot be more than $c\tau_1$ away from the BS, further minimization to the cost function is applied. So, the estimated position $(\hat{x}_{MT}, \hat{y}_{MT})$ of the MS is chosen as

$$(\hat{x}_{MS}, \hat{y}_{MS}) = \arg \min_{(x,y)\in D} \{F(x,y)\} \tag{9}$$

with

$$D = \left\{(x,y) \mid \sqrt{(x - x_{BS})^2 - (y - y_{BS})^2} \leq c\tau_1\right\} \tag{10}$$

Many techniques for minimizing the cost function are used to solve the nonlinear least square (NL-LS) in (8). The following case study will use MATLAB in simulating and computing the estimated MS locations.

3 Case study

The selected location technique has been estimated according to the difference between the actual and the expected position:

$$\varepsilon = \sqrt{(x_{MS} - \hat{x}_{MS})^2 - (y_{MS} - \hat{y}_{MS})^2} \tag{11}$$

where ε is the location error [3][4][6]. If the LOS condition is verified, then the location is estimated by (4) and otherwise by (9).

Different scenarios in a regular region will be considered, one assumes the BS is centered and another with it being outer most of the coverage. Also, the impairments of finite-resolution in estimating the channel-characteristics will be covered. Simulation results are compatible with those of U.S. FCC regulations. Our measurements were taken for Al-Dhahran using Google Earth as follows: There are 64 buildings considered. The scenario will take place in the following region:

- Cell coverage area $220 \times 220\ m^2$.
- Street width $20\ m$.
- Building height = 30 m, building width = $20\ m$.
- The MS height is 1.5 m.
- Five meters is assumed as spacing between any two points.

As a result, 492 actual subscriber locations are available. Six multipath components affecting on the BS have been chosen.
Two different positions are assumed for the BS antenna.

1- Three BSs are located on $(x_1 = 50,\ y_1 = 50)$, $(x_2 = 110,\ y_2 = 170)$ and $(x_3 = 170,\ y_3 = 50)$ see Figure 4.
2- BS is located at $(x = 110, y = 110)$ in the middle of a street junction Fig. 5.
3- BS is mounted on the rooftop of a building near the street junction.

Figure 4. 64-buildings in Al-Dhahran with Three BSs

Figure 5. 64-buildings in Al-Dhahran with Single BS located at the center of the scenario

4 Simulation and experimental results

There are two scenarios which are done using some numerical result as seen in following:
First: (The 64-buildings Al-Dhahran environment with BS in the middle of the scenario).

- The actual mobile location (492).
- The BS, MS, Building height.

- Compute the fundamental function distance to know where the MS is LOS (21 MS) or NLOS (471).
- Compute the position of the scatterers.
- Estimated the coordinates of the LOS MS by using eq. (4) and NLOS MS by minimizing the cost function using eq. (9).
- Compute the error in LOS, NLOS, and overall states.
- Then, to calculate the distribution function CDF.
- Finally, plot the error location with its distribution function.

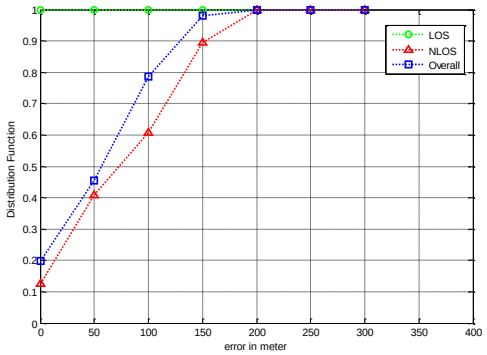


Figure 6. BS is located at center of the scenario

Second: (BS mounted on a corner of a building) as previously worked but change the position of BS. So, the LOS and NLOS MSs and the error value will change also.

The results shown in fig. 6 indicate that 88% of users will be known within125 m where less than this value within 79% of them will be identified when the BS located on the corner of one building around the center of the region as shown in fig. 7, so the simulated single-BS technique accuracy is comparable to that of the three-BS's method which uses different BS's. On the other hand, the minimization of the chosen cost function to get the best 6 MPCs is important where the farthest point at each building were selected, so that the signals will be diffracted from their edges, as a result fig. 8 and 9 will illustrate the cost function before and after minimization.

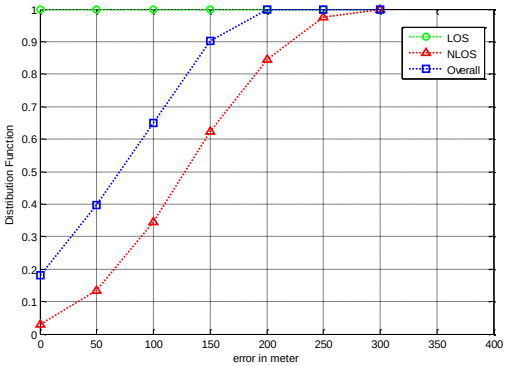


Figure 7. BS is located at the top roof of building in the scenario

5 Conclusions

In this paper, we highlighted the main techniques for estimating the location of MS by introducing a number of selected location-positioning metrics. We exploited an algorithm that utilizes a single-BS with directional antenna array, without doing modifications to the MS, in addition to some knowledge about the surroundings near the BS. The simulation results are shown to estimate the MS position with a satisfactory precision by applying the selected technique on a site of Al-Dhahran which has similar environment. Different scenarios were taken to explore the channel-parameter estimation. A deterministic ray tracer test was taken to verify the employment. In conclusion, the performance of the single-BS localization can be more accurate when having MIMO implemented on both the MS and BS.

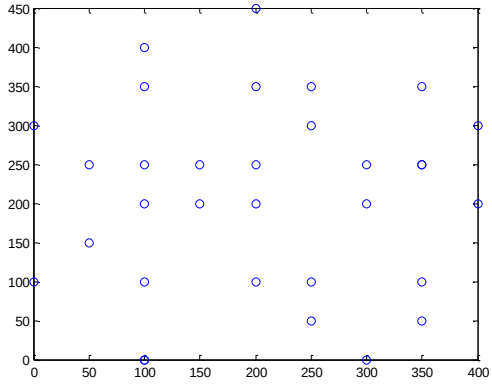


Figure 8. The cost function before minimizing

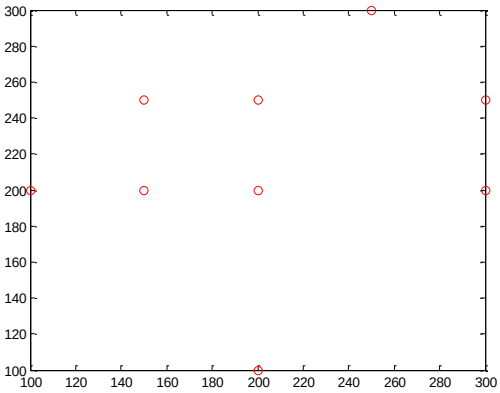


Figure 9. The cost function after minimizing

6 References

[1] "Revision of the commissions rules to ensure compatibility with en- hanced 911 emergency calling system." *Washington*, DC, RM-8143,CC Docket 94-102. U.S. FCC, 1996.

[2] J. J. Caffery and G. L. Stüber, "Overview of radiolocation in CDMA systems," *IEEE Commun. Mag*., vol. 36, pp. 38–45, Apr. 1998.

[3] Porretta, M.; Nepa, P.; Manara, G.; Giannetti, F.; Dohler, M.; Allen, B.; Aghvami, A.H.; , "A novel single base station location technique for microcellular wireless networks: description and validation by a deterministic propagation model," *Vehicular Technology, IEEE Transactions on* , vol.53, no.5, pp. 1502- 1514, Sept. 2004.

[4] Porretta, M.; Nepa, P.; Manara, G.; Giannetti, F.; Aghvami, A.H.; Dohler, M.; , "Validation of a novel radio location technique by a deterministic propagation model," *Antennas and Propagation Society International Symposium, IEEE* , vol.1, no., pp. 81- 84 vol.1, 22-27 June 2003.

[5] Wei-Yu Chiu; Bor-Sen Chen; , "Mobile location estimation in urban areas using mixed Manhattan/Euclidean norm and convex optimization," *Wireless Communications, IEEE Transactions on* , vol.8, no.1, pp.414-423, Jan. 2009.

[6] Shixun Wu; Jiping Li; Shouyin Liu; , "An Improved Reference Selection Method in Linear Least Squares Localization for LOS and NLOS," *Vehicular Technology Conference (VTC Fall), 2011 IEEE* , vol., no., pp.1-5, 5-8 Sept. 2011.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of King Fahd University of Petroleum & Minerals and Hadhramout Establishment for Human Development.

Samir Al-Bawri received the B.S. degree in 2003 from Hadhramout University, Yemen, in electronics and communications engineering, and the M.Sc. degree in 2009 from Yarmouk University, Jordan, with a focus on wireless communications. He is currently working toward the Ph.D. degree at KFUPM in Saudi Arabia. His research interest is in the area of Signal Processing and Location Estimation Techniques.

Dr. Abdelmalek Zidouri is a Senior IEEE member and an Associate Professor in the Department of Electrical Engineering at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. He holds a Master of Science in control engineering from Bradford University, UK, in 1984, and a Doctor of Engineering in applied electronics from Tokyo Institute of Technology Japan, in 1995. His research interests are in the field of signal processing and pattern recognition. In particular character recognition and document image analysis. Dr. Zidouri has published many refereed journal and conference papers. He is the head of the Digital Signal Processing Group at the EE Department. He is a member of Engineering Education Society, Signal Processing Society, and Communications Society.

Mobile Computing: Security Issues

Mohamad Ibrahim Ladan, Ph.D.
Computer Science Department, Haigazian University, *Beirut – LEBANON*
mladan@haigazian.edu.lb

Abstract- *Laptop computers, cell phones, mobile data storage devices, and similar mobile computing and communication devices have become very popular because of their convenience and portability. This has led to the creation of a new computing platform called mobile computing. However, the use of such devices in this new platform is accompanied by new security risks that must be recognized and addressed to protect the physical devices, the communication medium, and the information used. In this paper I will investigate and discuss the new security issues introduced by mobile computing, and summarize the current existing security measures and proposed solutions for these issues.*

Keywords*: Mobile computing security, wireless communications security, mobile devices security.*

1. Introduction

With the rapid growth in the wireless mobile communication technology, and because of their benefits, convenience, flexibility, and the ability to communicate with fixed network while in motion, small devices like laptop computers, smart cell phones, personal digital assistants (PDAs), tablets, mobile data storage devices, and similar mobile computing and communication devices have become very popular at all user and application levels. As a result, a new computing platform called mobile computing is becoming widely spread. Mobile computing is a frequently used term that can be defined as having access to computing resources from anywhere using mobile devices However, the use of such devices in such a platform comes with new security risks and challenges that must be recognized and addressed to keep this new computing environment safe and secure.

Mobile computing devices are capable of storing, processing, displaying, and communicating information. This information could be sensitive information, such as the identification and credit data of customers, and the mobile devices can move in and out of the boundaries of a networking environment. Mobile users have the ability to work from anywhere without being bound to any networking system. This flexibility extends the network boundary beyond a fixed point and makes security management a much more difficult task, as the users cannot be tracked down to a single location. Therefore, the implicit presumption that everything inside the company's network's firewall is secure turns out to be not true any more [1]. Mobile communication takes place mainly through the radio signals rather than wires, so it is easier to intercept or eavesdrop on the communication channels. Hence, traditional security technologies such as firewalls, authentication servers, biometrics, cryptography, intrusion detection, virus protection, and VPNs are not enough to tackle security issues in mobile computing.

Despite the fact that every mobile computing user is concerned one way or the other about security, the defense community personnel are the most concerned. "Cyber security and data protection is one of the current 'space races,'" said David Machuga, director of identity management and business solutions at Northrop Grumman, which supplies the Defense Department with secure mobile biometric data collection technology. "There are several challenges, but the three largest are how to certify the security, how to protect the information transmitted in a wireless mode and how to protect data at rest," he said. "For DOD, this is a constant problem, whether dealing with cell phones, handheld bar code scanners used in a warehouse or mobile computers used to validate biometrics." [2].

In addition to the general security issues like confidentiality, integrity, availability, legitimacy, and accountability that needs to be individually taken care of, new security issues need to be handled properly. These new issues will be discussed in section 3 of the paper and can be found under different areas of risks like physical risk, unauthorized access risk, application and operating environment risk, communication and network risk, and mobile data storage device risk. Although I cannot cover every security issue or problem in details, this paper provides a good discussion and a strong overall coverage and classification of the security problems and challenges arising in mobile computing environment, and it gives a good summary of the available techniques used in handling the different types of security risks and issues.

The paper is organized as follows: In section 2 is the traditional security issues found in most computing and communication systems are presented and discussed. In section 3, the new security issues and challenges introduced by mobile computing are presented, discussed, and classified. In section 4, different techniques and requirements used in dealing with the new security challenges are presented and discussed. Finally, in section 5, a summary and conclusion of the paper is given.

2 Traditional Security Issues

There are several traditional security issues in information systems in general that are addressed by any application developer and are more relevant in mobile computing systems in particular. In what follows, I will discuss briefly the most addressed general security issues that are common to most information systems.

- *Confidentiality*. This ensures that information stored on a system or transmitted over communication links, is only disclosed to those users who are authorized to have access to it. It protects the privacy of the information exchanged between any two or more devices or systems. Data should be encrypted so that if the communication was intercepted, then there should be no disclosure of information.
- *Integrity*. This prevents against intentional or unintentional data modifications during transmission. It ensures that information exchanged between different parties is accurate, complete and not altered during transmission. If data transmission is intercepted, data alteration and retransmission should not be allowed, inserting new messages or information into the communication link should be prevented.
- A*uthentication*. This enforces the verification and validation of the identities and credentials exchanged between mobile systems or a mobile device and a service provider. It ensures that the user accessing the information is the right person. The initiating service requester must be authenticated to prove its identity with reliable credentials to prevent masquerading attacks (when a user is deceiving about its real identity which may lead to impersonation).
- *Authorization*. This ensures that the service requester has the right to access the information on different network or mobile resources. It defines the policies associated with the required access control to the resources.
- *Non-repudiation*. This ensures that the different communicating parties cannot deny the exchange of information or the acceptance of a committed transaction at a later time. It ensures accountability.
- *Availability*. This ensures that the mobile computing environment or the services of the information systems are all the time available for users This could be threatened by a well known type of attacks, *Denial of Service (DoS)*, where an unauthorized user or a hacker tries to disrupt a service or a device by flooding it with useless traffic that consumes server/device resources and forces them to be unavailable.

3 Mobile Computing Security Issues

Mobile computing is a broad area that describes a computing environment where the devices are not restricted to a single place. It is the ability of computing and communicating while on the move. Wireless networks help in transfer of information between a computing device and a data source without a physical connection between them. These networks include wireless LAN, wireless access point, and cellular networks [3]. So some of the new security issues introduced in mobile computing are originated from the security issues of wireless networks and distributed computing systems. In addition, poorly managed mobile devices introduce new security issues involving information exposure and compromise especially when these devices like laptops, PDAs, iPhones, Blackberrys, and others are loaded with sensitive information and are stolen or fallen into the hands of an unauthorized person. Hence **t**he new types of threats and security challenges introduced by mobile computing can be classified into two main classes:

- ❖ Security issues related to wireless networks and the transmission of information 'over the air' between mobile units and mobile support stations and networks.
- ❖ Security issues related to the mobility of the devices and the information residing on them.

3.1 Wireless Networks Security Issues

Wireless networks have their own security issues and challenges. This is mainly due to the fact that they use radio signals that travel through the air where they can be intercepted by location-less hacker that are difficult to track down. In addition, most wireless networks are dependent on other private networks, owned and managed by others, and on a public-shared infrastructure where you have much less control of, and knowledge about, the implemented security measures. Although encryption aid to some extend in securing information moving across wireless networks, the moment the data leaves a mobile device and heads onto a communication network, it's the network operator's job to ensure that the information is securely transported to its final destination.

In what follows, I will list and discuss the main mobile computing security issues introduced by the use of wireless networks. Most of these issues can fall under one of the following categories: Availability where the availability of information and services could be disrupted, confidentiality where the privacy of information when it passes through the wireless medium can be compromised, and integrity of data where data interchanged can be modified and retransmitted [1, 3].

- *Denial of Service.* This attack is characterized by an explicit attempt by attackers to prevent legitimate users of a service from using that service. DOS attacks are common in all kinds of networks, but they are particularly threatening in the wireless context. This is because, the attacker does not require any physical infrastructure and he gets the necessary anonymity in the wireless environment. The attacker floods the communication server or access point with a large number of connection requests so that the server keeps responding to the attacker alone hindering legitimate users from connecting and receiving the normal service.
- *Traffic Analysis.* The attacker can monitor the transmission of data, measure the load on the wireless communication channel, capture packets, and reads the source and destination fields. In order to do this, the attacker only needs to have a device with a wireless card and listen to the traffic flowing through the channel. By doing such things, the attacker can locate and trace communicating users and gain access to private information that can be subject to malicious use.
- *Eavesdropping.* This is a well known security issue in wireless networks. If the network is not secure enough and the transmitted information is not encrypted then an attacker can log on to the network and get access to sensitive data, as long as he or she is within range of the access point.
- *Session Interception and Messages Modification.* The attacker can intercept a session and alter the transmitted messages of the session. Another possible scenario by an attacked is to intercept the session by inserting a malicious host between the access point and the end host to form what is called man-in-the-middle. In this case all communications and data transmissions will go via the attacker's host.
- *Spoofing.* The attacker may hijack a session and impersonate as an authorized legitimate user to gain access to unauthorized information and services.
- *Captured and Retransmitted Messages.* The attacker can capture a full message that has the full credential of a legitimate user and replay it with some minor but crucial modification to the same destination or to another one to gain unauthorized access and privileged to the certain computing facilities and network services.
- *Information Leakage.* This potential security issue lies in the possibility of information leakage, through the inference made by an attacker masquerading as a mobile support station. The attacker may issue a number of queries to the database at the user's home node or to database at other nodes, with the aim of deducing parts of the user's profile containing the patterns and history of the user's movements.

3.2 Device Security Issues

Mobile Devices are essential and key components of a mobile computing environment. A mobile device is any portable device that belongs to a specific user and has computing and storage capabilities. Mobile devices like laptops, cell phones, iPhones, Blackberrys, PDAs, USBs and other small devices can store vital and sensitive data outside office environment for convenient use by mobile users. But this convenience of mobility and portability is accompanied by several new security threats related to possible unintended data disclosure.

Mobile devices are easily stolen, and theft of such devices is on the rise. In most theft cases the aim was the data stored on the device rather than the device itself. One such well known case that happened in Beirut-Lebanon on Oct, 27, 2010 was the attack on the investigation team of the UN created International Tribunal for Lebanon, set up in 2007 to bring to justice those involved in the assassination of then Prime Minister Rafiq Hariri. The result of the attack was the confiscating of the laptop computers, cell phones, notebooks and other materials that were in the possession of the investigation team. One of the main goals of the attack was the sensitive and crucial data stored on these mobile and portable devices [4]. There is compelling evidence that mobile devices pose one of the fast growing areas of security concern. Since January 2008, Privacy Rights International's published Chronology of Data Breaches documents that 20 percent of the data breaches reported resulted from mobile device losses: Lost laptops, notebook computers, PDAs, portable drives, USBs, CDs, flash cards, SD cards, and disks [5]. As a result of these incidents, all of the major mobile devices makers have taken steps during the past few years to improve device security, such as by providing longer device unlock codes like the case of Apple iOS devices, and extending encryption support to SD cards and other mobile data storage devices. However, many defense technology experts feel that protection measures remain insufficient for defense needs and therefore must be strengthen with additional safety measures [2]. Mobile devices have extra stringent security needs and are vulnerable to new types of security threats

and attacks. They need to operate in foreign networks, such as coffee shops, airport kiosks, or other hotspots, and therefore, can't rely on the organization's firewall for protection. The organization needs a means of managing security configuration, patch deployment and antivirus updates on their devices in the field.

The main new mobile computing security issues introduced by the use of mobile devices include the following:

- *Pull Attacks:* The attacker controls the device as a source of propriety data and control information. Data can be obtained from the device itself through the data export interfaces, a synchronized desktop, mobile applications running on the device, or the intranet servers [3].
- *Push Attacks:* The attacker use the mobile device to plant a malicious code and spread it to infect other elements of the network. Once the mobile device inside a secure network is compromised, it could be used for attacks against other devices in the network [3].
- *Forced De-authentication*. The attacker transmits packets intended to convince a mobile end-point to drop its network connection and reacquire a new signal, and then inserts a crook device between a mobile device and the genuine network.
- *Multi-protocol Communication*. This security issue is the result of the ability of many mobile devices to operate using multiple protocols, e.g. one of the 802.11 family protocols, a cellular provider's network protocol, and other protocols which may have well-known security loop-holes. Although these types of protocols aren't in active usage, many mobile devices have these interfaces set "active" by default. Attackers can take advantage of this vulnerability and connect to the device, allowing them access to extract information from it or use its services.
- *Mobility and Roaming*. The mobility of users and data that they carry introduces security issues related to the presence and location of a user, the secrecy and authenticity of the data exchanged, and the privacy of user profile. To allow roaming, certain parameters and user profiles should be replicated at different locations so that when a user roams across different zones, she or he should not experience any degradation in the access and latency times. However, by replicating sensitive data across several sites, the number of points of attack is increased and hence the security risks are also increased.
- *Disconnections*. The frequent disconnections caused by hand-offs that occur when mobile devices cross different introduce new security and integrity issues [6]. The transition from one level of disconnection to another may present an opportunity for an attacker to masquerade either the mobile unit or the mobile support station.
- *Delegation.* The attacker can hijack mobile session during the delegation process. A delegation is a powerful mechanism to provide flexible and dynamic access control decisions [7]. It is a temporary permit issued by the delegator and given to the delegate who becomes limited authorized to act on the delegator's behalf [3]. Mobile devices have to switch connections between different types of networks as they move and some kind of delegation has to be issues to different network access points. Delegations may be issued and revoked frequently as mobile devices detach and reattach to different parts of the network system.

4 Mobile Security Requirements

The rise of mobile computing brings with it a rise in concerns about security issues in general and about data security in particular. In addition, the rise in the number of lost and stolen mobile computing devices raise the need to implement some protection for the data contained on the mobile devices. Organizations involved in mobile computing cannot rely on the traditional security controls of the mobile devices and network infrastructure, they must ensure that these devices, networks, and communication systems have sufficient integral security controls to protect exchanged and stored data. This is because the mobile devices, computers, and networks used for mobile computing may not be owned by these organizations and may be shared by anyone. Therefore, security controls implemented on the systems within the organizations are not enough and must be complemented by other security mechanisms on top of a mandatory good practice by their mobile users. Different security measures and requirements are implemented and suggested for both the mobile devices and the networks. Some of these measures include the following:

- *Encryptions:* If critical information is held on a mobile device, data encryption should be done to protect the data and prevent access by unauthorized persons.
- *Compliance*. Remote and wireless network access from mobile devices must be subject to the same organization's internal network security policies compliance and measures applied to inner users. Access and connection through public hotspots should be avoided.

- *Standards*. Mobile users must ensure that the mobile devices they use and the information they contains are well protected at all times and adhere to a set of requirements such as strong password protection, full disk strong encryption, locking, regular backups, current antivirus software, firewalls with similar configuration to the organization network's configuration [5].
- *Routing anonymity*. To prevent communication endpoints from being linked, anonymous routing may be used at the network layer. It is extremely useful as a building block for higher level applications as a security mechanism for general networked systems [8, 9].
- *VPN and Wireless Encryption Protocol*. A strong wireless encryption protocol should be used whenever possible, and all external connections to the internal organizational network must be over an encrypted virtual private network (VPN).
- *Network Access Control (NAC).* Network Access Control system should be in place to check and analyze mobile devices trying to connect to the organization network. This will protect the internal network from any system compromises or malicious code or infections the mobile device may have picked up while it was away. It could also ensure that the mobile device is patched, has the appropriate security software installed, running and up to date, and that it otherwise meets the organization's security policy requirements before allowing it to connect to internal network resources. Most NAC solutions offer an option between simply rejecting connections from noncompliant clients, or redirecting them to a site or server with information and resources to enable the device to become compliant [10].
- *Wireless Zero Configurations (WZC)*. WZC should be used so that mobile devices and network configuration setting will be done automatically without any intervention from the user [11]. WZC, also know as WLAN AutoConfig, is a wireless connection management utility included with Microsoft Windows XP and later operating systems as a service that dynamically selects a wireless network to connect to based on a user's preferences and various default settings [12].
- *Use Mobile IPV6* (MIPv6). Mobile IPv6 is a protocol developed as a subset of Internet Protocol version 6 (IPv6) to support mobile connections. Using MIPv6, the quality of services and the management of mobility issues in mobile computing environment will be taken care of in a very efficient and standard way that will aid in securing the data transmission [13].
- *Integration of IPSec/AAA and Hierarchical Mobile IPV6*. Internet Protocol Security (IPsec) is a protocol suite for securing Internet Protocol (IP) communications by authenticating and encrypting each IP packet of a communication session, and the Authentication Authorization Accounting (AAA) protocol helps in managing and controlling network accesses. This requirement can be considered as a fully integrated solution for securing data transmission and network access, and it could encompasses several previous mentioned security requirements [1].

5 Summary and Conclusion

Mobile Computing is an umbrella term used to describe technologies that enable people to access network services anyplace, anytime, and anywhere. It offers a lot of benefits for everyone especially the end users; however, it requires high security measures and introduces new security issues and challenges. In this paper, I have introduced the mobile computing systems, presented and discussed their new security issues and requirements, and presented some of the measures that should be taken to handle these security issues. I have classified these issues into two main classes: the first class includes those issues that are related to the wireless networks and communication systems, and the other class includes those issues that are related to the mobile device and data residing on it. Some of these issues are related or variants to the already existing security issues in other systems, and some are new.

Most security experts agree that users operating or transporting devices in an unsafe manner form the weakest link in the data security chain. They believe that creating and enforcing a mobile device use policy is the best way to ensure the highest possible level of data security [2]. The main ongoing challenges facing administrators and developers of mobile computing systems are related to how to maintain control over mobile device data with the rapid pace in the production of new mobile devices, mobile operating systems, mobile device applications, wireless network services/devices, and other new mobile technologies. New products typically have only a minimal security knowledge base and are more likely to contain undiscovered security vulnerabilities than technologies that have been around for several years. In addition to this, the ease of downloading and installing mobile devices applications adds to the above challenges in keeping mobile devices safe from malicious third-party applications and security vulnerability.

6 References

[1] Adrian Leunga,*,1, Yingli Shengb, Haitham Cruickshankb, "The security challenges for mobile ubiquitous services" Information Security Group, Royal Holloway, University of London, Egham, UKbCentre for Communication Systems Research, University of Surrey, Guildford, Surrey, UK ***, 2007***

[2] John Edwards, DOD tackles security challenges of mobile computing, Defense Systems, June 13, 2011

[3] Sowmya Shriraghavan , Srikanth Sundaragopalan ,Fan Yang ,Jinsuk Jun, "Introduction to Information Security Security in mobile computing"**,** November 5, 2003.

[4] http://www.naharnet.com/stories/en/676.

[5] http://www.nascio.org/publications/documents/NASCIO-SecurityAtTheEdge.pdf, july 2009.

[6] Imielinski, T. and Badrinath, B.R. "Data management for mobile computing." SIGMOD RECORD, **22(1)**, 34-39, 1993.

[7] Q. Pham , J. Reid , A. McCullagh , Ed Dawson, "Commitment issues in delegation process", Proceedings of the sixth Australasian conference on Information security, Jan. 2008, Wollongong, Australia.

[8] Li Zhuang, Feng Zhou, Ben Y. Zhao, Antony Rowstron, "*Cashmere: Resilient Anonymous Routing",*The 2nd Symposium on Networked Systems Design and Implementation, Boston, MA, 2005.

[9] J R Jiang, J P Sheu, C Tu, J W Wu, " A secure anonymous routing protocol for wireless sensor networks", IEEE Journal of Information Science and Engineering, Vol. 680, Issue 2, 2010, Pages: 657-680.

[10] Serrao, G.J., "Network access control (NAC): An open source analysis of architectures and requirements", IEEE International Carnahan Conference on Security Technology (ICCST), Oct. 5-8 2010 , pp 94 - 102 , San Jose, CA, USA.

[11] Adrian Leung and Chris J. Mitchell.: Towards Secure Zero Configuration. in *Proceedings of Western European Workshop on Research in Cryptography (WeWoRC 2005)*, Leuven, Belgium, July 5-7, 2005. pp. 34-36.

[12] Windows XP Wireless Auto Configuration: The Cable Guy, Nov. 2002".http://www.microsoft.com/technet/community/columns/cableguy/cg1102.mspx.

[13] H. Jang, J. Jee, Y. Han, S. Park and J. Cha, "Mobile IPv6 Fast Handovers over IEEE 802.16e Networks," Jun. 2008, IETF RFC 5270.

Low Complexity Quasi-Cyclic LDPC Decoder Architecture for IEEE 802.11n

Sherif Abou Zied[1], **Ahmed Tarek Sayed**[1], **and Rafik Guindi**[2]
[1]Varkon Semiconductors, Cairo, Egypt
[2]Nile University, Giza, Egypt

Abstract—*In this paper, we present a fully pipelined LDPC decoder for 802.11n standard that supports variable block sizes and multiple code rates. The proposed architecture utilizes features of Quasi-Cyclic LDPC codes and layered decoding to reduce memory bits and interconnection complexity through efficient utilization of permutation network for forward and backward interconnection routing. Permutation network reorganization reduced the overall resources required for routing, thus reducing the overall decoder dynamic power consumption. Proposed architecture has been synthesized using Virtex-6 FPGA and achieved 19% reduction in dynamic power consumption, 5% less logic resources and 12% increase in throughput.*

Keywords: LDPC, Iterative decoder, Error correction, Min-sum, Low power, Low complexity

1. Introduction

Low Density Parity Check (LDPC) codes, a class of error correction codes that can perform close to the Shannon limit proposed by Gallager in his 1962 PhD thesis [5], [8], [9].LDPC has been considered in different wireless standards due to its superior error correction performance, including digital video broadcasting (DVB-S2, DVB-T2) for satellite and terrestrial digital television, 802.11ad, 10 Gigabit ethernet (10GBASE-T), broadband wireless access (WiMax), wireless LAN (802.11n), deep space communications and magnetic storage in hard disk drives.

Reduced interconnect complexities, smaller die areas, lower power dissipation, and design reconfigurability (runtime) to support multiple code lengths and code rates are the main optimization areas required for efficient LDPC decoder [11]. Recently Quasi-Cyclic (QC) LDPC codes, a kind of architecture aware codes were adopted by several modern wireless communication standards such as IEEE 802.11n, IEEE 802.16e and IEEE 802.15.3c. On the one hand, QC-LDPC codes facilitate efficient high-speed decoding due to the regularity of their parity check matrices. On the other hand, random-like LDPC codes require complex routing for VLSI implementation, which not only consumes a large amount of chip area, but also significantly increases the computation delay and dynamic power consumption.

In this paper we present a low complexity fully pipelined QC-LDPC decoder based on layered minimum-sum algorithm with much less memory bits used and reduced routing overhead targeting wireless LAN 802.11n standard. Layered decoding is proved to converge approximately twice as fast as classical message passing with a flooding schedule [12]. Architecture of the proposed decoder reduces hardware overhead by utilizing only one permutation network rather than pre-processing and post-processing permutation network as in [2], [14].

2. Background

2.1 LDPC Codes and Decoding Algorithms

Low density parity check codes are a class of linear block codes defined by a sparse $M \times N$ parity check matrix H. N represents the number of bits in the code, called the block length, and M represents the number of parity checks.The information length K is $K = N - M$ for full-rank matrices, otherwise $K = N - rank$. The rate of the code R is defined as $R = K/N$ and gives the fraction of information bits in each codeword. Number of ones per column is called column weight W_c and same number of ones per row is called row weight W_r. A code is said to be regular if all rows of parity check matrix have same weight W_r and all columns have same W_c otherwise, it is irregular code.

LDPC codes can be described graphically using Tanner graph. A Tanner graph consists of check nodes representing check equations, variable nodes representing soft-bits and between each variable node i and check node j there exists an edge if and only if $H(j, i) = 1$. Connected pairs of variable node and check node are called neighbors. Figure 1 shows Tanner graph for generic irregular LDPC code where variable nodes (VN) drawn as circles and check nodes (CN) represented by squares. A path that can be traced from one node in the graph back to the same node in the graph while not passing through any other node more than once is called a cycle. The communication performance of an LDPC code is determined by the girth of its factor graph, where girth is the size of the smallest cycle in the graph.

Low-density parity-check codes are decoded iteratively using the belief propagation algorithm, also known as the message-passing algorithm [5]. LDPC codes have been shown to perform very close to the Shannon limit when decoded using the iterative message-passing algorithm.

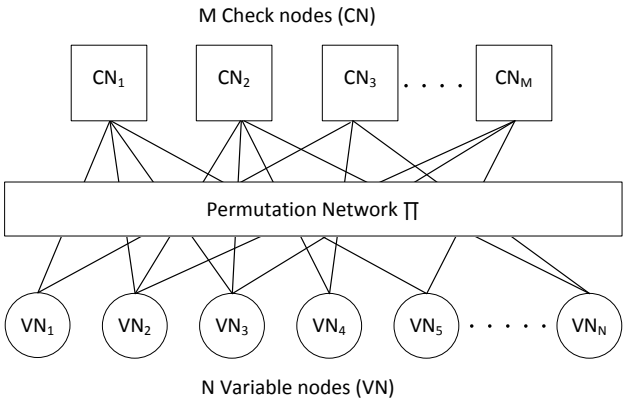


Fig. 1: Tanner graph representation for irregular LDPC codes

The iterative message-passing algorithm is the most widely used method for practical decoding .After receiving the corrupted information, algorithm begins by processing it and then iteratively corrects the received data. Message-passing algorithm can usually reach convergence within a small number of iterations when operating on graphs containing no short cycles. Message passing algorithm can be realised by Sum-Product (SPA) [8] or Min-Sum (MS) [4] algorithms which are near-optimum decoding algorithms and widely used in LDPC decoders.

2.1.1 Sum-Product Decoding Algorithm

The sum-product algorithm (SPA) is the traditional realization of the message passing algorithm. Assuming code word $(x_1, x_2, x_3, \cdots, x_n)$ was sent over a channel and received as $(y_1, y_2, y_3, \cdots, y_n)$. Sum-Product algorithm can be summarized as follows:

- Variable nodes are first initialized with soft information from channel. After initialization, each variable node updates its corresponding check nodes with variable to check messages.

$$\beta_{ij} = L_i = log\frac{Pr(x_i = 0|y_i)}{Pr(x_i = 1|y_i)} \tag{1}$$

  where β_{ij} is message from variable node $V(i)$ to check node $C(j)$.
- Check node update phase: check node messages are calculated using received β messages from variable node as follows:

$$\alpha_{ij} = \prod_{j' \in V(i)\backslash j} sign(\beta_{ij'}) \times \Phi(\sum_{j' \in V(i)\backslash j} \Phi(|\beta_{ij'}|)) \tag{2}$$

$$\Phi(x) = -log(tanh\frac{|x|}{2}) \tag{3}$$

  where α_{ij} is the message from check node $C(i)$ to variable node $V(j)$. Message from check node $C(i)$ to variable node $V(j)$ is calculated using all β messages from neighbouring variable nodes excluding β message from variable node $V(j)$. Φ is a non-linear function represented at equation 3 which indicates magnitude part of equation 2.
- Variable node update phase: Variable nodes calculates variable to check messages β_{ij} using its current value and received α_{ij} messages from neighbour check nodes as follows:

$$\beta_{ij_{new}} = \beta_{ij_{old}} + \sum_{i' \in C(j)\backslash i} \alpha_{i'j} \tag{4}$$

  same as in check to variable messages, Message from variable node $V(i)$ to variable node $C(j)$ is calculated using all α messages from neighbouring check nodes excluding α message from check node $C(j)$.
- Once variable nodes update is finished, reliability values for each bit is calculated same as in equation 4 but with all α messages received from all neighbouring check nodes as in equation 5.

$$z_i = \beta_{ij_{old}} + \sum_{i \in C(j)} \alpha_{ij} \tag{5}$$

  From estimated vector $Z = (z_1, z_2, z_3, \cdots, z_n)$ bit values are calculated by:

$$b_i = \begin{cases} 1, & \text{if } z_i \leqslant 0 \\ 0, & \text{if } z_i > 0 \end{cases} \tag{6}$$

  If $H.B^T = 0$, then B is a valid code word and therefore the iterative process has converged and decoding stops, this is called early termination. Otherwise the decoding repeats until a valid code word is obtained or the number of iterations reaches a maximum number, I_{max} , which terminates the decoding process.

2.1.2 Minimum Sum Algorithm

Minimum sum algorithm [6], [4] is a simplified version of sum product algorithm. Minimum sum algorithm simplifies the calculation of equation 2 even further by recognizing that the term corresponding to the smallest β_{ij} dominates the product term and so the product can be approximated by a minimum:

$$\alpha_{ij} = \prod_{j' \in V(i)\backslash j} sign(\beta_{ij'}) \times min|\beta_{ij'}| \tag{7}$$

Because check node processing requires the exclusion of $V(j)$ while calculating the $min|\beta_{ij'}|$ for α_{ij} , it necessitates finding both the first and second minimums (Min1 and Min2, respectively). In this case $min|\beta_{ij'}|$ is more precisely defined as follows:

$$min_{j' \in V(i)\backslash j}|\beta_{ij'}| = \begin{cases} Min1, & \text{if } j \neq argmin(Min1_i) \\ Min2, & \text{if } j = argmin(Min1_i) \end{cases} \tag{8}$$

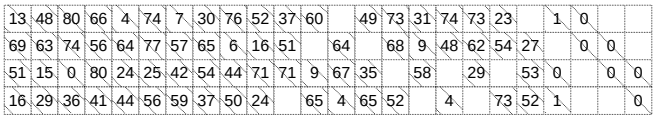


Fig. 2: Base matrix for the rate 5/6 code with codeword length 1944

Table 1: IEEE 802.11n LDPC parameters

Codeword size	Rates	Z
648	1/2 , 2/3, 3/4, 5/6	27
1296	1/2 , 2/3, 3/4, 5/6	54
1944	1/2 , 2/3, 3/4, 5/6	81

Then, minimum sum algorithm goes through same steps of sum-product algorithm till reaching a valid codeword or reaching maximum number of iterations.

2.2 *IEEE 802.11n* LDPC Code

The IEEE 802.11n defines four irregular LDPC codes of rates 1/2, 2/3, 3/4, and 5/6 with three different codeword lengths 648, 1296 and 1944 bits. Figure 2 defines the base H-matrix for the rate 5/6 code with codeword length 1944. In general, H-matrices for QC-LDPC codes are divided into number of macro layers and each macro layer is further divided into number of sub-matrices which are cyclic-permutations of the identity matrix or null sub-matrices. For the 802.11n standard, the base matrices for all codes consists of 24 columns and $R \times 24$ rows of sub-matrices. Each entry in the base matrices defines a $Z \times Z$ sub-matrix. The first $R \times 24$ columns correspond to information bits, the second $(1-R) \times 24$ columns for the parity information which have a fixed structure required by the encoder design. The sub-block size Z of the sub-matrices is variable according to codeword size. Table 1 summarizes different parameters for 802.11n LDPC.

All code rates are suitable for layered decoding. This layered decoding approach which requires a special message scheduling is explained in Section 2.3.

2.3 LDPC Decoding Schedule

Decoding of sum-product algorithm was based on flooding algorithm where in one iteration, all check nodes compute check-to-variable messages and send the results to their neighbouring variable nodes. Then, all variable nodes compute variable-to-check message and send them to their neighbouring check nodes.

Flexible message scheduling was then introduced which is called layered decoding. Layered decoding is done by dividing parity check matrix into L layers and decoding is done row by row and this is called row-layered scheduling investigated in [10], [7] or column by column which is column-layered scheduling investigated in [15]. Layered message scheduling is able to significantly speed up the decoding convergence as each layer passes already updated messages to next layer [10]. Due to its fast convergence speed and implementation regularity , layered message passing has attracted extensive attention for VLSI implementation and it is widely utilized in recent LDPC decoder designs [13].

3. LDPC Decoder Architecture

Decoder architecture must be designed flexible enough to support different code rates and codeword sizes. In general, higher decoder parallelism results in large area and capacitance but low operating frequency and high routing congestion. Fully parallel LDPC decoders maps each CN and VN in tanner graph to a single processing element. Total number of CNs is $M \times W_r$ and number of VNs is $N \times W_c$ while partially parallel LDPC decoders divides the parity check matrix into less number of CNs and VNs such that a set of CN and VN updates can be done at a time. On the one hand, the advantage of high parallel decoders is increased throughput and energy efficiency defined as energy consumed to decode one codeword compared to other less parallel architectures due to fewer clock cycles per iteration [3]. On the other hand, less parallel decoders achieve lower throughput and provide less hardware area and less routing congestion problems also they provide shorter wires than high parallel architectures.

For Layered decoding, layers are processed sequentially since each layer depends on output messages of upper layer. Dependencies between layers impose less parallelism flexibility for layered decoders compared to non-layered ones at which maximum parallelism can be achieved by processing a whole macro layer at a time since all check equations in a macro layer are non-overlapped and independent. Processing of a macro layer can be made more parallel by processing more than one sub-matrix at a time which indicates more variable node processing elements.

Check node parallelism level of $P = 81$ was selected which indicates the maximum number of check equations processed at a time. This size P was chosen from the maximum sub-matrix size presented in the 802.11n standard. Each CN can process various number of input soft bits which indicates number of sub-matrices that can be processed per clock cycle. Each sub-matrix of the simultaneously processed sub-matrices requires its own permutation network. While increasing number of processed sub-matrices per clock increases decoder's complexity and degrades frequency, it increases throughput by reducing number of clock cycles per iteration. To allow flexibility for processing different codeword sizes, CNs are grouped into three groups each group contains 27 CNs which corresponds to minimum sub-block size. For codeword size 648, only first CN group will be active and others will be idle, these idle groups can be utilized to increase throughput by accepting other codewords while processing previous one and this is valid for this codeword size only. Also power reduction can be

done using clock-gating by shutting down idle CN groups in codeword sizes 648 and 1296.

Figure 3 shows overall architecture for proposed decoder. In order to reduce decoder complexity hardware implemented and routing overhead, CNs are designed to process one input at a time. Serialization in CN processing degrades throughput as processing time for one layer will take number of clock cycles equal to number of non-zero sub-matrices per layer but this is still a reasonable trade-off as throughput requirements for decoder can be achieved at a reasonable frequency. CN block functionality will be further illustrated in section 3.1. VNs are used to compute equations 4 and 5. To reduce routing and calculation overhead of VNs, equation 5 can be simply calculated by adding input soft bits of CN block with current CN output value while exclusion of CN previous value that is in equation 4 is done by subtraction of CN previous output value which is stored in internal RAM from current input within CN block . This concept for variable node calculation was introduced in different decoders as in [1].

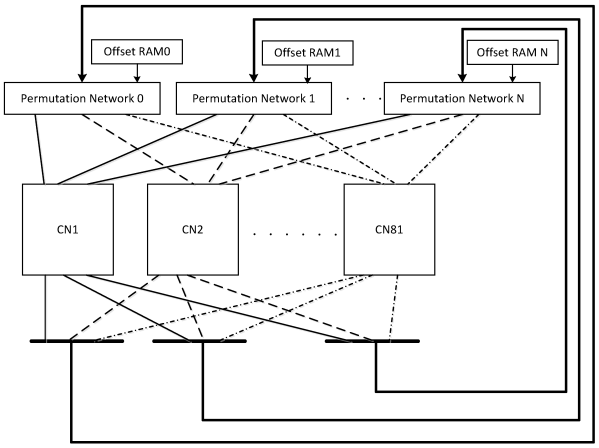


Fig. 3: Overall architecture for partially parallel LDPC decoder

3.1 Check Node Block

Check node block shown in figure 4 is considered the core processor of the decoder. In order to lower number of clock cycles per iteration, layer processing is done in a pipelined fashion inside the CN block. CN performs three main operations divided into three pipeline stages as shown in figure 5. First stage is the marginalization of reliability bits stage. Marginalization is used to do the exclusion of CN's own message resulted from previous iteration from input reliability bits which resembles $\alpha_{i'j}$ in equation 4. Second pipeline stage is detecting first and second minimums of input soft bits and calculating sign bit multiplication which is done using simple XOR gate. Finally, CN to VN message is calculated according to equation 7.

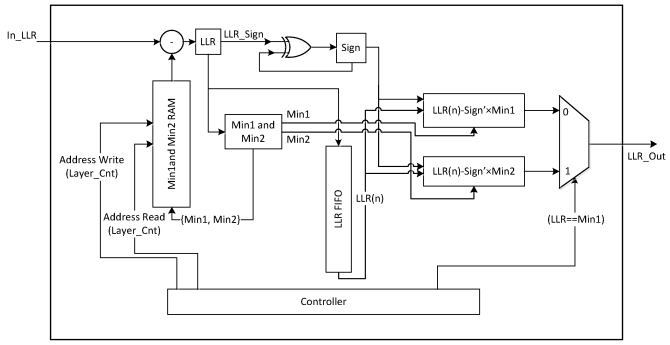


Fig. 4: Check Node block architecture

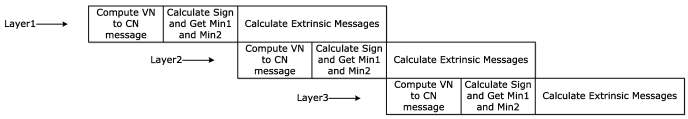


Fig. 5: Check Node pipeline stages

3.2 Permutation Network

For LDPC decoder design, implementation of interconnection between check and variable nodes in the Tanner Graph of the code is always a critical problem, especially in case of high parallelism decoder. A reconfigurable QC-LDPC decoder requires a programmable shift network for different sub-matrix sizes, code rates, and block lengths. Though, an architecture that minimizes the interconnect complexity and, hence, the interconnect delay is more desirable.

Connectivity network between CNs and VNs can be implemented using different scenarios. In [1], only one network was used for routing of messages from VNs to CNs while the route back from CNs to VNs is done by storing these messages inside a RAM and circulation can be done by controlling sequence of memory reads. In [2], [14], one permutation network was used for routing of VNs to CNs messages and another one for reverse direction.

A simple logarithmic barrel shifter composed of modular cells which provide wrap-arounds on 27, 54 and 81 positions for each supported block size was utilized for both routing directions. Since the outputs of CNs are already updated messages, we propose they can be appended directly to their inputs. For the first layer in the first iteration, reliability bits are passed from the wireless channel to the decoder's input directly to permutation network which passes these messages to CNs input permuted by offsets presented at first layer. The output of CNs is then appended back to permutation network which must restore original order of the messages and then perform permutations of layer 2, this is simply

done by rotating these messages by the difference of rotation value of previous layer with respect to the current one. This approach of using one permutation network instead of two reduced routing resources by half also it reduced memory bits required compared to implementation in [1].

4. Results

Figure 6 shows the bit error rate (BER) curves for all code rates with codeword size 648 and six iterations. Due to fixed quantization with 6-bits for internal soft bits inside CNs, a small quantization loss (<0.2dB) can be observed compared to floating point performance for lower rate codes while higher rate codes do not show any significant loss.

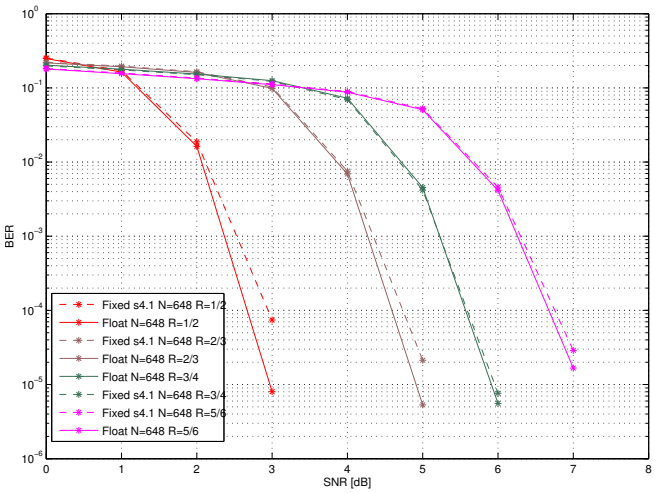


Fig. 6: Fixed-point vs. Floating point simulations for codeword 648-bits

Two designs were implemented in order to verify resources reduction and power optimization of using one permutation network using Xilinx Virtex-6 FPGA. First design uses only one permutation network for both routing directions, second design was implemented with two routing networks same as in [2], [14]. As shown in table 2, logic slices and LUT utilization is reduced by 5% of total FPGA resources. Also as can be seen utilizing one switching network reduced overall switching resources and wirings inside the decoder which reduced dynamic power consumption by 19% and reduced maximum path delay by 1.3ns. Throughput achieved can be calculated as in equation 9.

$$Throughput = \frac{F \times N \times R}{Iter. \times L \times Clk_{layer}} \qquad (9)$$

Where F is the achieved frequency , $Iter.$ is number of iterations, L is number of layers and Clk_{layer} is number of clocks per layer which depends on pipeline overhead which is 5 clock cycles (constant for all rates)in addition to number of non-zero H-matrices per layer.

Table 2: FPGA implementation results of the multi-rate decoder, Device Xilinx Virtex 6 (xc6vlx240t-3-ff1156)

	Design 1	Design 2	Savings
Slices	13,229(35%)	15,322(40%)	5%
LUT	35,668(23%)	42,572(28%)	5%
Dyn. Power(mW)	1066.70	1324.22	19%
Frequency(MHz)	100	88	12%
Throughput(Mbps)	$37.5 \sim 281.25$	$33 \sim 247$	12%

5. Conclusions

A fully pipelined multi-rate, multi-codes low complexity QC-LDPC decoder was implemented for the 802.11n standard. Interconnection complexity was greatly reduced by implementing reusable permutation network for forward and backward message routing. Decoder was implemented on Xilinx Virtex-6 FPGA and achieved 19% reduction in dynamic power, 5% reduction in resources utilized and 12% increase in throughput compared to architecture designed with same routing approach in [2], [14]. We conclude that the engineering of the permutation network has a very significant effect on the LDPC implementation results without compromising its BER. This effect will be more significant with future processes.

Future work will explore a more parallel decoder to achieve higher throughput and the effect of wiring congestion and high switching activity will be further investigated and optimized to achieve best power consumption.

References

[1] T. Brack, M. Alles, F. Kienle, and N. Wehn. A synthesizable ip core for wimax 802.16e ldpc code decoding. In *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on*, pages 1–5. IEEE, 2006.

[2] Z. Chen, X. Peng, X. Zhao, Q. Xie, L. Okamura, D. Zhou, and S. Goto. A macro-layer level fully parallel layered ldpc decoder soc for ieee 802.15. 3c application. In *VLSI Design, Automation and Test (VLSI-DAT), 2011 International Symposium on*, pages 1–4. IEEE, 2011.

[3] A. Darabiha, A. Chan Carusone, and F. Kschischang. Power reduction techniques for ldpc decoders. *Solid-State Circuits, IEEE Journal of*, 43(8):1835–1845, 2008.

[4] M. Fossorier, M. Mihaljevic, and H. Imai. Reduced complexity iterative decoding of low-density parity check codes based on belief propagation. *Communications, IEEE Transactions on*, 47(5):673–680, 1999.

[5] R. Gallager. Low-density parity-check codes. *Information Theory, IRE Transactions on*, 8(1):21–28, 1962.

[6] J. Hagenauer, E. Offer, and L. Papke. Iterative decoding of binary block and convolutional codes. *Information Theory, IEEE Transactions on*, 42(2):429–445, 1996.

[7] D. Hocevar. A reduced complexity decoder architecture via layered decoding of ldpc codes. In *Signal Processing Systems, 2004. SIPS 2004. IEEE Workshop on*, pages 107–112. IEEE, 2004.

[8] D. MacKay. Good error-correcting codes based on very sparse matrices. *Information Theory, IEEE Transactions on*, 45(2):399–431, 1999.

[9] D. MacKay and R. Neal. Near shannon limit performance of low density parity check codes. *Electronics letters*, 32(18):1645, 1996.

[10] M. Mansour and N. Shanbhag. A 640-mb/s 2048-bit programmable ldpc decoder chip. *Solid-State Circuits, IEEE Journal of*, 41(3):684–698, 2006.

[11] T. Mohsenin and B. Baas. Trends and challenges in ldpc hardware decoders. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 1273–1277. IEEE, 2009.

[12] E. Sharon, S. Litsyn, and J. Goldberger. An efficient message-passing schedule for ldpc decoding. In *Electrical and Electronics Engineers in Israel, 2004. Proceedings. 2004 23rd IEEE Convention of*, pages 223–226. IEEE, 2004.

[13] Z. Wang, Z. Cui, and J. Sha. Vlsi design for low-density parity-check code decoding. *Circuits and Systems Magazine, IEEE*, 11(1):52–69, 2011.

[14] M. Weiner, B. Nikolic, and Z. Zhang. Ldpc decoder architecture for high-data rate personal-area networks. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 1784–1787. IEEE, 2011.

[15] J. Zhang and M. Fossorier. Shuffled iterative decoding. *Communications, IEEE Transactions on*, 53(2):209–213, 2005.

Performance Analysis of a WiMAX/Wi-Fi System whilst streaming different Video Conference applications with varying network loads

P.O Umenne[1], **Odhiambo Marcel O**[2]
[1]Electrical and Mining Engineering, University of South Africa, Johannesburg, Gauteng, South Africa
[2]Electrical and Mining Engineering, University of South Africa, Johannesburg, Gauteng, South africa

Abstract - *WiMAX and Wi-Fi are considered as the promising broadband access solutions for wireless MAN's and LANs, respectively. In the recent works WiMAX is considered suitable as a backhaul service to connect multiple dispersed Wi-Fi 'hotspots'. Hence a new integrated WiMAX/Wi-Fi architecture has been proposed in literatures. In this paper the performance of an integrated WiMAX/Wi-Fi network has been investigated by streaming different video conference applications which vary the network load. The difference in performance between the two network connections WiMAX and Wi-Fi is compared with respect to video conferencing. The Heterogeneous network was simulated in the OPNET simulator. Results show that such a heterogeneous network can support a high resolution video conference application and a low resolution video conference application but not a VCR quality video conference application.*

Keywords: Throughput; delay; delay variance; Packet loss; QoS – Quality of Service.

1 Introduction

WiMAX is a popular technology for broadband access in Wireless Metropolitan Area Networks (WMAN) environment. It offers a rich set of features and flexibilities in terms of deployment options and it supports new applications. The physical layer of WiMAX is based on Orthogonal Frequency Division Multiplexing (OFDM), which is widely recognised as the modulation technique for mitigating multipath fading problem associated with broadband wireless system. WiMAX is capable of supporting very high peak data rates. In fact a peak rate of 74Mbps can be achieved when operating with a 20MHz wide spectrum. Under very good signal conditions, even higher peak rates may be achieved by using multiple antennas and spatial multiplexing [1].

One of the potential applications of WiMAX is to provide backbone support for mobile Wi-Fi hotspots. Traditionally wired connections are used as backhaul support for Wi-Fi hotspots. But wired infrastructure is always considered expensive and it should be replaced by wireless backbones. Heterogeneous wireless networks consisting of WiMAX and Wi-Fi have been proposed in the literatures [2], [3]. The architecture of this type of network is shown in Fig. 1.

In this network model a WiMAX base station (BS) serves both WiMAX subscriber and Wi-Fi access points in the coverage area.

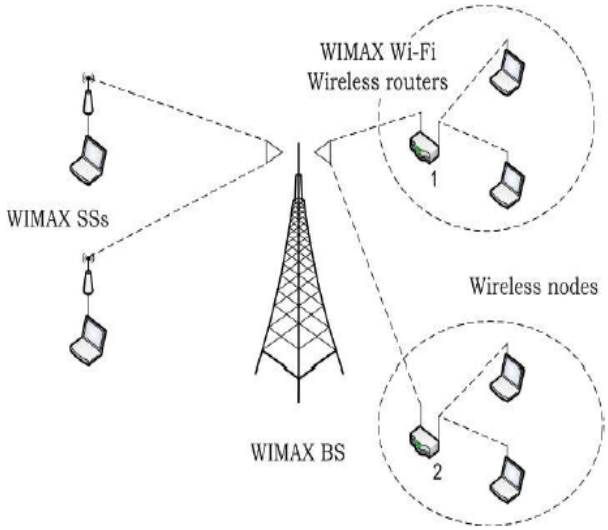


Figure 1 Heterogeneous Network architecture (WiMax/WiFi)

The connection between the WiMAX base station and the WiMAX subscriber station is based on the WiMAX protocol and the connection between the Wireless LAN access points and the Wireless LAN nodes is based on Wi-Fi protocol. Several QoS provisioning mechanisms for integrating WiMAX /Wi-Fi systems have been proposed in literatures [4], [5].

The Quality of Service (QoS) of a Video Conference application is determined by the following parameters; Packet loss: it's a comparative measure of packets received to the total number of packets that were transmitted, Delay: it's a finite amount of time that a packet takes to reach the receiving end point after being transmitted from the sending endpoint, throughput and delay variance (jitter).

In this article we investigated the performance of the WiMAX/Wi-Fi network for different network loads generated by three different Video conference applications.

2 QoS Requirements of a Video Conference Application

QoS parameters for a video conference application are as follows:

Bandwidth and throughput: Bandwidth is the available capacity of connection between two terminals. Throughput slightly differs from bandwidth as it stands for effective bandwidth that is provided by the network.

Delay or latency: It specifies the time it takes for a packet to leave the source and reach the destination.

Jitter (delay variation): Jitter is an interval between subsequent packets. It is caused by network congestion, route alteration etc.

Application	Bandwidth	Sensitivity to:		
		Delay	Jitter	Loss
VOIP	Low	High	High	Med
Video Conferencing	High	High	High	Med
Streaming Video	High	Med	Med	Med
Streaming Audio	Low	Med	Med	Med
Client/Server Transactions	Med	Med	Low	High
Email	Low	Low	Low	High
File Transfer	Med	Low	Low	High

Figure 2 Applications QoS metrics Sensitivity

As can be seen from figure 2 Video Conference applications are highly sensitive to the factors of Delay, Jitter and packet loss. Hence this factors need to be kept at minimum values in order for the QoS to be as high as possible in transmitting or streaming a video application.

For best quality of a picture the above mentioned factors should be kept at the following values [6].

End to end delay should be below 150mS

Jitter should be kept under 30mS.

3. Simulation Methodology

In order to investigate the performance of the integrated WiMAX /Wi-Fi network with respect to a video conference application the OPNET modeller simulation tool was used. The OPNET modeller supports both WiMAX and Wi-Fi technology. Three video conference applications were applied over the network to represent different network loads [6] namely;

- Low resolution video – 45Mbps
- High resolution video – 99Mbps
- VCR quality video – 840Mbps

The network consists of a centrally placed BN_ASN router that has 12 Point-to-point (PPP) links. The router is connected to an application server running the video conference application and 4 logical subnets. Within each logical subnet there is a Base station (BS) based on the WiMAX protocol. Each base station is connected to a WiMAX subscriber Station (SS) which connects to 4 Wireless LAN subscribers such as Laptops etc. The WiMAX subscriber station has two interfaces, the WiMAX interface to communicate with the WiMAX base station and the Wireless LAN interface to communicate with the Wireless LAN based nodes.

The application profile is running in serial mode which means that each application initiates packet generation in a serial manner. The whole process of packet generation lasts till the end of the simulation.

All traffic is discrete. The WiMAX layer was configured with the rtPS (real time polling services) scheduling technique with a maximum sustained traffic rate of 10Mbps and a minimum reserved traffic of 0.5Mbps.

Figure 3 shows the overall network topology.

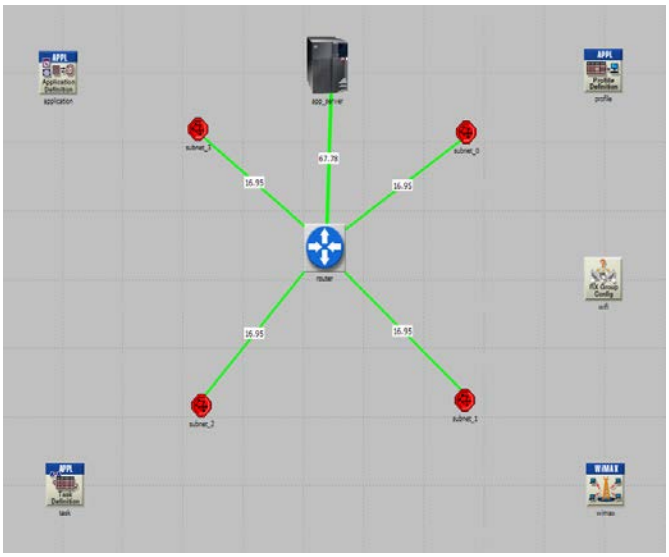


Figure 3 the Network Topology

The topology inside a subnet is shown in detail in figure 4.

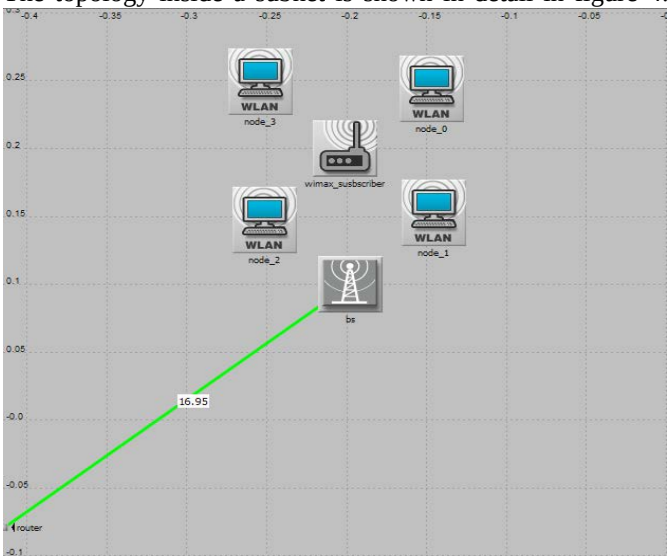


Figure 4 Subnet Topology

The parameters of the WiMAX system are shown in table 1,

Table 1 Parameters of the WiMAX system

Parameters Selected	Values Set
Max No of SS Nodes Supports	100
Transmits Power (W)	0.5
Physical Profile	OFDMA 20MHz
Modulation	Adaptive
Average SDU Size (bytes)	1500
Connection Retries	16
Antenna Gain	15dbi
Service Class Used	Silver
Scheduling type	rtPS

Essentially the parameters set up for the Wi-Fi hotspots are; data rate of between 65Mbps and 600Mbps, physical layer technology of 802.11n 2.4GHz and buffer size of 1,024,000 are all done to enable a video conference application stream over the network without much packet loss and delay, the reason being that the video conference application of high resolution requires high bandwidth of as much as 150Mbps in this case.

The Parameters of the Wi-Fi Hotspots are shown in table 2.

Table 2 Parameters of the Wi-Fi Hotspot

Parameters Selected	Values Set
Physical layer Technology	HT PHY 2.4GHz (802.11n)
Data Rates bits/sec	65Mbps (base)/600 Mbps max
Transmit power	0.005W
Packet received power	-95
Large Packet Processing	Drop
Antenna Gain	14dbi
Access point Functionality	Enabled
Buffer Size	1,024,000
Antenna Gain	14dbi

4. Results

QoS of the Video Conference Application:

As earlier on stipulated the requirements of a video conference application for good picture quality should be as follows [8];

End to end delay should be below 150mS

Jitter should be kept under 30mS.

The performance of the integrated network with respect to the above mentioned factors is as follows;

4.1 Packet end-to-end delay for the whole path WiMAX-Wi-Fi

The packet end-to-end delay for the whole path WiFi-WiMAX is shown in Figure 5 for the different network loads. For the high resolution video conference application the packet delay stabilises on 60mS whilst the low resolution video application stabilises on 45mS. The high resolution video has a higher delay than low resolution because the load of the high resolution video on the network is higher and requires more time to transverse the network.

The VCR quality video application curve does not appear because the load of that application is 840Mbps which leads to a high packet loss on this network. In certain sections of the network the VCR quality video application is not sustained and packet drop is too high, hence it's not possible to get the packet delay for the whole path for this application.

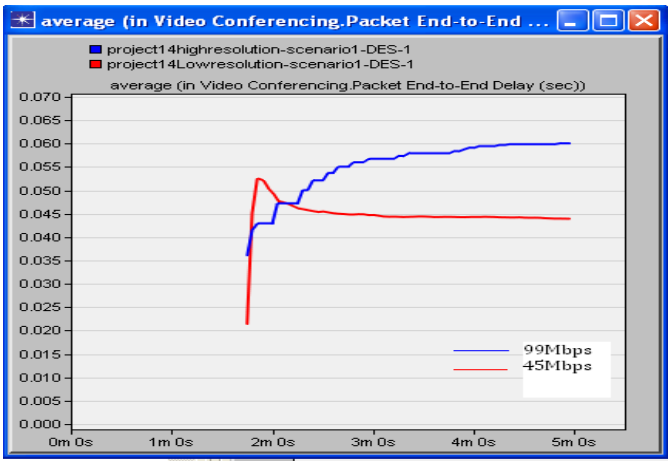


Figure 5 Packet end-to-end delay for the whole path WiMAX-Wi-Fi

4.2 Jitter or Packet Delay Variance for the whole path WiMAX-Wi-Fi

The packet delay variation for the whole path WiFi-WiMAX is shown in figure 6 for the different network loads. For high resolution video the jitter settles on 0.1mS whilst the low resolution video the jitter is about 0.4mS. Generally the high resolution video is a better quality video signal hence producing less jitter or deviation from the signal than the low

resolution video signal. Again VCR quality video does not appear in the graph because of its high network load that leads to high packet drop hence it's impossible to measure the overall network jitter for this application since the application does not complete the path.

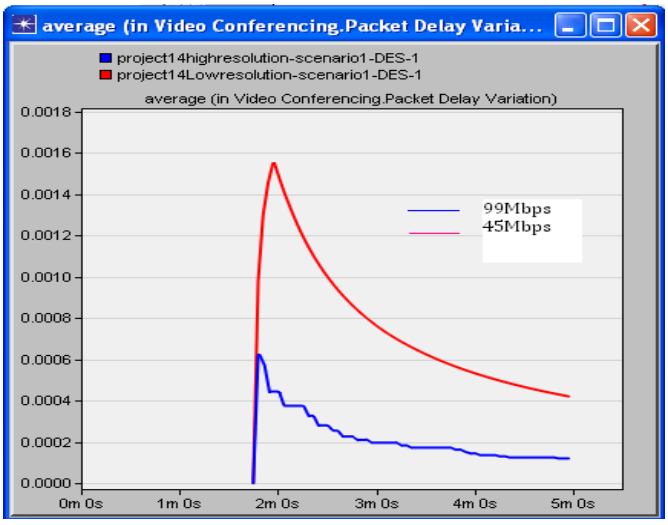


Figure 6 Packet delay variation (jitter) for whole path WiMAX-Wi-Fi

4.3 Throughput of the Wi-Fi connection

The throughput of the Wi-Fi connection is shown in the figure 6 for the different network loads. The throughput for the VCR video quality is higher than for the high resolution video and the low resolution video. However this is because the load on the network from the VCR quality video is highest. The packet dropped for the VCR quality video is also very high.

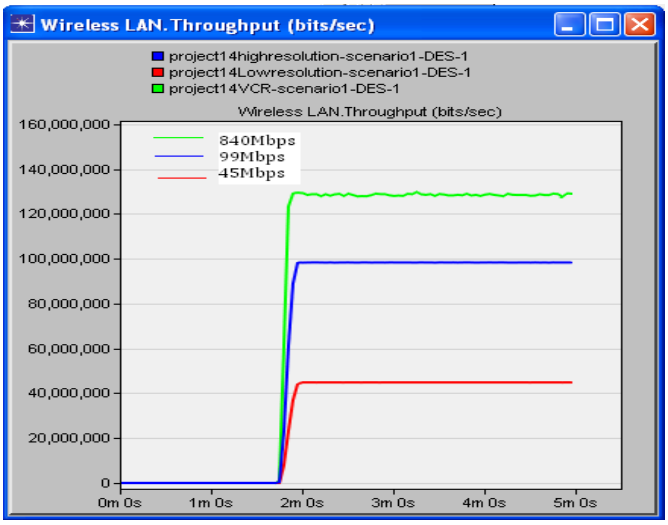


Figure 7 Throughput of the Wi-Fi connection

The curves in figure 7 are total values for the throughput whilst the key chart shows the network loads of those applications.

4.4 Throughput for the WiMAX connection

Essentially the throughput on the WiMAX connection shows a similar pattern to the throughput on the Wi-Fi connection the only difference being that across the WiMAX connection the video conference applications drop more packets hence the throughput on the WiMAX connection for the different video conference applications is less than for Wi-Fi.

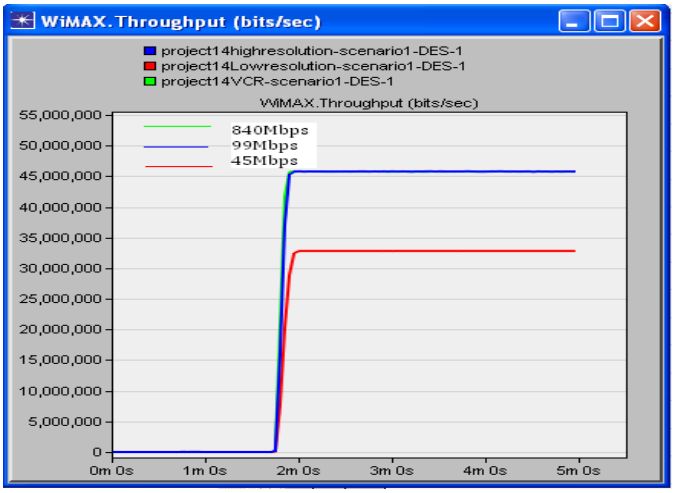


Figure 8 Throughput on the WiMAX connection

4.5 Packet Delay on the Wi-Fi connection

The packet delay on the Wi-Fi connection for the VCR quality video approaches 130mS as can be seen in figure 9 which is relatively high. For the high resolution video it's about 10mS and for the low resolution video it's about 5mS. Again the delay for the VCR quality video is higher than for the other video conference applications due to the fact that the throughput for this application is higher than the other video conference applications.

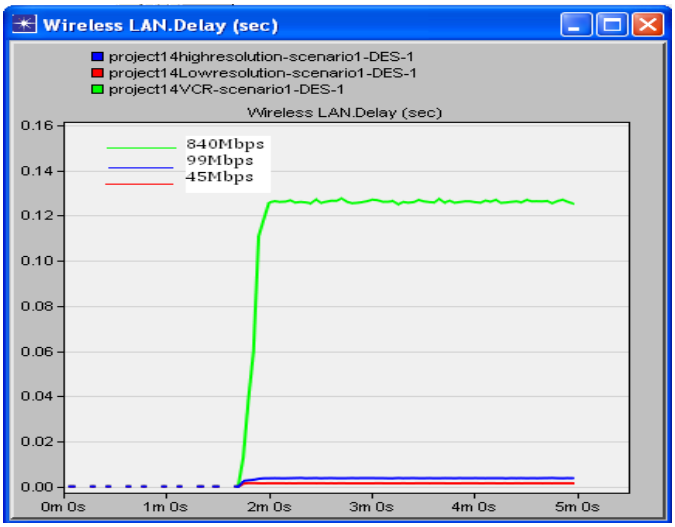


Figure 9 Packet Delay on the Wi-Fi connection

4.6 Packet Delay on the WiMAX connection

The packet delay in the WiMAX connection for the VCR video quality application is about 96mS whilst for high resolution video it's about 94mS and finally for low resolution video its 72mS.

In the WiMAX connection the VCR quality video application has a lower delay as compared to the Wi-Fi connection but relatively high for a throughput of 46Mbps because in this section of the network the VCR quality video drops more packets hence the throughput is lower but the delay is relatively high for such a low throughput because there is congestion in the network. Similarly the delay for the high resolution video and the low resolution video in this WiMAX connection is very high 94mS and 72mS respectively. This are increased values of delay from the values in the Wi-Fi connection due to the fact that there is more packets being dropped in the WiMAX connection due to congestion hence increasing the delay.

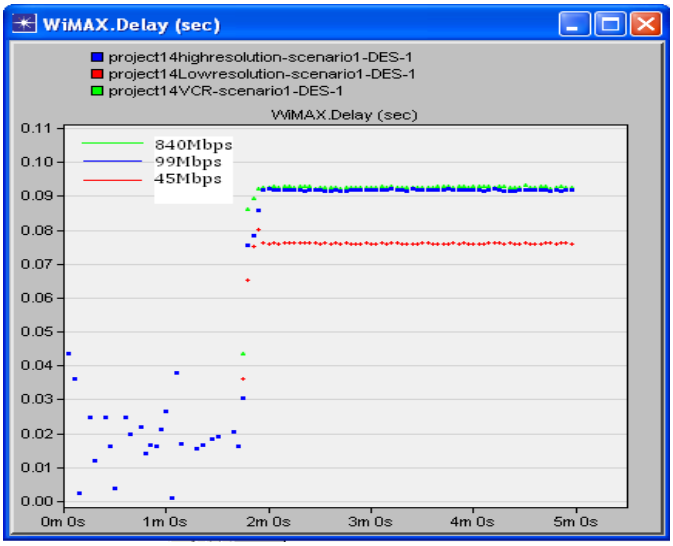


Figure 10 Packet delay on the WiMAX connection

5. Conclusions

In this paper an Integrated WiMAX /Wi-Fi network was modelled whilst streaming video conference applications to determine the performance of the integrated network with respect to the QoS requirements. The network loads were varied by changing the type of video conference application between low resolution video, high resolution video and VCR quality video conference applications. In addition the WiMAX and Wi-Fi connections were compared with respect to throughput and packet delay.

In conclusion a high resolution video application experiences a higher overall network delay as compared to a low resolution video application but has a lower jitter value or packet variance.

Also it was determined that when a network connection experiences congestions and high packet drop rate that network connection would have a higher packet delay than other sections.

Table 3 summarises the results discussed in the conclusion earlier and according to the table the main QoS parameters were satisfied for the high resolution video application and the low resolution video application except for the VCR quality video application which experienced excessive packet drop rate in certain sections of the network exceeding the required QoS parameters necessary for a good video conference transmission.

Table 3 QoS Parameters for Video Conference Application

Parameters	Expected values for a good QoS	High resolution video	Low resolution video
Overall Packet delay	<150mS	<60mS	<45mS
Jitter(Delay Variance)	<30m	0.1mS	0.4mS

6. References

[1] Jeffrey G. Andrews, Ph.D. Arunabha Ghosh, Ph.D "Fundamentals of WiMAX Understanding Broadband Wireless Networking", First Edition, Prentice Hall.

[2] Dave Cavalcanti and Dharma Agrawal, "Issues in Integrating Cellular Networks, WLANS, and MANETs: A futuristic Heterogenous Wireless Networks", IEEE Communication Magazine, June 2005, pp. 30-41.

[3] Dusti Niyato and Ekram Hossain, "Integrating of WiMax and Wi-Fi: Optimal pricing for Bandwidth Sharing" IEEE Communication Magazine, May 2007, pp. 140-146

[4] Kamal Gakhar, Annie Gravey and Alain Leroy, "IROISE: A new QoS Architecture for 802.16 and IEEE 802.11e Internetworking", in the proceedings of IEEE International Conference on Broadband networks", pp. 607-612, October 2005.

[5] Pedro Neves, Susana Sargento Rui and L. Aguiar, "Support of Real-Time Services over integrated 802.16 Metropolitan and Local Area networks", In the proceedings of IEEE ISCC, pp. 15-22, June 2006.

[6] Ahmad Khalifeh, Ashkan Gholamhosseinia, "QOS for Multimedia Applications with emphasis on Video Conferencing", masters project, February 2011, Halmstad University.

[7] Dr. Mayyada Hammoshi, "WiMax Simulation Model to Investigate Performance Factors", Journal of Convergence of Information Technology, January 2011, Volume 6, Number 1.

[8] R.Fantaci, D. Tarchi, "Bridging solutions for a Heterogenous WiMAX-WiFi Scenario, journal of communications and Networks, 2006, 8 pp. 369-377".

Security in Ubiquitous Computing: A Work in Progress

Hatim Aboalsamh Fatimah Alanizi Mashael Bin Sabbar Sumayah AlRabiaah
Computer Science Department
King Saud University
Riyadh, KSA
{hatim, falanizi, msabbar, salrabiaah}@ksu.edu.sa

***Abstract*— Ubiquitous computing change normal physical space into intelligent active space with enhanced services to users. This new field of computing still has some security limitations. In this paper we provide an overview of Ubiquitous Computing security concerns .Furthermore we discuss some issues such as access control and privacy preserving in authentication models. We present access control models and show why the traditional methods are not suitable to use in ubiquitous computing environments. We discuss a trust based access control model and propose a simple UCE example with a simple pseudo code. Also we argue the privacy problem with authentication in Ubiquitous Computing Environment and exhibit two models and compare their security properties.**

Keywords-Ubiquitius Computing; Pervasive Computing; Security; Access control; Authentication; Ubicomp

Introduction

Ubiquitous computing is about placing computing everywhere around human. It focuses on users and accomplishing task rather than machines.. we can say that "Ubiquitous Computing is a technology that resides in the human world and weaves itself into the fabric of everyday life "[1]. This vision is so close to reality [2]. There are some factors which helps to make ubiquitous computing relate to society, technology, and markets [3]. We own and interact with microprocessors embedded in everyday devices. For example cell phones, home appliances, home video systems, cameras, cars and washing machines. [4]. All of these examples show our acceptance and awareness, as a *society*, of ubiquitous computing value. Also *technology* has major participation to facilitate the vision of ubiquitous computing. The final factor affecting the ubiquitous computing is the *market* that has movements towards decreasing cost, increasing capacity and demand for ubiquitous computers.[3]

Nowadays, with the big propagation of the ubiquitous computing, people need some technique to gain the privacy, integrity, and to deal with this kind of computing with more confidence. At the same time, organizations need such technique be reliable.

It is expected for user to interact with hundreds of invisible ubiquitous computers with normal and unremarkable ways, instead of one personal computer per user. If the problem grows hundred times, then the old security solutions are not guaranteed to work in the same way. Security solutions like authentication works with PC's, laptops, and other machines, possibly do not work in the same efficiency with hundreds of ubiquitous computers. The solutions need to be reshaped and give a while to rethink them over [4]. The big challenge that's facing security in ubiquitous computing, is that the security solutions and concepts that applied in the basic interne are not enough to be applied here [5].

In the *PCEs* we have many computers and services embedded inside everyday devices which are shared and available to us.. We should ask: What are the assets? What are the risks? What safeguards and countermeasures to avoid and defend against those security risks? [4]

A good practise to assessing security issues in a *PCE* is through evaluating vulnerabilities from attacker viewpoint.

Security properties such as confidentiality, integrity ,privacy , authentication, access controls increase *PCE*s vulnerability to attacks in in many ways. such as wireless networking. It's known that wireless networks are vulnerable to passive eavesdropping attacks which threatens the Confidentiality. Also, wireless communications increase the chance of Integrity valuations.

When we search for how we can protect information security properties in *PCE*, we found that research papers focused on the use of one of the following three ways:

- Developing a framework and defining requirements of general security.

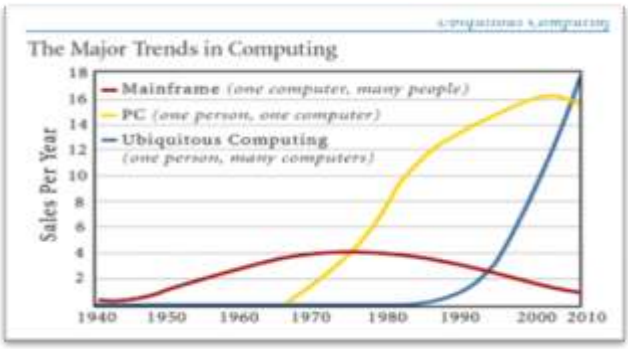


Figure1: Ubiquitous Computing progress

- Developing security protocols that can be work with many PCEs.
- focusing on some security properties to ensure them by designing security infrastructures

I. ACCESS CONTROL

Traditional computer systems depend on access policies and user identities; such as user name, password, or user certificates, to grant or deny users to access resources [6]. But the problem is these identities are not enough in ubiquitous computing for the following issues: (1) in ubiquitous computing, people can visit resources and obtain services any time any where. (2) Ubiquitous computing environments include not only software and hardware entities, but also related with the context and the physical space which dynamically change over time. (3) The requester and the sender do not know each other in advance [7], [8]. For that, access control is not only based on user identification, but rather, it is primarily based on the properties of the environment such as the location and the access time. So, it is more efficient for ubiquitous computing environment in which access control based on implicit trust relationship authentication instead of traditional identification methods.

A. Access Control Methods

In general, there are three access control methods [6].First method is called "Traditional Access Control Methods" which directly grant or cancel user rights after identifying him. These methods are not suitable to be used with ubiquitous computing because in the ubiquitous computing environments there are a hug number of users which make the matching process; of subjects and objects; and management of rights very complex.

The second types of methods are known as "Role Based Access Control Methods". Here, the system first identifies the user. Each user has an associated role. By the given role, group of access rights are assigned to this user. RBAC methods based upon users' identity without considering context information [6]. For that, these types of methods are not good to be used with ubiquitous computing environment.

"Trust Based Access Control Methods" is the last type of access control methods. These methods are based on trust establishment between two strangers. Then, a set of roles are assigned to the strangers in order to give permissions. These methods are the most suitable to be used with ubiquitous computing environment.

B. Trust Policy

Trust and security are closely related. Ubiquitous computing environment need a sufficient level of trust. But there is no clear definition to explain trust, but rather, trust can be understood by its characteristics. The following are the main characteristics of trust:

Subjectivity: Two strangers have different trust degree on the same matter.

Transitivity: trust commonly known to be non-transitive. But to simplify the trust model, trust can be considered to be transitive.

Anti-symmetry: the trust degree between two strangers may differ.

Measurability: trust can be measured like any information by trust value.

Multidimensionality: The trust degree to the same target may differ.

Dynamic of trust: trust has relations with time; different time may have different trust degree.

Context relativity: when trust is given to the other strangers, it is happen for particular purposes.

C. Trusted Based Access Control

In this section we introduce how to apply access control in ubiquitous computing environment using a "Trust Based Access Control Model".There are two trust values which togather give the final trust degree. T_a is the initail trust and it is determine according to the authintication process. This part of trust is static because in one session the system either trust the user or do not. The second part of trust is the dynamic trust we denote it as T_c . This dynamic trust can differ in one session according to the context information; for example physical boundries. By the value of T_a and T_c , The final trust degree T will take a value in between. Acoording to this value T, a subset of roles with an associated access rights will assigned to the user.The life cycle of this model is consisting of four steps:

step1. The two strangers build the initial trust relationship according to their properties (static trust).

$$Ta = f(A1,A2, \ldots., An)$$

step2. The dynamic part of trust is a function of context information:

$$Tc = h(c)$$

step3. Compute the value of trust between the two strangers according to Ta and Tc:

$$T = (Ta, Tc)$$

step4. The user is mapped a subset of roles by the value of T.

D. Our Proposed Model

We propose a simple model which consider both the authentication process and the context information. To explain our model, let's take this example. Suppose a simple ubiquitous computing environment where Prof. Bob can be authenticated using his mobile device, after hie initialy enters a pasward to access his office. In his office, there are sensors to detect his mobile and give him a faculty status in" he will get an access to all facuilty rights. If some "student" holds Prof. Bob's mobile and try to enter his office, he can not because the authentication process will failed; for example, wrong password, and he will not get the initial trust. As figure2 shows, the initial trust can get two values either "trusted" or "un-trusted". Also the dynamic trust will take two values:

	In	out
Trusted	prof.	faculty
Un-trusted		student

Figure 2: A model of Initial Trust

```
Tc = Get location();
if( Tc = "out"){
   Ta = Authenticate user();
    If (Ta = "Trusted"){
          Open the door;
          Tc = Get location();

          If (Tc = "in"){
            Set role = "professor";
            Set Printer usage = "enable";
          }

          Else{
            Set role = "faculty";
            Set Printer usage = "disable";
          }
    }
    Else{
          Set role = "student";
          Set Printer usage = "disable";
}
```

Figure3: pseudo code of the proposed model

"in" or "out". If Ta = "Trusted" and Tc = "In", "Professor" role will assigned to the user with the associated access rights which is in our example "enable printer usage". "Faculty" role will assigned to the user if Ta = "Trusted" and Tc = "out". And "student" role will assigned based only to the value of Ta = "un-trusted". In figure 3, will write a pseudo code to explain our example.

II. PRIVACY PRESERVING AUTHENTICATION METHODS

Ubiquitous computing environments (UCEs) or Active *Information Spaces* encourage the propagation of hundreds or thousands of embedded devices, which are distributed and accomplish tasks. Also, the abundant services provided promise great incorporation of the digital infrastructure into many aspects of our everyday life. Individuals, companies and organizations depend on networking technologies to transfer and process data and to provide services, to exploit the ubiquitous computing environment advantages. As might be expected, many of these resources and services will be sensitive and critical for users. [9][10]

Some services provided in ubiquitous computing environment are automated and resources could be accessed anytime anywhere, trying to increase the productivity of users and makes services always available. These properties give hackers, cyber-attacker and unauthorized intruders more opportunities to break into the system and make great damages. Also, the ubiquitous computing environment integrates the physical world and the virtual world, this integration create more threats and vulnerabilities in system security. For this reason, resources and services should be allowed to be access only by legitimate users. In some papers the authors mentioned that, the deployment of ubiquitous computing environment in real life is held back by weak and insufficient security measures, particularly, two security properties authentication and access control techniques.[9][10]

A. Authentication

Authentication can be defined as verifying whether the user identification is correct. Most traditional authentication methods cannot be applied as it is in ubiquitous computing environment. There are several reasons of why traditional authentication methods not fit? One of these reasons is; these methods cannot scale well with hundreds or thousands of embedded devices that placed in highly distributed environment such as ubiquitous computing environment. Another reason is; they are not convenient for users walking around within ubiquitous computing environment. Furthermore traditional authentication method that focus on identity authentication, possibly will fail to work in ubiquitous computing environment, since it conflicts with privacy protection which is one of the most important user's concerns in ubiquitous computing environment. Authentication in ubiquitous computing environment

requires different methods to cope with its different requirements, context and applications. Also authentication requirements are highly varied for different applications. [9][10]

There is still one issue that is considered as an essential security concerns and identified explicitly by a series of laws, it is the user privacy. Privacy is about protecting the personal confidential information, and that includes the invasion of personal space. Privacy preserving is another challenge that facing the deployment of pervasive computing services on large scale. Because in such environment with such requirements like focusing on invisibility of the computing devices that could be gathering information about users identities, their locations and users transactions, will make users concerns about their privacy. Beside, the integration of the physical world makes the task of preserving privacy in ubiquitous computing environment more difficult. [9][10]

B. Authentication methods in UCE:

In this section, in brief we describe some authentication devices that can be used integrated or alone in some Active Information spaces (Ubiquitous Computing Environment).

Some of authentication devices used in authentication framework in some Ubiquitous Computing Environment are [9]:

- *Active Badges*: In some Ubiquitous Computing Environment, each person has an active badge that can transmit the information of identity.
- *Smart Jewelry***:** People can ware Jewelry at all times, so it is harder to be stolen and does not necessitate a user to carry other gear. For that reasons, programmed jewelry can offer a convenient authentication method. The iButton is an example; it is a 16mm computer chip in a stainless steel case. Also it allows up-to-date information to move with user or object. The steel button is strong enough to resist insensitive outdoor environments,
- *Smart Watches:* A wristwatch is another wearable device that is worn by people almost all the day. A "smart" watch can be considered as an interactive wearable device, it provides a higher degree of security in authentication. In contrast to the previous wearable devices, smart watches store more information, have more processing power, have display features and make possible for user to interact with the device. Smart watch considered as secure authentication device because of these features make.
- *PDAs:* Larger PDAs are also used for authentication purposes as well as the wearable gadgets. The PDAs devices provide more feature i.e. more storage capacity and more processing power. Even as PDAs can be stolen or lost more easily than wearable devices like gadgets, they can be utilized to provide better authentication according to their processing, storage and interactive displays.
- *Passwords:* The traditional authentication method uses username and password pairs can be usable as a supplementary authentication method that can leverage other authentication methods.
- *Biometrics:* Biometrics could be used as an efficient mean of authentication. The users will be authenticated based on their distinctive physical characteristics, in order that users are identified according to "what they are." This may include retina, fingerprints, and face or voice recognition.

III. COMPARE PRIVACY PRESERVING AUTHENTICATION AND ACCESS CONTROL MODELS

In this section we give a brief overview of two models of preserving privacy authentication and access control models and compare between some of their security properties.

A. Model 1: A Flexible, Privacy-Preserving Authentication Framework for Ubiquitous Computing Environments

In this paper authors proposed general security framework, which builds over Kerberos and establish new enhancements that let it to blend nicely into pervasive computing environments, and identify general security requirements. They focused on designing specific infrastructure for security to protect user context privacy from the service providers. They have used MIST infrastructure which provides anonymity for user through an overlay network also it keeps all information of all the users using what they call "Lighthouse". [9]

The authentication framework is able to scale to highly distributed system, while providing convenient and flexible authentication and access control services for ubiquitous computing environment. Also it uses different embedded and wearable devices and authenticates users in a convenient, transparent and private approach. In this model users can authenticate themselves to the system using a multiple means furthermore some of these means are reliable more than others and could provide stronger authentication than others. The strength of these means deduced from the assigning varied confidence values to variety authentication methods. This value is considered as a measurement of how the system is confident that the user, who just authenticates himself using some means, is definitely who he claims himself to be. The overall level of confidence increase when a user uses more than one authentication method. In this model a Confidence-builder module has been established to handle the use of several authentication methods by employing algorithm for combining many confidence values, and then produce net confidence value. [9] .

In this model, designers make the Active Space able to detect the presence of users and objects actively. They believe that, these features are necessary to make spaces active and to enable context-based applications. So they used a method that allows users to be authenticated to the surrounding environment and simultaneously preserve their

privacy. They establish Mist which is communication infrastructure in ubiquitous computing environments that preserves location privacy, while allowing users and objects to be authenticated at the same time [9].

- *Mist:* consists of a hierarchy of *Mist Routers* that structure an overlay network. This network allows private communicate for users. The Mist Routers route packets using a hop-by-hop, handle-based routing protocol with limited encryption using public key cryptography, as a consequence, the communication become untraceable by eavesdroppers. .

- *Authentication Protocol*: authentication protocol in this model extends Kerberos authentication protocol to support user devices and make use of the location privacy that provided by Mist. In each Active Space, they assume the existence of "*Space Authentication Portals*" (SAPs), which are special types of Portals that could be located at the Active Space entrance, or other suitable places. The SAP will feature a set of wired and wireless base stations and device readers that allow users to be authenticated with the Active Space using any authentication devices they are wearing or carrying.

In this model, all users have active badges. The badge programmed to store unique ID number, for user identification, and store ID for user's Lighthouse identification. Then user comes close to one of the available SAPs for authentication. Some of the authentication devices possibly will require the intervention of user, e.g. insert the iButton into the corresponding designated receptor. Then the communication is done through Mist communication, so the lighthouse communicates with the Security Server. Note that SAP does not have sufficient information for user's authentication. Upon authentication success, the AS, like Kerberos protocol, produces a ticket granting ticket (TGT) for that user. The TGT is issued for a user is encrypted and stored in the users Lighthouse. The AS remembers the user's previous authentication methods and then calculates the net confidence of all authentication methods of the user being there to issue new TGT with the new value. After that, the user can access the service, but the service needs to check the user first by contacting with the user's Lighthouse. Using the TGT that are stored in the Lighthouse of user, the Lighthouse will communicate with the TGS and request for tickets to access the requested service. These tickets are encrypted and do not contain any indications to the real identity or name of the user; they incorporate a pseudonym. Also, they contain the net confidence level and the security privileges of the user, so the service can make access control decision whether to authorize that user or not. When the user exit from the room the badge reader at the exits can discover that and automatically it will log off the user and destroy the stored tickets in his Lighthouse. [9]

B. Model 2: A Novel Privacy Preserving Authentication and Access Control Scheme for Pervasive Computing Environments

In this model, the authors propose a scheme to secure the interactions between services and mobile users in ubiquitous computing environment. The scheme integrates two fundamental cryptographic primitives; they are the hash chain and the blind signature, into authentication protocol.

We provide brief description of the two techniques, as follows:

- *Blind Signature*: The blind signature is one of the digital signature variations where the message content disguised from the signer. It can be implemented based on some well-known digital signature schemes. a user first use a random "blinding function" *f*, to "blinds" the message before sign it from third party. So the signer will sign the message without having any idea about its content, and then send it back to user. The user unblinds the message and obtains the signature on the original message. Blind signature used for nonlinkability property, and this property is helpful when anonymity is required. [10]
- *Hash Chain:* also called one-way hash function is one of the powerful cryptographic tools; it takes a message of any size as input and outputs a fixed size hash. A chain of hash outputs can be obtained by applying repeatedly on an initial message. And the outputs of hash can be used in the reverse order of generation for authentication purpose. [10]

Sample system architecture of a ubiquitous computing environment, generally, consists of three types of entities: the Mobile users, the Services and the Back-end authentication servers, besides, the underlying wireless and wired communication infrastructures. While the wireless network access is a service by itself. Protecting the user privacy includes protection from the outsiders and from the network service providers. The proposed access control in this model is designed to secure the interactions among these three types of entities. [10]

The design considerations in this model include:

1) Providing precise mutual authentication between the service and the mobile user;

2) Allowing mobile users to interact with the service anonymously

3) Enabling differentiated service access control, by classifying mobile users into different service groups.

4) Providing scalability and flexibility to both service and user sides.

5) To secure the interaction, generate fresh session keys.

6) Having high effectiveness in terms of computation, management, and communication overheads.

7) Providing simple accountability.

8) Providing correctness verification in formal manner, based on Burrows–Abadi–Needham logic. [10]

The model in this paper consists of two protocols: user authorization protocol and user operational protocol. The steps of the two protocols are described below but first notice Table I which lists the notations used all through the description of the protocols for ease of reference:

User authorization protocol*:* The point of the user authentication protocol is to launch security credentials between service providers and mobile users, which can be used in the later mutual authentication processes at any time a mobile user wants to access a service.

User operational protocol*:* The user operational protocol permits the mobile user to have the benefits of different types of services safely which is authorized to in ubiquitous computing environments from anywhere in anytime without disclosing any of his context information unless he is prepared to do so and it is absolutely needed.

Comparison

We compare the security properties between model1 (A Flexible, Privacy-Preserving Authentication Framework for Ubiquitous Computing Environments) and model2 (A Novel Privacy Preserving Authentication and Access Control Scheme for Pervasive Computing Environments) in the Table I.

C. Results and Discussion

From the TableII we conclude the following results:

- Model 2 provides Mutual authentication, Concrete protocol while model 1 dose not.
- Both models provide User context privacy and Differentiated service access control.
- Both models use Encryption and Digital Signature to achieve confidentiality and integrity.
- Both models have been proved.

1-Mutual authentication: means that both parties have to be authenticated to each other. In Model2, the mobile user is authenticated, based on their authorized credential, to the service. The service also authenticates itself to the user. While Model1, does not provide this property. The mutual authentication property is essential in ubiquitous computing environments, since it prevents the potential malicious attacks from both sides.

2-Concrete protocol: Model1 protect user context privacy by designing specific security infrastructures and identifying general security requirements, but does not provide a concrete security protocol. On the other side, Model2 provides a concrete security protocol.

3-User context privacy The users' context privacy is protected in both models; only the necessary information is recognized by the service, to grant proper access. In Model1 the user context privacy is preserved using Mist Infrastructure with Lighthouse. In Model2, the blind signature technique is responsible for authenticate users anonymously to provide context privacy. So while protecting the context privacy the user could be authenticated without disclosing any context information.

Table I
MODELS SECURITY FEATURES COMPARISON

Security Property	Model1	Model2
Mutual authentication	No	Yes
Concrete protocol	No	Yes
User context privacy	Yes	Yes
Differentiated service access control	Yes	Yes
Integrity	Yes (Mist Communication)	Yes
Confidentiality	Yes (Mist Communication)	Yes
Provable security	Yes (employed in Gaia research project)	Yes (BAN logic)

4-Differentiated service access control: allows different users to access different services according to their privileges. In Model1 the concept of multiple authentication confidence levels are used by services in access control decisions. In Model2 differentiated service access control is enabled by means of classifying the mobile users into different types of service. Based on the types of service different mobile users are authorized according to which one they belong. Therefore user authorization is achieved in a differentiate way.

5-Integrity and Confidentiality: Using encryption and digital signature can provide confidentiality and integrity respectively. In Model1 and Model2, Encryption and digital signatures are used in many aspects aiming to provide confidentiality and integrity protection for the communications between the service and the mobile user. Also in Model1, we found that confidentiality is achieved using Mist Infrastructure. [11] It also provides confidentiality and integrity protection for the communications between the mobile user and the service.

6-Provable security:

The model has to be secure against both passive and active attacks and this security has to be verified. According to the important of verifying the correctness of model, we check this property for both models. Model1 have been employed in the Gaia research project in real life and the security is

proved. Also, Model2 correctness has been verified using BAN logic.

IV. CONCLUSION

In this paper we have presented Security issues in Ubiquitous Computing Environments. And we go through some of these issues; which are Access Control and Authentication. Traditional access control models can not be used in UCEs because of the large number of users in this environments and also because UCEs have a relation with time. Trust based access control is the best model that can be used with UCEs. We propose a simple ubiquitous computing environment and we write a pseudo code foe our proposed environment. Then, we show how user privacy could be great risk in UCEs and how the conflict between user privacy and authentication makes security design in UCEs a challenging task. Finally, we display and compare two models that preserve location and context privacy for authenticated users.

REFERENCES

[1] M. Weiser, " The Computer of the 21st Century," *Scientific American*, vol. 265, no. 3, Sept. 1991, pp. 66–75 (reprinted in this issue, see pp. 19–25).

[2] F. Stajano, "Security for whom? the shifting security assumptions of pervasive computing", in *Software security: theories and systems*,LNCS 2609, M. Okada , B. Pierce , A. Scedrov, H. Tokuda , and A. Yonezawa Eds. New York:Springer, 2003, pp.16- 27.

[3] G. Sweden, "Workshop on Security in Ubiquitous Computing", in Proc. UBICOMP, 2002. http://www.teco.edu/~philip/ubicomp2002ws/proceedings.htm

[4] F. Stajano, "Security issues in ubiquitous computing," in *Handbook of Ambient Intelligence and Smart Environments* H. Nakashima , H. Aghajan and J. Augusto Eds. New York :Springer, 2009, pp. 281-309.

[5] F. Stajano, "The Security Challenges of Ubiquitous Computing", invited talk in CHES 2003.

[6] G. Ya-Jun, H. Fan, Z. Qing-Guo, and L. Rong, " An Access Control Model for Ubiquitous Computing Application ", In: Proc. of 2nd International conference on Mobile Technology, Applications and Systems, **IEEE** conferences, Nov. 2005.

[7] L. Cao, " A Flexible, Autonomous and Non-redundancy Access Control for Ubiquitous Computing Environment", In: Proc. Of International Symposium on Information Science and Engineering (ISISE '08), IEEE conferences, Dec. 2008, pp 446 – 450.

[8] C. Miao and L. Chen," Trust-based dynamic access control policy for ubiquitous computing", In: Proc. of 3rd IEEE International Conference on Ubi-media Computing (U-Media), IEEE conferences, Augt. 2010, pp 277 – 281.

[9] J. Al-Muhtadi, A. Ranganathan, R. Campbell, and M. Mickunas, "A flexible, privacy-preserving authentication framework for ubiquitous computing environments," in *Proc. ICDCS Workshops*, 2002, pp. 771–776.

[10] K. Ren, W. Lou, K. Kim and R. Deng, "A Novel Privacy Preserving Authentication and Access Control Scheme for Pervasive Computing Environments", vol. 55, no. 4, IEEE Transactions on Vehicular Technology, 2006, pp. 1373-1384.

[11] J. Al-Muhtadi, R. Campbell, A. Kapadia, D. Mickunas, and S. Yi, "Routing through the mist: Privacy preserving communication in ubiquitous computing," in *Proc. ICDCS*, Vienna, Austria, 2002, pp. 65–74.

Wireless Overlay Sensor Networks

Sumesh J. Philip
School of Computer Sciences
Western Illinois University
Macomb IL 61455

***Abstract*—Wireless Sensor Networks (WSN) have been in the forefront of distributed autonomus network research in recent years. A majority of the research in this area has used the physical topology of the radio network as the underlying platform for understanding the many issues in WSNs. Unlike traditional networks where an individual node is assumed to be a reliable component of the network, a single node in a wireless sensor network may not hold much importance. Thus, a WSN model in which a collection of nodes that cooperate with each other to carry out a specific network function is a more suitable model in terms of utilizing redundancies in sensing, data aggregation and forwarding. To this end, we have introduced the concept of Overlay Sensor Networks (OSN), which treats the sensor network as an abstract set of vertices and edges based on the position of nodes in the network. In a past work, we showed that planar graph routing in an an Overlay Sensor Network performs better than routing on the physical network topology. In this work, we would like to investigate the possibility of using traditional routing protocols (such as Distance Vector) in an OSN. Answers from this study will help us better understand the tradeoffs in using such abstractions for efficient operation of a Wireless Sensor Network.**

I. Overview

A Wireless Sensor Network (WSN) is an autonomous, self organizing set of nodes in which the nodes have the individual capability of sensing/monitoring the ambient environment in which they are deployed. Individual nodes are equipped with sensor/actuators for environment monitoring, microcontrollers for processing, memory (RAM, flash) for running programs and recording data, radio transciever for communication and a power source for operation [1]. Due to limitations in sensing and transmission ranges of an individual node, a collection of such nodes operating in a cooperative manner is better suited for monitoring large terrains or widespread phenomena. Many applications have been envisaged for WSNs such as monitoring environmental changes, emergency/crisis management, medical observations, logistics and transportation, security and smart spaces. Tremendous research has been carried out in order in this area to better understand the inherent characteristics of WNSs and the challenges to overcome before WNSs become a reality.

A majority of the research in WSNs have used the physical topology of the WNS as the underlying platform to study issues such as routing [2]. However, unlike a traditional network, in which an individual node is an integral component of the network, an individual node in a WSN may not be as important. Nodes are expected to fail due to hazardous conditions, or be inoperable due to lack of battery power etc. Despite such failures, the network is required to carry on its function, relying on the inherent redundancy available via the remnant nodes. In order to take advantage of such redundancy, we use the notion of an abstract topology using the collective node positions to model the WSN, and call it an Overlay Sensor Network (OSN). An OSN is defined as follows:
Given a set of sensor nodes N and their Euclidean co-ordinates, the network terrain is tiled into an uniform infinite grid such that the unit square regions within the grid are of diameter r, where r is the radio range of a node. The center point (x_i, y_i) of each unit region R_i uniquely represents the unit region and forms an abstract vertex in the overlay network. A total ordering on the vertices are defined as follows:

- $(x_1, y_1) < (x_2, y_2)$ if $x_1 < x_2$ or $x_1 = x_2$ and $y_1 < y_2$
- $R_i < R_j$ if $(x_i, y_i) < (x_j, y_j)$

A sensor node $u \in N$ can now be uniquely assigned to one of the vertices by defining a total ordering on the node locations on the grid using the following three rules:

- A node $u(x_u, y_u)$ is assigned to R_i if $(x_u, y_u) \in R_i$
- A node $u(x_u, y_u)$ is assigned to min(R_i, R_j) if $(x_u, y_u) \in R_i, R_j$
- A node $u(x_u, y_u)$ is assigned to min(R_i, R_j, R_k, R_l) if $(x_u, y_u) \in R_i, R_j, R_k, R_l$

The last two rules assign nodes to a unique vertex if they fall on the boundary of two unit regions or the corner of four unit regions.

Once the vertices are defined, the adjacency matrix of the overlay network $G(V, E), R_i \in V$ is defined as follows: two vertices R_i and R_j are adjacent, and a bidirectional edge $e_{R_i \to R_j} \in E$ exists between them if there is a set of sensor nodes u, v $(u, v \in N, u \in R_i,\ v \in R_j)$ that are directly connected to each other in the physical topology. Due to this construction, a vertex R_i can be adjacent to at most 20 other vertices as shown in figure 1.

In a past work [3], we showed that planar graph routing on such an abstraction outperformed routing on the physical topology. Main reasons for the performance improvement were a) routing from region to region places less emphasis on individual nodes and the actual topology, leading to greater recovery from node failures, and b) quick exit from the current region as soon as the absence of the destination in that region is known, leading to faster delivery of packets. With these positive results, we are interested in finding out if the overlay concept may be useful in other scenarios as well. Specifically, we are interested in the performance of Distance Vector (DV)protocols in OSNs. Studies in the past have shown that DV protocols modified for wireless networks

This paper was presented as a poster paper.

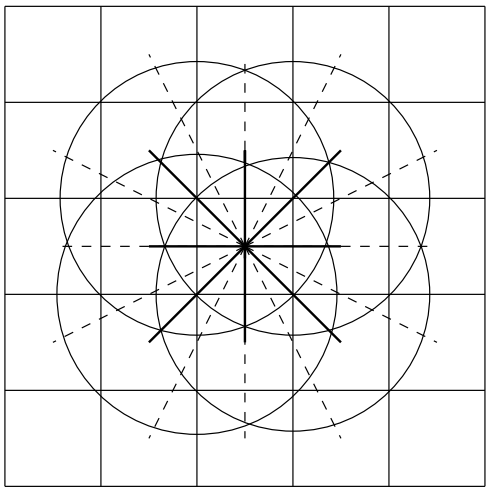


Fig. 1. Example of a vertex in the OSN. An OSN vertex is adjacent to at most 20 other vertices.

(such as WRP) perform rather poorly when links/nodes fail in the network. This is predominantly due to routing loops created by the famous "Counting-to-infinity" problem inherent to DV protocols [4]. Although this a distinct disadvantage, DV protocols are appealing in wireless networks due to their localized nature of operation. Nodes only need to communicate with their immediate neighbors for topology inference, thereby saving on route updates, channel contention, battery power etc. In addition, the destination of sensor node packets being Base Stations (BS), and since the number of such BSs are much smaller than the number of sensor nodes, this will amount to a small number of routing entries in a node's routing table, in contrast with storing route information to the entire network. The objective of this research will try to answer the following questions:

- What modifications do DV protocols need to operate in an OSN?
- Can we find optimal routes to BSs without needing complete topology information of the WSN?
- Can we prevent routing loops in the network due to neighbors advertising obsolete routes? If so, can it done with partial topology information exchange between neighbors?

Finding an optimal route to a BS without having the WSN topology information is trivial in DV protocols since each node chooses the best option from amongst the routes presented by each of its neighbors, and advertises the choice further down the network. However, knowing if the optimal route is still valid in the presence of node/link failures is not so easy. Modifications to DV protocols to advertise the predecssor information (for e.g. WRP) of the path is one possible solution to this problem [5]. However, this requires each node to maintain and advertise their chosen routes to **all** nodes in the network, and not just the routes to the BSs. This would invalidate the savings obtained from just advertising routes to BSs which was one of the appealing features of DV protocols tailored for WSNs as we noted earlier.

Our solution to this problem would be to assign a unique label to each edge in the OSN, and use edge labels to create a labelled path for each unique path from a node to a BS in the OSN. One of the benefits of doing this in an OSN (instead of the WSN topology) is the limited set of edges (at most 20) in the OSN compared to the WSN. In addition, while one cannot predict the neighborhood of a node in the WSN before deployment (unless the deployment is planned), the planned neighborhood of a vertex in the OSN is fixed a priori. The actual OSN topology that results from the WSN deployment can only remove edges from an OSN vertex neighborhood (due to lack of nodes in an adjacent region), but never add to it. This simplifies the label distribution problem in an OSN significantly compared to that of edge label assignment in a WSN topology.

Each sensor node broadcasts periodic beacons to their neighbors that contain their locations, as well as its OSN vertex ID. This allows each node to discover its OSN neighborhood as well as its neighbors in the WSN topology. In order to find routes to the BSs in the WSN, we will assume that the BSs will also periodically send out beacons to announce their presence in the network. Nodes that receive a beacon from a BS will add an entry in their routing table for the BS. The route to the BS is identified by an integer label l_{ij} which is computed from the OSN edge label connecting the sensor node's OSN vertex R_i to that of the BS's OSN vertex R_j. The route will then be advertised to the node's neighbors with the route label, and the cost to reach the BS. A neighbor node *u* that receives an update from node *v* processes it as follows:

- If *u* has no route to this BS or if the route cost is better than the entry in the routing table, it adds/updates an entry in its routing table by computing the path label as
  $l_{u \to BS} = f(l_{uv}, l_{v \to BS})$
  where $f(x, y)$ is a monotonically increasing function which preserves the individual label values. *u* then readvertises the route to its neighbors.
- Otherwise, *u* discards the update from *v*

We will show that the protocol will result in nodes finding an optimal path to the BSs without routing loops.

Node failures or link failures in the network can result in routes being invalidated during the operation of the network. Periodic beacons allow nodes to determine if a neighbor is alive/dead, and we will use this feature to detect if an existing route has failed or not. If a node i in R_i detects the failure of a route due to a failed neighbor j in R_j, it triggers a route deletion phase. This will be done by advertising the failed route, along with the label l_{ij} which indicates the link associated with the route failure. Nodes that receive this update will use l_{ij} to check the path label to the BS in their routing table. Since the label function $f(x, y)$ preserves each edge label in path, nodes can discover if the route deletion phase affects existing routes in their routing table. If the failed link affects any of the routes in a node's table, it deletes those routes and readvertises them to its neighbors. This continues till the updates reach all the nodes, or a node whose table is not affected by the link failure. This node then updates its neighbors with the valid route, and allows nodes to rebuild the route similar to the route formation phase. Note that this protocol does not require nodes to keep table entries for all

the nodes in the network, but only to the BSs. We will show that this protocol will effectively deal with the "Counting-to-infinity" problem in DV protocols with nodes needing only partial network information unlike previous solutions. Simulation studies will be used to see the practical aspects of the protocol in terms of protocol effectiveness, overhead, delay and other quantitative metrics.

REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, March 2002.

[2] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad hoc networks*, vol. 3, no. 3, pp. 325–349, 2005.

[3] Sumesh J. Philip, Joy Ghosh, Hung Q. Ngo and Chunming Qiao, "Routing on Overlay Graphs in Mobile Ad hoc Networks," *IEEE Globecom*, Nov. 2006.

[4] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach, 6/E*. Pearson Education, 2012.

[5] S. Murthy and J. Garcia-Luna-Aceves, "A Routing Protocol for Packet-Radio Networks," *Proceedings ACM/IEEE MobiCom*, November 1995.