International Journal of Communications, Network and System Sciences

ISSN: 1913-3715

Volume 2, Number 2, May 2009

www.scirp.org/journal/ijcns/

Scientific
Research
Publishing

JOURNAL EDITORIAL BOARD

ISSN 1913-3715 (Print)   ISSN 1913-3723 (Online)
Http://www.scirp.org/journal/ijcns/

***I. J. Communications, Network and System Sciences*, 2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 2 Number 2 May 2009

International Journal of Communications, Network and System Sciences (IJCNS)

Journal Information

SUBSCRIPTIONS

The *International Journal of Communications, Network and System Sciences* (Online at Scientific Research Publishing, www.SciRP.org) is published quarterly by Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603-3334, USA.

E-mail: ijcns@scirp.org

Subscription rates: Volume 2 2009
Print: $50 per copy.
Electronic: free, available on www.SciRP.org.
To subscribe, please contact Journals Subscriptions Department, E-mail: ijcns@scirp.org

Sample copies: If you are interested in subscribing, you may obtain a free sample copy by contacting Scientific Research Publishing, Inc at the above address.

SERVICES

Advertisements
Advertisement Sales Department, E-mail: ijcns@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc. 5005 Paseo Segovia, Irvine, CA 92603-3334, USA.
E-mail: ijcns@scirp.org

COPYRIGHT

Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assumes no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ijcns@scirp.org

I. J. Communications, Network and System Sciences, 2009, 2, 91-168
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Cooperative Self Encoded Spread Spectrum in Fading Channels

Kun HUA, Won Mee JANG, Lim NGUYEN
University of Nebraska-Lincoln, Omaha, NE, USA
Email: {*khua, wjang, nguyenl*}*@unlnotes.unl.edu*

Received November 15*,* 2008; *revised February* 26*,* 2009; *accepted March* 3*,* 2009

Abstract

Self-encoded spread spectrum (SESS) is a unique realization of random spread spectrum. SESS eliminates the need for the traditional transmitting and receiving PN code generators. Instead, the time varying, random spreading sequence is obtained from the data source. Cooperative diversity (CD) has been attracting increasing attention as a novel and promising diversity technique. This paper analyzes the cooperative SESS for Amplify and Forward CD links in Rayleigh channels. The results show that our cooperative SESS improves the system performance significantly over MRC-based cooperative systems.

Keywords: Cooperative Diversity, Spread Spectrum, Maximum Ratio Combiner

1. Introduction

Cooperative diversity receives increasing attention as a diversity enabler, whereby several partner terminals around a given mobile terminal form a distributed cooperative network and transmit information collaboratively [1]. The advantages of CD are similar to existing diversity technique like MIMO to combat the detrimental effects of multipath fading. Sendonaris [2,3] has proposed a user cooperation model that achieved an increase in capacity. As spread spectrum can effectively deal with multipath fading, direct-sequence spread spectrum transmissions have been considered for implementing a novel spectrally efficient cooperative protocol [4]. SESS is a unique random spread spectrum that eliminates the need for traditional transmit and receive PN code generators [5]. In this paper, we consider SESS cooperative diversity (SESS-CD) communication over fading channels and analyze its performance in fading channels. Expressions for the average bit error rate (BER) are derived and the result is compared with the repetition scheme with maximum ratio combiner (MRC). The mobile radio channel suffers from multipath fading, implying that, within the duration of any given call, mobile users could experience severe variations in signal attenuation. Spread spectrum and diversity are methods for combating the detrimental effects of fading. Iterative detection with SESS-CD receiver is shown to achieve remarkable performance improvement reducing the BER significantly. SESS-CD with iterative detection provides both temporal and spatial diversity while MRC exploits only spatial diversity gain.

In Section 2, we describe the system model. Section 3 analyzes the performance of SESS-CD and MRC. The analytical and simulation results based on SESS-CD schemes are presented in Section 4. The conclusion follows in Section 5.

2. System Model

Consider the cooperative network where information is communicated between a source (S=R_1) and a destination (D=R_0) over a complex channel with fading parameter f_{10}. Two relay nodes, R_2 and R_3, are willing to cooperate to provide repeated signals through the complex channels with flat fading channel parameters (f_{12}, f_{13}) from (S) to (R_2, R_3), and (f_{20}, f_{30}) from (R_2, R_3) to (D), respectively. Without loss of generality, we assume the relays and destination have the same additive white Gaussian noise (AWGN) power. We also assume that the values of random variables, f_{10}, f_{12}, f_{13}, f_{20} and f_{30} have been determined at the receiver ends by training. We consider the

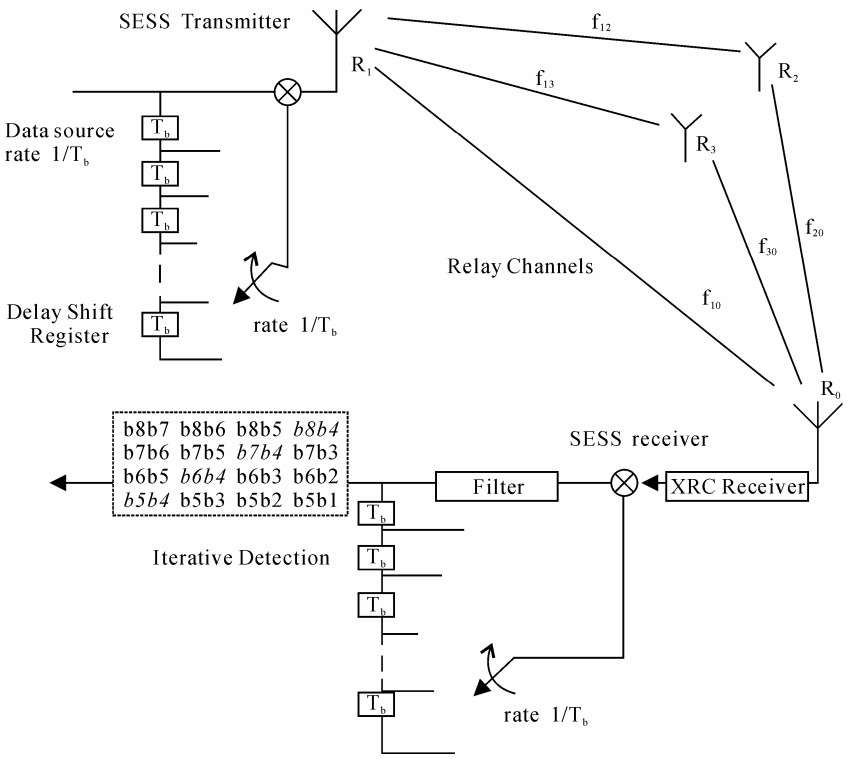


Figure 1. Cooperative self-encoded spread spectrum structure.

Amplify and Forward (AF) model with a constant average power. The basic idea in our proposed spatially cooperative spread spectrum is to implement SESS across a cooperative relay network. Figure 1 shows the block diagram of SESS-CD system. At the transmitter, the delay registers are constantly updated from *N*-tap serial delay of the data to generate the spreading sequence of length *N*. The current bit is spread by the time varying *N* chip sequence that has been obtained from the previous *N* data bits [6]. The SESS data bit will be transmitted through the direct and relay paths simultaneously with different fading coefficients as shown in Figure 1. The self-encoding operation at the transmitter is reversed at the receiver. The recovered data are fed back to the *N*-tap delay registers that provide an estimate of the transmitter spreading code required for signal de-spreading. The SESS-CD receiver employs iteration decision. The receiver thus exploits the additional time diversity as well as the spatial diversity inherent in relay systems. The transmitted signal can be expressed as:

$$x=d_iS_i \tag{1}$$

where d_i and S_i are the data bit and the SESS spreading sequence, respectively, during *i*-th bit duration. In MRC scheme, x is a simple data bit. Let the fading amplitude be f_{ij} with the corresponding mean of K_{ij}. Then, the received signals can be expressed as:

$$y_1 = f_{10}x + n_1 \tag{2}$$

$$y_2 = f_{20}A_2(f_{12}x + n_{r2}) + n_2 \tag{3}$$

$$y_3 = f_{30}A_3(f_{13}x + n_{r3}) + n_3 \tag{4}$$

where n_{ri} is the noise at the relay, and n_i is the noise at the destination. n_{ri} and n_i are statistically independent Gaussian noise which is distributed as $N(0,\sigma_0^2)$, where we assume the same noise power σ_0^2 at relays and the destination. A_2 and A_3 are amplification factors to maintain constant average power output of the relays:

$$A_2 = \sqrt{(E_b/N_o)/(f_{12}{}^2(E_b/N_o)+1)}$$

$$A_3 = \sqrt{(E_b/N_o)/(f_{13}{}^2(E_b/N_o)+1)} \tag{5}$$

Then, the output of the decorrelator at the receiver is given by

$$r_i = \psi_1 y_1 S_i^* + \psi_2 y_2 S_i^* + \psi_3 y_3 S_i^* \tag{6}$$

where S_i^* is the recovered spreading sequences at the receiver, which may be different from S_i due to detection errors. ψ_1, ψ_2 and ψ_3 are the normalization factors for fading and noise power:

$$\psi_1 = f_{10}/\sigma_0{}^2$$

$$\psi_2 = f_{20}A_2 f_{12}/((f_{20}{}^2 A_2{}^2 + 1)\sigma_0{}^2)$$

$$\psi_3 = f_{30}A_3 f_{13}/((f_{30}{}^2 A_3{}^2 + 1)\sigma_0{}^2) \tag{7}$$

We can write SESS signals as

$$S=\underbrace{\begin{bmatrix} S_1= & d_0 & d_{-1} & d_{-2} & \cdots & d_{-N+1} \\ S_2= & d_1 & d_0 & d_{-1} & \cdots & d_{-N+2} \\ S_3= & d_2 & d_1 & d_0 & \cdots & d_{-N+3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_N= & d_{N-1} & d_{N-2} & d_{N-3} & \cdots & d_0 \\ S_{N+1}= & d_N & d_{N-1} & d_{N-2} & \cdots & d_1 \end{bmatrix}_{[(N+1)\times N]}}_{spread\quad sequence\quad block} \tag{8}$$

where d_i are the data bits delayed to form the SESS spreading sequences. Since the current bit is spread by N previous bits, we can observe that current detecting bit d_1 is also related to previous N information bits, which are stored in the delay shift register $d_{-N+1},\ldots, d_0$. By incorporating previous detected bits, we expect to improve the performance. Therefore signal energy can be retrieved from previous estimated bits(c_i) as

$$\xi_i=\sum_{k=1}^{N} r_{i-k}c_{i-k} \tag{9}$$

and the bit decision can be made based on

$$Y_i=r_i+\xi_i \tag{10}$$

For MRC scheme, we obtain

$$Y_i=\psi_1 y_1+\psi_2 y_2+\psi_3 y_3 \tag{11}$$

at the receiver for bit detection. We assume that each relay path and direct path are isolated. The isolation can be achieved by time division multiplexing.

3. Performance

1) BER for Relay Channel (MRC): As shown in Figure 1, $f_{10}, f_{12}, f_{13}, f_{20}$ and f_{30} are the fadings on the relay and direct paths. Let the mean and the second moment (power) of the fading, f_{ij} are equal to K_{ij} and ζ_{ij}, respectively. Then, the signal-to-noise ratio (SNR) at different nodes can be calculated as:

$$\gamma_{ij}=\xi_{ij}\frac{P_x}{N_o} \tag{12}$$

where P_x/N_o is the received SNR in AWGN channels without fading. The SNR at receiver with diversity can be derived from [1] as

$$\gamma_z=\sum_{k=2}^{3}\frac{\gamma_{1k}\gamma_{k0}}{1+\gamma_{1k}+\gamma_{k0}}+\gamma_{10} \tag{13}$$

which is reduced to

$$\gamma_z=\sum_{k=2}^{3}\frac{\gamma_{1k}\gamma_{k0}}{\gamma_{1k}+\gamma_{k0}}+\gamma_{10}=\sum_{k=2}^{3}\frac{1}{\frac{1}{\gamma_{1k}}+\frac{1}{\gamma_{k0}}}+\gamma_{10} \tag{14}$$

at high SNR. In MRC cooperative scheme, information bits are repeated in relay paths. We assume binary phase-shift keying modulation (BPSK) over Rayleigh fading channels. Therefore, the bit error rate with M relay branches is [1]:

$$P_e\approx\frac{C(M)(K+1)^{M+1}}{k^{M+1}}\frac{1}{\gamma_{10}}\prod_{m=1}^{M}\left(\frac{1}{\gamma_{1m}}+\frac{1}{\gamma_{m0}}\right) \tag{15}$$

where K denotes the factor in non-central Chi-squared distribution, and K=0 for exponential distribution. The constant k depends on the type of modulation, and k =2 for phase shift keying. $C(M)$ can be obtained as

$$C(M)=\frac{\prod_{k=1}^{M+1}(2k-1)}{2(M+1)!} \tag{16}$$

If the relay nodes number M=2, then

$$P_e=\frac{5}{32}\left(\frac{1}{E_b/N_o}\right)^3\left(\frac{1}{\xi_{12}}+\frac{1}{\xi_{20}}\right)\left(\frac{1}{\xi_{13}}+\frac{1}{\xi_{30}}\right)\frac{1}{\xi_{10}} \tag{17}$$

We observe that the error probability P_e is the function of $(E_b/N_o)^{-(M+1)}$ where M is the number of relay nodes. Therefore, the cooperative network can achieve the full diversity order of M +1.

2) BER for Self-encoded Spread Spectrum Cooperative Diversity (SESS-CD): The performance of SESS-CD with iterative detection can be considered as

$$P_e=\int_0^\infty Q(\sqrt{k\gamma_z})p_{\gamma_z}(\gamma_z)d\gamma_z \tag{18}$$

where $p_{\gamma_z}(\gamma_z)$ is the probability density function of γ_z. In this cooperative SESS-CD performance analysis, we do not consider the self-interference that comes from the erroneous despreading sequences due to the incorrect bit decision at the receiver. The self-interference was shown to be dominant at low SNR or with small spreading factors [7]. The received energy in each path can be considered as

$$y=\alpha_0+\sum_{i=1}^{N}\alpha_i \tag{19}$$

where α_i for i=1, ...,N is an exponential random variable (r.v) with parameter $1/\gamma_c$, i.e.,

$$p_{\alpha_i}(\gamma)=\frac{1}{\gamma_c}\exp\{-\gamma/\gamma_c\} \tag{20}$$

where γ_c is the chip energy to noise ratio with fading. The α_0 is an exponential r.v. with parameter $1/\gamma_c$. The first term in Equation (19) is the output of the current bit despreading and the second term is the iterative detection output. We apply the central limit theorem to find the approximate probability density function (pdf) of y. Since the mean and variance of α_i, for i = 1,...N, is γ_c and γ_c^2, respectively, we can approximate the mean and variance of *y* in Equation (19) as

$$m_y = N\gamma_c + N\gamma_c = 2N\gamma_c \quad \textbf{(21)}$$

$$\sigma_y^2 = N^2\gamma_c^2 + N\gamma_c^2 = N(N+1)\gamma_c^2 \quad \textbf{(22)}$$

Therefore, the approximate pdf of the r.v. *y* can be obtained as

$$p_y(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\{-(y-m_y)^2/(2\sigma_y^2)\} \quad \textbf{(23)}$$

Since the first term in Equation (19) is a dominant term, Equation (23) may not be the best approximation. However, we will find that the result can provide a useful insight regarding the SESS-CD diversity gain. For high SNR, $p_y(0)$ tends to be zero. Therefore, we will find the $\partial p(0)/\partial y$ to be applied to the initial value theorem of Laplace Transforms [1] as

$$\frac{\partial p(0)}{\partial y} = \frac{1}{\sqrt{2\pi\sigma_y^2}} \frac{m_y}{\sigma_y^2} \exp\{-m_y{}^2/2\sigma_y^2\} \quad \textbf{(24)}$$

$$\approx \sqrt{\frac{2}{\pi}} \exp\{-2\} \frac{1}{N^2} \frac{1}{\gamma_c^2} = \sqrt{\frac{2}{\pi}} \exp\{-2\} \frac{1}{\gamma_b^2}, \quad \text{for large } N \quad \textbf{(25)}$$

where γ_b is the bit energy to noise ratio with fading. The SNR at the different nodes can be represented as γ_{ij}. With *M* cooperating branches, the probability of bit error with BPSK can be obtained as

$$P_e \approx \frac{C(M)(K+1)^{2(M+1)}}{k^{2(M+1)}a^{M+1}} \frac{1}{\gamma_{10}^2} \prod_{i=1}^{M} \left(\frac{1}{r_{1i}^2} + \frac{1}{r_{i0}^2}\right) \quad \textbf{(26)}$$

where $a = (\sqrt{2/\pi}\exp(-2))^{-1}$ from Equation (25). *C*(*M*) can be obtained as

$$C(M) = \frac{\prod_{k=1}^{2(M+1)}(2k-1)}{2(2(M-1))!} \quad \textbf{(27)}$$

Comparing Equations (15) and (26), we find that the effective SNR in SESS-CD with iterative detection is the square of the actual SNR.

4. Simulations and Numerical Results

In Figure 2, we can see that the performance of SESS-CD is superior to MRC. The result can be predicted from Equations (15) and (26). The BER difference between SESS-CD simulation and analysis comes from the gaussian approximation of the received signal power. The exact pdf and its gaussian approximation of the received signal power over random fading channels are shown in Figure 3. We can observe that the gaussian approximation shifts the probability of low received signal power to high received signal power at both E_b/N_o equal to 5 dB and 10 dB, while maintaining the same mean and variance as the exact pdf. However the slope of SESS-CD simulation BER and analytical BER agrees well. The diversity gain determines the slope of the BER versus average SNR curve, at high SNR, in a log-log scale. On the other hand, coding gain (in decibels) determines the shift of curve in SNR relative to the benchmark BER curve in uncoded communication over a random fading channel [8]. We see that the Gaussian approximation exhibits a rather accurate diversity gain but not coding gain. The diversity gain in Figure 2 portrays well the square term of the SNR enhancement in SESS-CD in Equation (26). Figure 4 shows the performance of

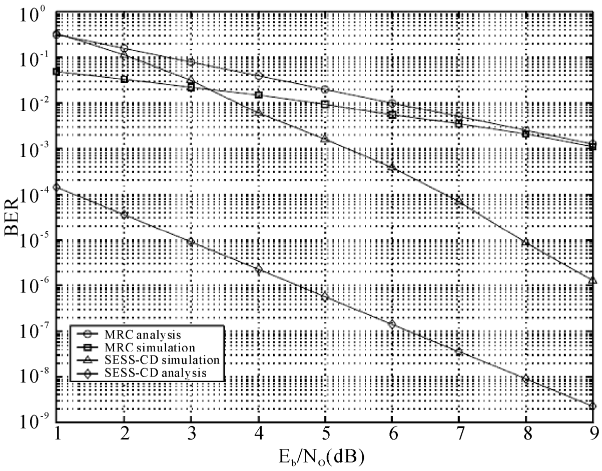


Figure 2. Simulation BER, SESS-CD (64 chips/bit) and MRC, K_{10}=K_{20}=$_{30}$=1, K_{12}=K_{13}=1.

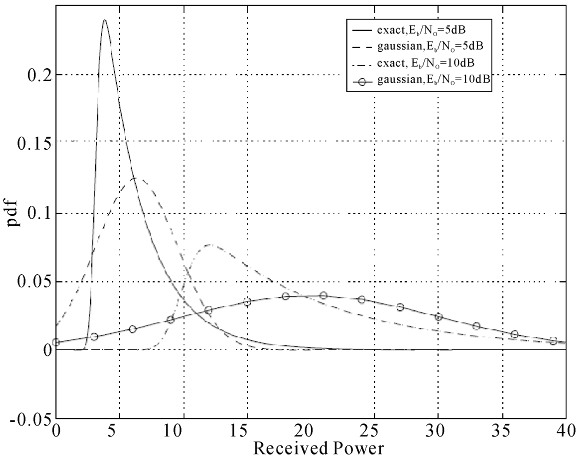


Figure 3. Probability density function of exact pdf and gaussian approximation, 64 chips/bit, E_b/N_o=5 and 10 dB.

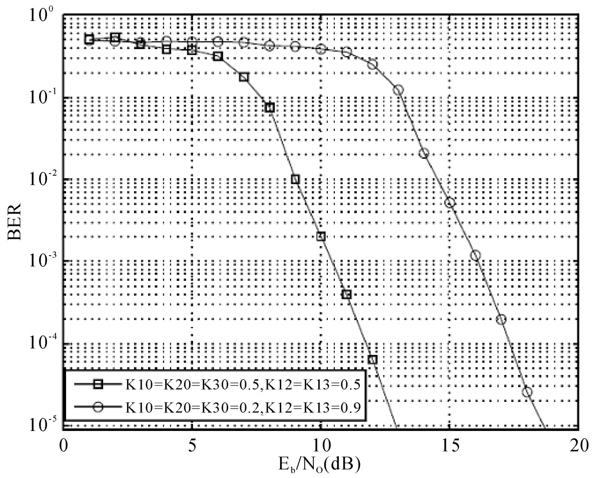


Figure 4. Simulation BER of SESS-CD, 64 chips/bit.

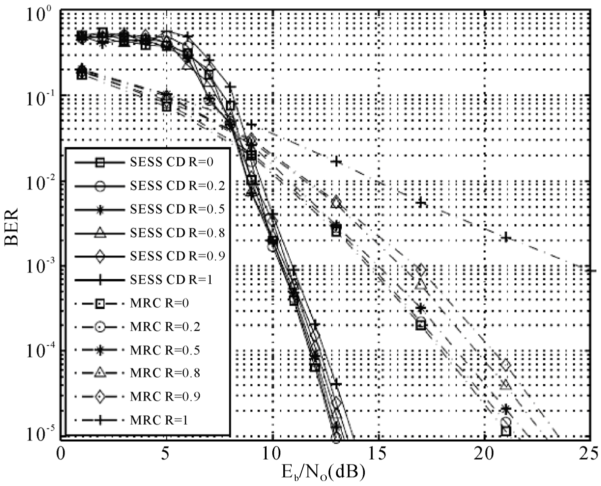


Figure 5. Simulation BER of MRC and SESS-CD (64 chips/bit) with K_{10} = K_{20} = K_{30} = 0.5, K_{12} = K_{13} = 0.5, for various correlation values of correlated channel.

SESS-CD with different relay locations. The relay location in the middle of the source and destination (K_{12} = 0.5, K_{20}=0.5) exhibits a better BER than the relay location near to the source (K_{12}=0.9,K_{20}=0.2). We can also see in Figure 5 that SESS-CD is stable in correlated channels but MRC degrades rapidly as the channel correlation increases. A similar effect can be observed in hostile channels with bit losses in Figure 6 where SESS-CD displays much stable BER performance compared to the MRC.

5. Conclusions

We incorporated SESS with CD in this paper. SESS-CD diversity gain is linked to the square of the received SNR. The SESS-CD BER is inversely proportional to the square term of the SNR while the MRC BER is inversely proportional to the SNR only. We observe that SESS-CD is very stable in highly correlated channels as well as in severely fading channels. SESS combined with CD is obviously a promising CD technique for the future generation of wireless communications.

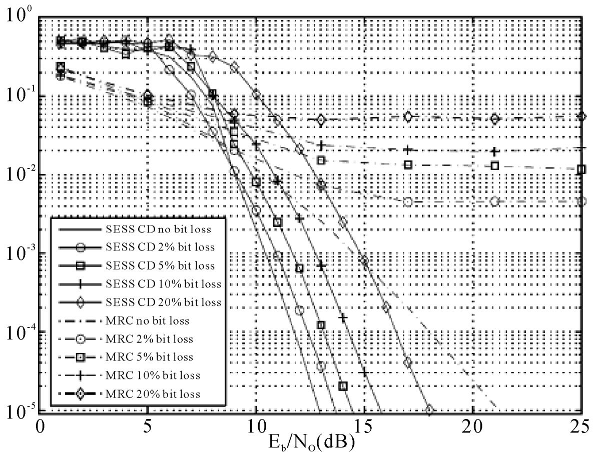


Figure 6. Simulation BER of MRC and SESS-CD (64 chips/bit) with K_{10}=K_{20}=K_{30}=0.5, K_{12}=K_{13}=0.5, under different bit loss percentage.

6. Acknowledgment

This work was supported in part by contract award FA9550-08-1-0393 from the U.S. Air Force Office of Scientific Research. Thanks are due to Dr. J. Sjogren whose support has allowed the authors to investigate the field of cooperative self encoded spread spectrum.

7. References

[1] A. Ribeiro, C. X. Cai, and G. B. Giannakis, "Symbol error probabilities for general cooperative links," IEEE Transactions on Wireless Communications, Vol. 4, No. 3, pp. 1264–1273, May 2005.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity–Part I: System description," IEEE Transactions on Communications, Vol. 51, No. 11, pp. 1927–1938, November 2003.

[3] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part II: Implementation aspects and performance analysis," IEEE Transactions on Communications, Vol. 51, No. 11, pp. 1939–1948, November 2003.

[4] A. Ribeiro , X. Cai, and G. B. Giannakis, "Opportunistic multipath for bandwidth–efficient cooperative networking," IEEE International Conference on Acoustics Speech and Signal Processing, Montreal, Canada, May 2004.

[5] L. Nguyen, "Self–encoded spread spectrum and multiple

access communication,” IEEE 6th International. Symposium on Spread−Spectrum Techniques & Applications, New Jersey, September 2000.

[6] K. Hua, L. Nguyen, W. M. Jang, “Self−encoded spread spectrum synchronization with genetic algorithm and markov chain analysis,” IEEE 42th Conference on Information Science and Systems, Princeton, New Jersey, March 2008.

[7] Y. Kong, L. Nguyen, and W. M. Jang, “On the BER of self−encoded spread spectrum communication systems,” Proceedings of the IASTED International Conference, Wireless and Optical Communications, Banff, Alberta, Canada, June 27−29, 2001.

[8] Z. Wang and G. B. Giannakis, “A simple and general parameterization quantifying performance in fading channels,” IEEE Transactions on Communications, Vol. 51, No. 8, pp.1389−1398, August 2003.

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

MATLAB Simulink Simulation Platform for Photonic Transmission Systems

Le Nguyen BINH
Centre for Telecommunications and Information Engineering, Department of Electrical and Computer Systems Engineering, Monash University, Clayton Campus, Clayton Victoria, Australia
Email: *le.nguyen.binh@eng.monash.edu.au*

Received October 25, 2008; *revised December* 20, 2008; *accepted December* 31, 2008

Abstract

High speed and ultra-high capacity optical communications have emerged as the essential techniques for backbone global information transmission networks. As the bit rate of the transmission system gets higher and higher 40 Gb/s to 100 Gb/s the modeling of proposed modulation techniques is very important so as to avoid costly practical demonstration. The search for a universal modeling platform for such systems is urgent. Matlab Simulink has become the universal mathematical and modeling tools in most universities and research laboratories around the world. This paper thus describes the modeling techniques for advanced photonic transmission systems and Simulink is proven to be very effective platform for development of photonic communications systems due its comprehensive blocksets. The simulation is based mainly on the physical phenomena and understanding of its concepts of communications and photonics. Simulink models are given as examples of various sub-systems of the photonic transmission systems. Some simulated transmission performances are demonstrated as examples of final results obtained from Simulink models of the transmission systems.

Keywords: Communications Systems, MATLAB Simulink Simulation, Optical Communications

1. Introduction

1.1. Overview of a Digital Photonic System

Any study on digital photonic transmission systems requires in-depth understanding of operational principles of system components which involve: 1) modulation/ demodulation or generation/detection of the optical signals modulated by proposed formats and the detection here implies the incoherent direct detection; 2) impairments in either electronic or photonic domains, especially the dynamics of optical fiber and the noise sources contributed by optical amplifiers and receiver electronic noise; 3) effects of optical and electrical filters. The schematic diagram of a DWDM digital photonic system is illustrated in Figure 1.

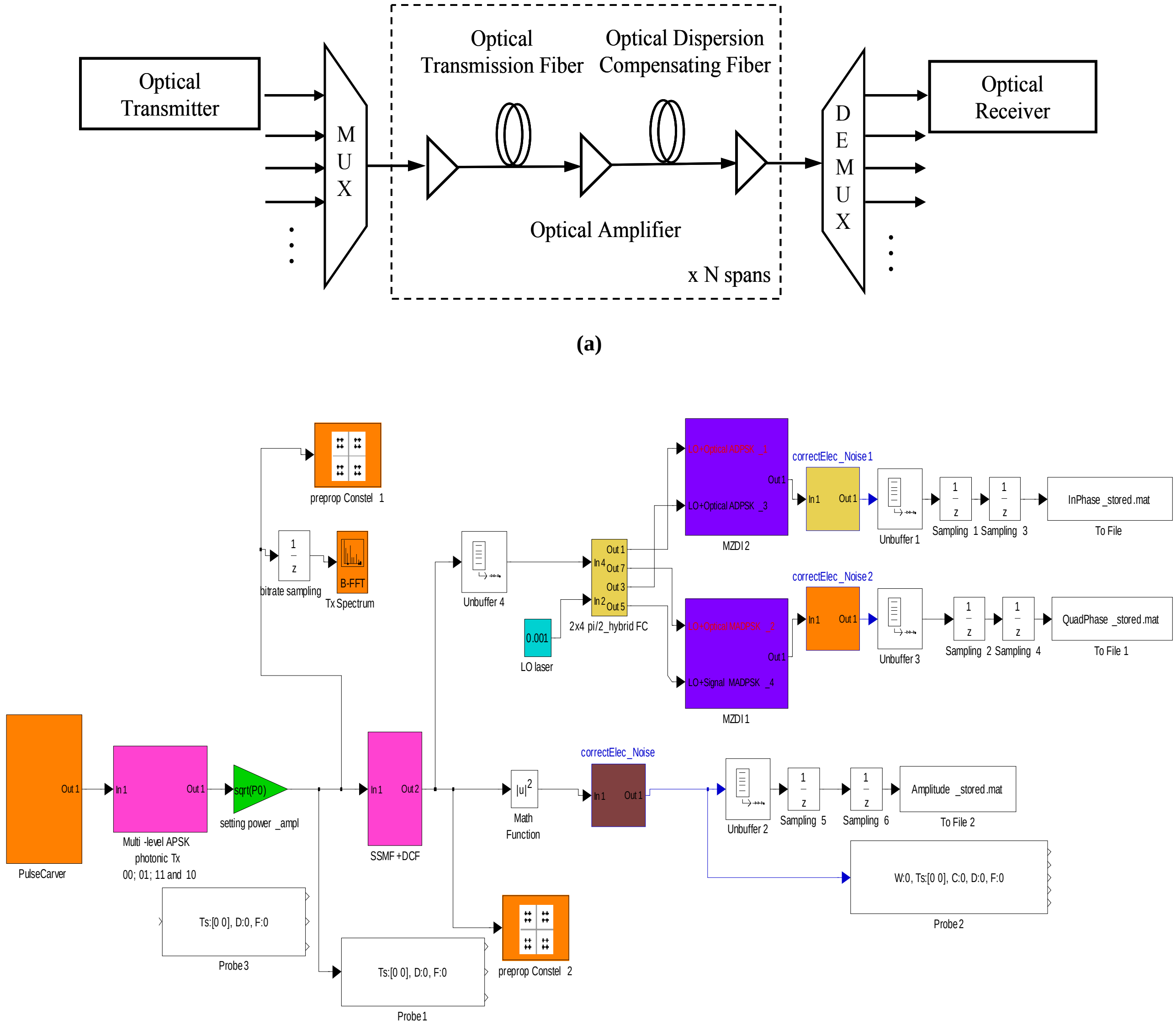
The transmission medium may consist a variety of fiber types such as the standard SMF ITU-G.652 or non-zero dispersion shifted fibers (NZ-DSF) ITU-G.655 or the new type of fiber: Corning Vascade fiber. The dispersion and distortion of the lightwave signals are usually compensated by dispersion compensating fibers (DCF). The DCFs are normally accompanied by two discrete optical amplifiers, the Erbium-doped optical amplifiers (EDFA), one is for pre-amplification to compensate the attenuation of the transmission span, and the other is a booster amplifier for boosting the optical power of the channels to an acceptable, below the nonlinear limit level. It is assumed in this work that the amplifiers are operating in the saturation region.

The receiving sub-system would take on: 1) single detector direct detection optical receiver 2) the balanced detector receiving structure. The first type of the receiver is widely used for detection of ASK modulated optical signals. For the later case, the structure acts as an optical phase comparator employing a delay interferometer. Detailed description of these direct detection receivers for novel modulation formats are presented. In addition, especially for contemporary systems with high capacity, high bit rate and requiring high performance, electronic equalizers can be employed as part of the receiver. Section 5 gives insight and performance of one of the most effective electronic equalizers which is maximum likelihood sequence estimation (MLSE) with Viterbi algorithm.

1.2. Matlab Simulink® Modeling Platform

High speed and high capacity modern digital photonic systems require careful investigations on the theoretical performance against various impairments caused by either electronics or fiber dynamics before they are deployed in practice. Thus, the demand for a comprehensive modeling platform of photonic systems is critical, especially a modeling platform that can structure truly the photonic sub-systems. A simulation test-bed is necessary for detailed design, investigation and verification on the benefits and shortcomings of these advanced modulation formats on the fiber-optic transmission systems

Furthermore the modeling platform should take advantage of any user-friendly software platform that are popular and easy for use and further development for all the operators without requiring very much expertise in the physics of the photonic systems. Besides, this platform would offer research community of optical communication engineering a basis for extension and enhance the linkages between research groups.



(a)

(b)

Figure 1. (a) General block diagram of the DWDM optical fiber transmission system. (b) MATLAB Simulink model.

Thus, it is the principal incentive for the development of a simulation package based on Matlab Simulink® platform[1] To the best of my knowledge, this is the first Matlab© Simulink-platform photonic transmission testbed for modeling advanced high capacity and long-haul digital optical fiber transmission systems. The simulator is used mainly for investigation of performance of advanced modulation formats, especially the amplitude and/or phase shift keying modulation with or without the continuity at the phase transition. Here, a single channel optical system is of main interest for implementation of the modeling in this paper.

Several noticeable advantages of the developed Matlab Simulink® modeling platform are listed as follows:

- The simulator provides toolboxes and blocksets adequately for setting up any complicated system configurations under test. The initialization process at the start of any simulation for all parameters of system components can be automatically conducted. The initialization file is written in a separate Matlab file so that the simulation parameter can be modified easily.
- Signal monitoring is especially easy to be carried out. Signals can be easily monitored at any point along the propagation path in a simulation with simple plug-and-see monitoring scopes provided by Simulink®.
- Numerical data including any simulation parameters and the numerical results can be easily stored for later processing using Matlab toolboxes. This offers a complete package from generating the numerical data to processing these data for the achievement of final results.
- A novel modified fiber propagation algorithm has been developed and optimized to minimize the simulation processing time and enhance its accuracy.
- The transmission performance of the optical transmission systems can be automatically and accurately evaluated with various evaluation methods. These methods, especially proposal of novel statistical evaluation techniques are to be presented in Section 6.

Several Matlab Simulink® modeling frameworks are demonstrated in the Appendix of this paper. A Simulink model of a photonic transmission system can be shown in Figure 1(b).

2. Optical Transmitters

The transmitters would consist of a narrow linewidth laser source to generate lightwaves of wavelength conformed to the ITU grid. These lightwaves are combined and then modulated. This form is for laboratory experiments only. In practice each laser source would be modulated by an external modulation sub-system. The MZIM can bee a single or dual drive type. The schematic of the modulator is shown in Figure 2(a) and the Simulink model is in Figure 2(a) for generation of photonic signals by multi-level amplitude and phase shift keying modulation formats.

In 1980s and 1990s, direct modulation of semiconductor lasers was the choice for low capacity coherent optical systems over short transmission distance. However, direct modulation induces chirping which results in severe dispersion penalties. In addition, laser phase noise and induced from non-zero laser linewidth also limit the advance of direct modulation to higher capacity and higher bit rate transmission.

Overcoming the mentioned issues, external modulation techniques have been the preferred option for digital photonic systems for over the last decade. External modulation can be implemented using either electro-absorption modulator or electro-optic modulators (EOM). The EOM whose operation is based on the principles of electro-optic effect (i.e. change of refractive index in solid state or polymeric or semiconductor material is proportional to the applied electric field) has been the preferred choice of technology due to better performance in terms of chirp, extinction ratio and modulation speed. Over the years, the waveguides of the electro-optic modulators are mainly integrated on the material platform of lithium niobate (LiNbO3) which has been the choice due to their prominent properties of low loss, ease of fabrication and high efficiency [4].

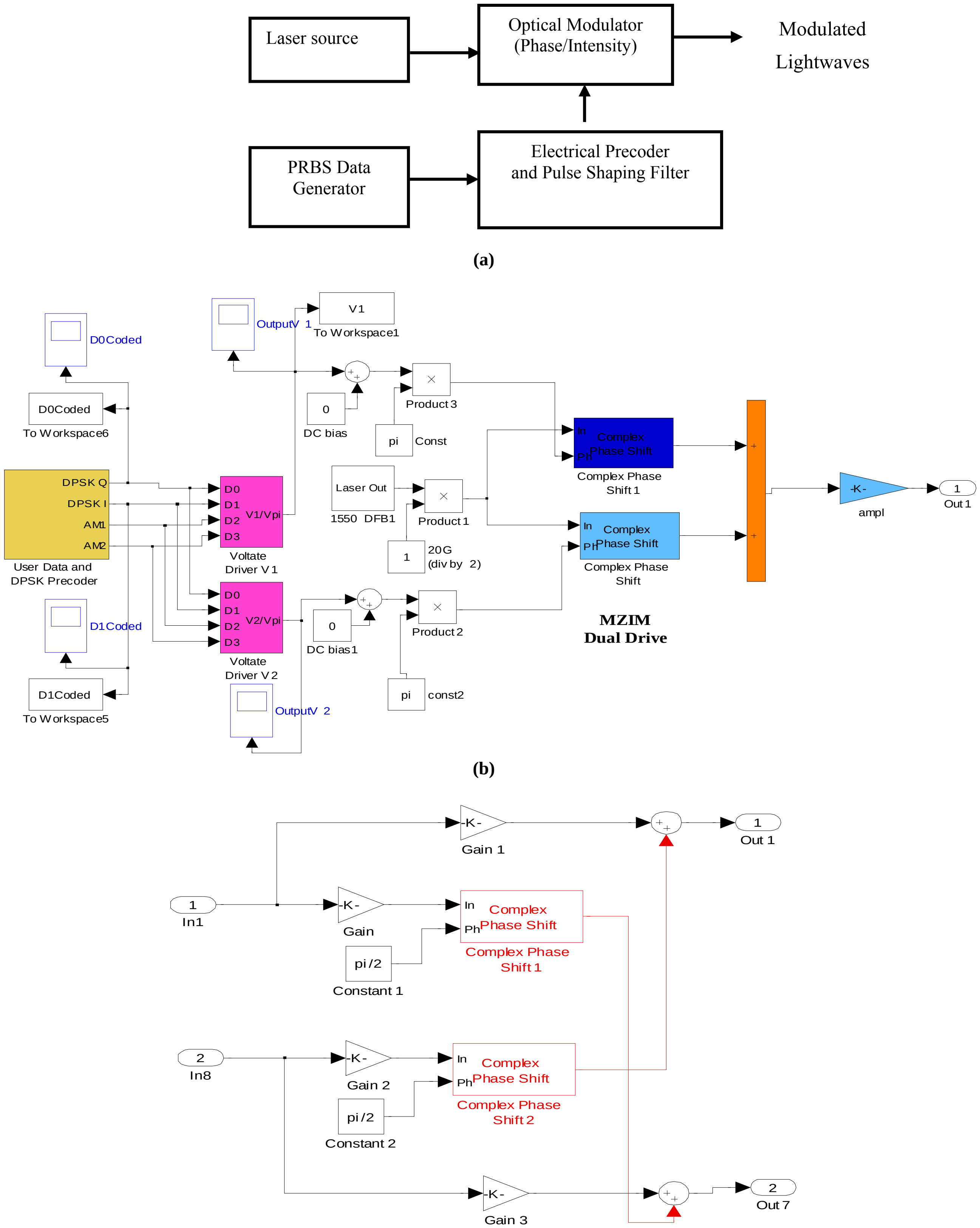
These LiNbO3 modulators have been developed in the early 1980s, but not popular until the advent of the Erbium-doped optical fiber amplifier (EDFA) in the late 1980s. Prior to the current employment of LiNbO3 modulators for advanced modulation formats, they were employed in coherent optical communications to mitigate the effects of broad linewidth due to direct modulation of the laser source. These knowledges have recently been applied to the in-coherent advanced modulation formats for optically amplified transmission systems.

EOMs are utilized for modulation of either the phase or the intensity of the lightwave carrier. The later type is a combination of two electro-optic phase modulators (EOPMs) forming an interferometric configuration.

2.1. Optical Phase Modulator

Electro-optic phase modulator employs a single electrode as shown in Figure 3. When a RF driving voltage is applied onto the electrode, the refractive index changes accordingly inducing variation amount of delays of the propagating lightwave. Since the delays correspond to the phase changes, EOPM is used to carry out the phase modulation of the optical carrier.

[1] http://www.mathworks.com/, access date: August 22, 2008.



(a)

(b)

(c)

Figure 2. Structure of external modulation for generation of advanced modulation format lightwave signals. (a) Schematic. (b) Simulink model of pre-coder and modulation. (c) Details of MZIM.

The induced phase variation is governed by the following equation:

$$\varphi_1(\mathrm{t}) = \pi \frac{V_{RF}(t)}{V_\pi} \quad \textbf{(1)}$$

where V_π is the RF driving voltage required to create a π phase shift of the lightwave carrier and typically has a value within a range of 3V to 6V. The optical field at the output of an EOPM is generated given in following equation:

$$E_0 = E_\mathrm{i} \mathrm{e}^{j\varphi_1(t)} \quad \textbf{(2)}$$

where E_o is the transmitted optical field at the output the MZIM and noted in the low pass equivalent representation i.e the carrier is removed from the expression; V(t) is the time-varying signal voltage, Vbias is the DC bias voltage applied to the phase modulator.

Recently, EOPMs operating at high frequency using resonant-type electrodes have been studied and proposed in [2,3]. Together with the advent of high-speed electronics which has evolved with the ASIC technology using 0.1µm GaAs P-HEMT or InP HEMTs [4], the contemporary EOPMs can now exceed 40Gb/s operating rate without much difficulty.

Such phase modulation can be implemented in MATLAB Simulink as shown in Figure 2(b) using a phase shift block of the Common Blockset. The phase bias is in one phase shift block and then the signal modulation or time dependent is fed into another phase shift block. The signals of the two parallel phase shift/modulation blocks are then combined to represent the interferometric construction and destruction, thus an intensity modulation can be achieved as described in the next sub-section.

2.2. Optical Intensity Modulator

Optical intensity modulation is operating based on the principle of interference of the optical field of the two lightwave components. A LiNbO3 optical intensity modulator thus employs the interferometric structure as shown in Figure 4 and is most popularly well-known as the Mach-Zehnder interferometer (MZIM). The operational principles are briefly explained in the following paragraph. For the rest of the chapters in this paper, unless specifically indicated, the term of optical modulator is referred to the external LiNbO3 MZIM modulator.

The lightwave .s split into two arms when entering the modulator. The power slitter is normally a 3-dB type i.e equally splitting the power of the optical signals. Each arm of the LiNbO3 modulator employs an electro-optic phase modulator in order to manipulate the phase of the optical carrier if required. At the output of the MZIM, the lightwaves of the two arm phase modulators are coupled and interfered with each other. The transfer curve of an MZIM is shown in Figure 4(c). A LiNbO3 MZIM modulator can be a single or dual drive type.

In the case of single-drive MZIM, there is only a single RF voltage driving one arm of the MZIM. For instance, there is no RF driving voltage on arm 1, hence $V_1(t)$ = 1 and the RF voltage $V_2(t)$ applied on arm 2 is noted as $V(t)$. The transmitted optical field E(t) at the output a single-drive MZIM as a function of the driving voltage $V(t)$ and a bias DC voltages V_{bias} can be written as

$$\begin{aligned} E_0 &= \frac{E_i}{2}\left[1 + e^{j\pi \frac{(V(t)+V_{bias})}{V_\pi}}\right] \\ &= E_i \cos\left[\frac{\pi}{2}\frac{(V(t)+V_{bias})}{V_\pi}\right] e^{-j\left[\frac{\pi}{2}\frac{(V(t)+V_{bias})}{V_\pi}\right]} \end{aligned} \quad \textbf{(3)}$$

where V_π. is the required driving voltage to obtain a π phase shift in the lightwave carrier.

It can be seen that the phase term in Equation (1) implies the existence of the modulation of the optical carrier phase and commonly known as the chirping effect. Thus, by using a single-drive MZIM, generated optical signals is not chirp-free. Furthermore, it is reported that a z-cut LiNbO3 MZIM can provide a modest amount of chirping due to its asymmetrical structure of the electrical field distributions whereas its counterpart x-cut

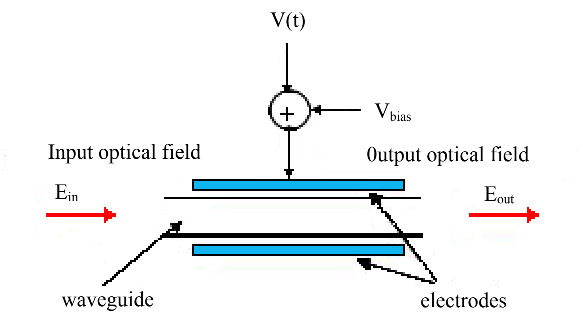


Figure 3. Electro-optic optical phase modulator.

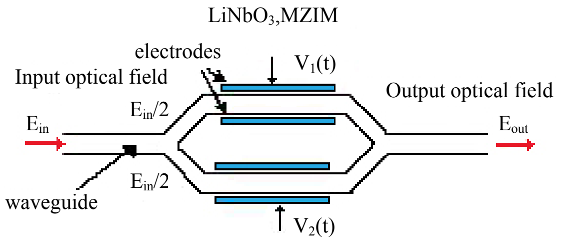


Figure 4. Optical intensity modulator based on Mach-Zehnder interferometric structure.

MZIM is a chirp-free modulator thanks to the symmetrical or push-pull configuration of the electrical fields. Furthermore, also having a push-pull arrangement, complete elimination of chirping effect in modulation of the lightwave can be implemented with use of a dual-drive MZIM. The transmitted optical field E(t) at the output a MZIM as a function of the driving and bias voltages can be written as

$$E_0 = \frac{E_i}{2}\left[e^{j\pi\frac{(V(t)+V_{bias})}{V_\pi}} + e^{j\pi\frac{-(V(t)+V_{bias})}{V_\pi}}\right] = E_i \cos\left[\frac{\pi}{2}\frac{\left(V(t)+V_{bias}\right)}{V_\pi}\right] \quad \textbf{(4)}$$

In a dual-drive MZIM, the RF driving voltage $V_1(t)$ and $V_2(t)$ are inverse with each other i.e $V_2(t)=-V_1(t)$. Equation (4) indicates that there is no longer phase modulation component, hence the chirping effect is totally eliminated.

3. Fiber Transmission Dynamics

3.1. Chromatic Dispersion (CD)

This section briefly presents the key theoretical concepts describing the properties of chromatic dispersion in a single-mode fiber. Another aim of this section is to introduce the key parameters which will be commonly mentioned in the rest of the paper.

The initial point when mentioning to the chromatic dispersion is the expansion of the mode propagation constant or "wave number" parameter, β, using the Taylor series:

$$\beta(\omega) = \frac{\omega n(\omega)}{c} = \beta_0 + \beta_1 \Delta\omega + \frac{1}{2}\beta_2 \Delta\omega^2 + \frac{1}{6}\beta_3 \Delta\omega^3 \quad \textbf{(5)}$$

where ω is the angular optical frequency, n(ω) is the frequency-dependent refractive index of the fiber. The parameters $\beta_n = \left(\frac{d^n\beta}{d\omega^n}\right)\Big|_{\omega=\omega_0}$ have different physical meanings as 1) β_o is involved in the phase velocity of the optical carrier which is defined as $v_p = \frac{\omega_0}{\beta_0} = \frac{c}{n(\omega_0)}$; 2) β₁determines the group velocity v_g which is related to the mode propagation constant β of the guided mode by [5,6]

$$v_g = \frac{1}{\beta_1} = \left(\frac{d\beta}{d\omega}\Big|_{\omega=\omega_0}\right)^{-1} \quad \textbf{(6)}$$

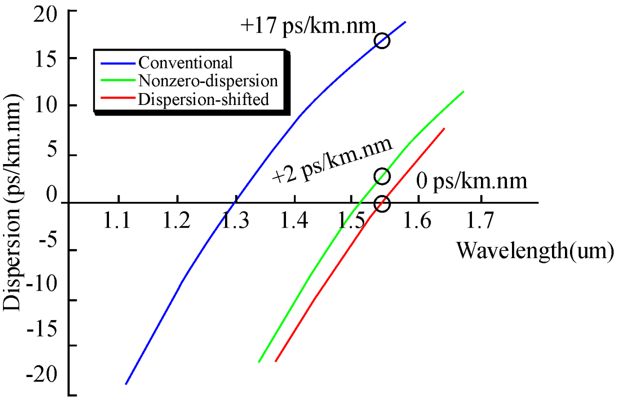


Figure 5. Typical values of dispersion factor for different types of fiber.

And 3) β2 is the derivative of group velocity with respect to frequency. Hence, it clearly shows the frequency-dependence of the group velocity. This means that different frequency components of an optical pulse travel at different velocities, hence leading to the spreading of the pulse or known as the dispersion. β_2 is therefore is known as the famous group velocity dispersion (GVD). The fiber is said to exhibit normal dispersion for β_2>0 or anomalous dispersion if β_2<0.

A pulse having the spectral width of $\Delta\omega$ is broadened by $\Delta T = \beta_2 L \Delta\omega$. In practice, a more commonly used factor to represent the chromatic dispersion of a single mode optical fiber is known as D (ps/nm.km). The dispersion factor is closely related to the GVD β2 and given by: $D = -\left(\frac{2\pi c}{\lambda^2}\right)\beta_2$ at the operating wavelength λ; where β_3 defined as $\beta_3 = \frac{d\beta_2}{d\omega}$ contributes to the calculations of the dispersion slope, $S(\lambda)$, which is an essential dispersion factor for high-speed DWDM transmission. $S(\lambda)$ can be obtained from the higher order derivatives of the propagation constant as

$$S = \frac{dD}{d\lambda} = \left(\frac{2\pi c}{\lambda^2}\right)\beta_3 + \left(\frac{4\pi c}{\lambda^3}\right)\beta_2 \quad \textbf{(7)}$$

A well-known parameter to govern the effects of chromatic dispersion imposing on the transmission length of an optical system is known as the dispersion length LD. Conventionally, the dispersion length LD corresponds to the distance after which a pulse has broadened by one bit interval. For high capacity long-haul transmission employing external modulation, the dispersion limit can be estimated in the following Equation [8].

$$L_D = \frac{10^5}{D.B^2} \quad \textbf{(8)}$$

where B is the bit rate (Gb/s), D is the dispersion factor (ps/nm km) and L_D is in km.

Equation (8) provides a reasonable approximation even though the accurate computation of this limit that depends the modulation format, the pulse shaping and the optical receiver design. It can be seen clearly from (8) that the severity of the effects caused by the fiber chromatic dispersion on externally modulated optical signals is inversely proportional to the square of the bit rate. Thus, for 10 Gb/s OC-192 optical transmission on a standard single mode fiber (SSMF) medium which has a dispersion of about ±17 ps/nm.km, the dispersion length L_D has a value of approximately 60 km i.e corresponding to a residual dispersion of about ±1000 ps/nm and less than 4 km or equivalently to about ± 60 ps/nm in the case of 40Gb/s OC-768 optical systems. These lengths are a great deal smaller than the length limited by ASE noise accumulation. The chromatic dispersion therefore, becomes the one of the most critical constraints for the modern high-capacity and ultra long-haul transmission optical systems.

3.2. Polarization Mode Dispersion (PMD)

Polarization mode dispersion (PMD) represents another type of the pulse spreading. The PMD is caused by the differential group delay (DGD) between two principle orthogonal states of polarization (PSP) of the propagating optical field.

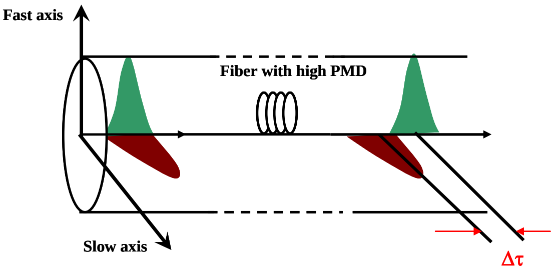


Figure 6. Demonstration of delay between two polarization states when lightwave propagating optical fiber.

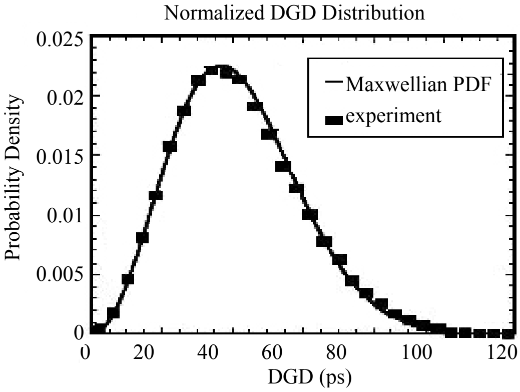


Figure 7. The Maxwellian distribution is governed by the following expression: Equation (9).

One of the intrinsic causes of PMD is due to the asymmetry of the fiber core. The other causes are derived from the deformation of the fiber including stress applied on the fiber, the aging of the fiber, the variation of temperature over time or effects from a vibration source. These processes are random resulting in the dynamic of PMD. The imperfection of the core or deformation of the fiber may be inherent from the manufacturing process or as a result of mechanical stress on the deployed fiber resulting in a dynamic aspect of PMD.

The delay between these two PSP is normally negligibly small in 10Gb/s optical transmission systems. However, at high transmission bit rate for long-haul and ultra long-haul optical systems, the PMD effect becomes much more severe and degrades the system performance [9−12]. The DGD value varies along the fiber following a stochastic process. It is proven that these DGD values complies with a Maxwellian distribution as shown in Figure 7 [10,13,14].

$$f(\Delta\tau)=\frac{32(\Delta\tau)^2}{\pi^2\langle\Delta\tau\rangle^3}\exp\left\{-\frac{4(\Delta\tau)^2}{\pi\langle\Delta\tau\rangle^2}\right\}\Delta\tau\geq 0 \qquad \textbf{(9)}$$

where $\Delta\tau$ is differential group delay over a segment of the optical fiber δz. The mean DGD value $\langle\Delta\tau\rangle$ is commonly termed as the “fiber PMD” and normally given by the fiber manufacturer.

An estimate of the transmission limit due to PMD effect is given as:

$$L_{\max}=\frac{0.02}{\langle\Delta\tau\rangle^2\cdot R^2} \qquad \textbf{(10)}$$

where R is the transmission bit rate. Therefore, $\langle\Delta\tau\rangle$=1 ps/km (older fiber vintages); Bit rate = 40 Gbit/s; Lmax=12.5 Km; Bit rate =10 Gbit/s; Lmax=200 Km; $\langle\Delta\tau\rangle$=0.1 ps/km (contemporary fiber for modern optical systems); Bit rate = 40 Gbit/s; Lmax=1250 Km ; thence for Bit rate = 10 Gbit/s ; Lmax=20.000 Km.

Thus PMD is an important impairment of ultra long distance transmission system even at 10 Gb/s optical transmission. Upgrading to higher bit rate and higher capacity, PMD together with CD become the most two critical impairments imposing on the limitation of the optical systems.

3.3. Fiber Nonlinearity

The fiber refractive index is not only dependent of wavelength but also of intensity of the lightwave. This well-known phenomenon which is named as the Kerr effect is normally referred as the fiber nonlinearity. The power dependence of the refractive index n_r is shown in the following expression

$$n_r^{'} = n_r + \overline{n}_2 (P / A_{\text{eff}}) \quad \textbf{(11)}$$

P is the average optical intensity inside the fiber, $\overline{n}_2$ is the nonlinear-index coefficient and A_{eff} is the effective area of the fiber.

There are several non-linearity phenomena induced from the Kerr effects including intra-channel self-phase modulation (SPM), cross phase modulation between inter-channels (XPM). four wave mixing (FWM), stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS). SRS and SBS are not main degrading factors compared to the others. FWM effect degrades performance of an optical system severely if the local phase of the propagating channels are matched with the introduction of the ghost pulse. However, with high local dispersion parameter such as in SSMF or even in NZ-DSF, effect of the FWM becomes negligible. XPM is strongly dependent on the channel spacing between the channels and also on local dispersion factor of the optical fiber [refs]. [ref] also report about the negligible effects of XPM on the optical signal compared the SPM effect. Furthermore, XPM can be considered to be negligible in a DWDM system in the following scenarios: 1) highly locally dispersive system e.g SSMF and DCF deployed systems; 2) large channel spacing and 3) high spectral efficiency [15−19]. However, the XPM should be taken in to account for the systems deploying Non-zero dispersion shifted fiber (NZ-DSF) where the local dispersion factor is low. The values of the NZ-DSF dispersion factors can be obtained from Figure 5. Among nonlinearity impairments, SPM is considered to be the major shortfalls in the system.

In this paper, only the SPM non-linearity is generally considered. This is the main degradation factor for high bit rate transmission system where the signal spectrum is broadened. The effect of SPM is normally coupled with the nonlinear phase shift which is defined as

$$\begin{aligned} \phi_{NL} &= \int_0^L \gamma P(z) dz = \gamma L_{\text{eff}} P \\ \gamma &= \omega_c \overline{n}_2 / (A_{\text{eff}} c) \\ L_{\text{eff}} &= (1 - e^{-\alpha L}) / \alpha \end{aligned} \quad \textbf{(12)}$$

where ω_c is the lightwave carrier, L_{eff} is the effective transmission length and α is the attenuation factor of a SSMF which normally has a value of 0.17−0.2 dB/km for the currently operating wavelengths within the 1550nm window. The temporal variation of the non-linear phase ϕ_{NL} while the optical pulses propagating along the fiber results in the generation of new spectral components far apart from the lightwave carrier ω_c implying the broadening of the signal spectrum. The spectral broadening $\delta\omega$ which is well-known as frequency chirping can be explained based on the time dependence of the nonlinear phase shift and given by the expression:

$$\delta\omega = -\frac{\partial \phi_{NL}}{\partial T} = -\gamma \frac{\partial P}{\partial T} L_{eff} \quad \textbf{(13)}$$

From (13), the amount of $\delta\omega$ is proportional to the time derivative of the signal power P. Correspondingly, the generation of new spectral components may mainly occur the rising and falling edges of the optical pulse shapes, i.e. the amount of generated chirp is larger for an increased steepness of the pulse edges.

4. Modeling of Fiber Propagation

4.1. Non-linear Schrodinger Equation (NLSE)

Evolution of the slow varying complex envelope $A(z,t)$ of the optical pulses along a single mode optical fiber is governed by the well-known nonlinear Schroedinger equation (NLSE):

$$\begin{aligned} &\frac{\partial A(z,t)}{\partial z} + \frac{\alpha}{2} A(z,t) + \beta_1 \frac{\partial A(z,t)}{\partial t} + \frac{j}{2} \beta_2 \frac{\partial^2 A(z,t)}{\partial t^2} \\ &- \frac{1}{6} \beta_3 \frac{\partial^3 A(z,t)}{\partial t^3} = -j\gamma |A(z,t)|^2 A(z,t) \end{aligned} \quad \textbf{(14)}$$

where z is the spatial longitudinal coordinate, α accounts for fiber attenuation, β_1 indicates the differential group delay (DGD), β_2 and β_3 represent 2nd and 3rd order factors of the group velocity dispersion (GVD) and γ is the nonlinear coefficient. Equation (14) involves the following effects in a single-channel transmission fiber: 1) the attenuation, 2) chromatic dispersion, 3) 3rd order dispersion factor i.e the dispersion slope, and 4) self phase modulation nonlinearity. Other critical degradation factors such as the non-linear phase noise due to the fluctuation of the optical intensity caused by ASE noise via Gordon-Mollenauer effect [20] is mutually included in the equation.

4.2. Symmetrical Split Step Fourier Method

In this Paper, solutions of the NLSE and hence the model of pulse propagation in a single mode optical fiber is numerically solved by using the popular approach of the split step Fourier method (SSFM) [5] in which the fiber length is divided into a large number of segments of small step size δz.

In practice, dispersion and nonlinearity are mutually interactive while the optical pulses propagate through the fiber. However, the SSFM assumes that over a small length δz, the effects of dispersion and the nonlinearity on the propagating optical field are independent. Thus, in SSFM, the linear operator representing the effects of fiber dispersion and attenuation and the nonlinearity operator taking into account fiber nonlinearities are defined separately as

$$\hat{D} = -\frac{i\beta_2}{2}\frac{\partial^2}{\partial T^2} + \frac{\beta_3}{6}\frac{\partial^3}{\partial T^3} - \frac{\alpha}{2} \qquad \textbf{(15)}$$

$$\hat{N} = i\gamma |A|^2$$

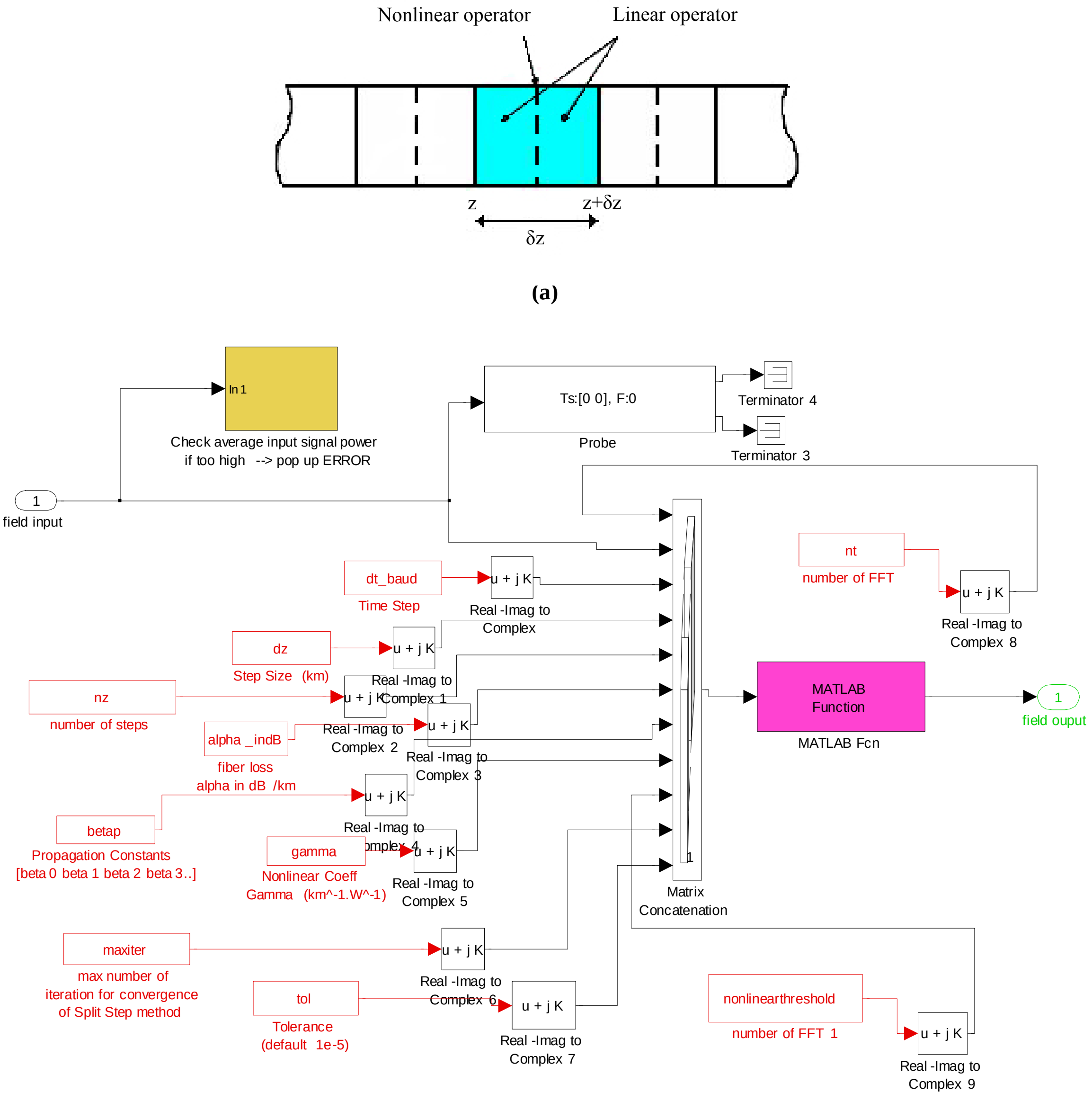
where A replace s $A(z,t)$ for simpler notation and T=t-z/v_g is the reference time frame moving at the group velocity. The NLSE Equation (14) can be rewritten as

$$\frac{\partial A}{\partial z} = (\hat{D} + \hat{N})A \qquad \textbf{(16)}$$

and the complex amplitudes of optical pulses propagating from z to z+ δz is calculated using the approximation as given:

$$A(z+h,T) \approx \exp\left(h\hat{D}\right)\exp\left(h\hat{N}\right)A(z,T) \qquad \textbf{(17)}$$

Equation (14) is accurate to second order in the step size δz. The accuracy of SSFM can be improved by including the effect of the nonlinearity in the middle of the segment rather than at the segment boundary as illustrated in Equation (17) can now modified as



(a)

(b)

Figure 8. (a) Schematic illustration of the split-step Fourier method. (b) MATLAB Simulink model.

$$A(z+\delta z,T)$$
$$\approx \exp\left(\frac{\delta z}{2}\hat{D}\right)\exp\left(\int_{z}^{z+\delta z}\hat{N}(z')dz'\right)\exp\left(\frac{\delta z}{2}\hat{D}\right)A(z,T) \quad \textbf{(18)}$$

This method is accurate to third order in the step size δz. The optical pulse is propagated down segment from segment in two stages at each step. First, the optical pulse propagates through the first linear operator (step of $\delta z/2$) with dispersion effects taken into account only. The nonlinearity is calculated in the middle of the segment. It is noted that the nonlinearity effects is considered as over the whole segment. Then at z+$\delta z/2$, the pulse propagates through the remaining $\delta z/2$ distance of the linear operator. The process continues repetitively in executive segments δz until the end of the fiber. This method requires the careful selection of step sizes δz to reserve the required accuracy.

The Simulink model of the lightwave signals propagation through optical fiber is shown in Figure 8(b). All parameters required for the propagation model are fed as the inputs into the block. The propagation algorithm split-steps and FFT are written in .m files in order to simplify the model. This demonstrates the effectiveness of the linkage between MATLAB and Simulink. A Matlab program is used for modeling of the propagation of the guided lightwave signals over very long distance is given in the Appendix.

4.3. Modeling of Polarization Mode Dispersion (PMD)

The first order PMD effect can be implemented by splitting the optical field into two distinct paths representing two states of polarizations with different propagating delays $\Delta\tau$, then implementing SSFM over the segment δz before superimposing the outputs of these two paths for the output optical field.

The transfer function for first-order PMD is given by [21].

$$H_f(f) = H_{f+}(f) + H_{f-}(f) \quad \textbf{(19)}$$

$$H_{f+}(f) = \sqrt{\gamma}\, exp\left[j2\pi f\left(-\frac{\Delta\tau}{2}\right)\right] \quad \text{and}$$

$$H_{f-}(f) = \sqrt{\gamma}\, \exp\left[j2\pi f\left(-\frac{\Delta\tau}{2}\right)\right]$$

with γ is the splitting ratio. The usual assumption is γ =1/2. Finite impulse response filter blocks of the digital signal processing blocksets of Simulink can be applied here without much difficulty to represent the PMD effects with appropriate delay difference.

4.4. Fiber Propagation in Linear Domain

Here, the low pass equivalent frequency response of the optical fiber, noted as *H(f)* has a parabolic phase profile and can be modeled by the following equation, [22]

$$H_c(f) = e^{-j\alpha_D f^2} \quad \textbf{(20)}$$

where, $\alpha_D = \pi^2\beta_2 L, \beta_2$ represents the Group Velocity Distortion (GVD) parameter of the fiber and L is the length of the fiber. The parabolic phase profile is the result of the chromatic dispersion of the optical fiber [23]. The 3rd order dispersion factor β_3 is not considered in this transfer function of the fiber due to negligible effects on 40Gb/s transmission systems. However, if the transmission bit rate is higher than 40Gb/s, the β_3 should be taken into account.

In the model of the optical fiber, it is assumed that the signal is propagating in the linear domain, i.e. the fiber nonlinearities are not included in the model. These nonlinear effects are investigated numerically. It is also assumed that the optical carrier has a line spectrum. This is a valid assumption considering the state-of-the-art laser sources nowadays with very narrow linewidth and the use of external modulators in signal transmission.

A pure sinusoidal signal of frequency f, propagating through the optical fiber, experiences a delay of $|2\pi f_D\beta_2 L|$. The standard fibers used in optical communications have a negative β_2 and thus, in low pass equivalent representation, sinusoids with positive frequencies (i.e. frequencies higher than the carrier) have negative delays, i.e. arrive early compared to the carrier and the ones with negative frequencies (i.e. frequencies lower than the carrier) have positive delays and arrive delayed. The dispersion compensating fibers have positive β_2 and so have reverse effects. The low pass equivalent channel impulse response of the optical fiber, $h_C(t)$ has also followed a parabolic phase profile and is given as,

$$h_c(t) = \sqrt{\frac{\pi}{j\alpha_D}}e^{-j\alpha^2 t^2/\alpha_D} \quad \textbf{(21)}$$

5. Optical Amplifier

5.1. ASE Noise of Optical Amplifier

The following formulation accounts for all noise terms that can be treated as Gaussian noise

$$N_{ASE} = mn_{sp}hv(G-1)B_o \quad \textbf{(22)}$$

G =amplifier gain; *nsp* = spontaneous emission factor; *m* =number of polarization modes (1 or 2); *PN* =mean noise in bandwidth; OSNR at the output of EDFA.

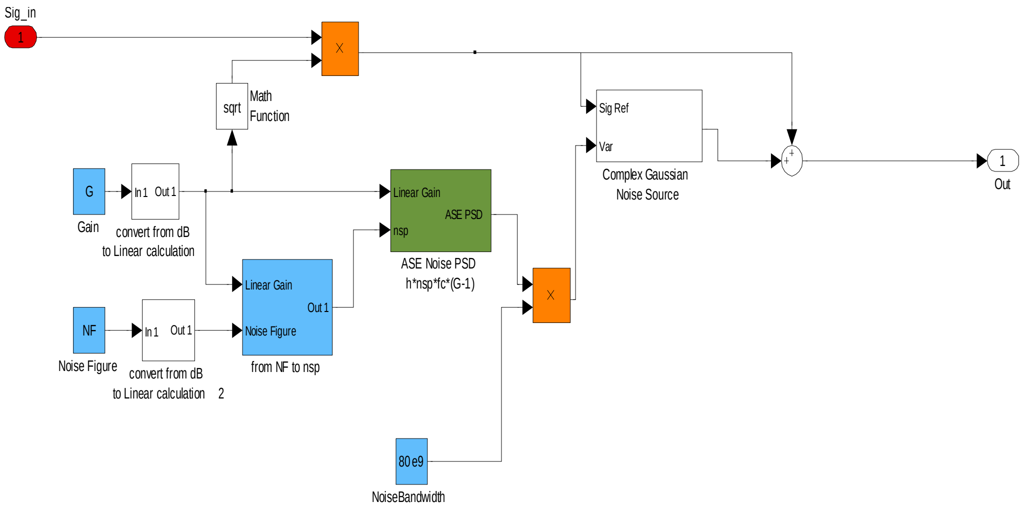


Figure 9. Simulink model of an optical amplifier with gain and NF.

5.2. Optical Amplifier Noise Figure

Amplifier Noise Figure (NF) is defined at the output of the optical amplifier as the ratio between the output OSNR on the OSNR at the input of the EDFA.

$$F_N = \frac{OSNR_{in}}{OSNR_{out}} \approx 2n_{sp} \text{ for } G \gg 1 \qquad \textbf{(23)}$$

A Simulink model of the optical amplifier is shown in that represents all the system operational parameters of such amplifier. Only blocks of the Common Blockset of Simulink are used.

6. Optical Filter

In this paper, optical filtering of the noise-corrupted optical signals is conducted with a Gaussian-type filter whose 3dB bandwidth is governed by

$$h_{Gauss}(t) = \frac{1}{\sqrt{2\pi}\xi} e^{\left(\frac{-t^2}{2\xi^2}\right)} \qquad \textbf{(24)}$$

where $\xi = \frac{\sqrt{\ln(2)}}{2\pi BT}$ in which B is the Gaussian filter's 3-dB bandwidth and T is the bit rate. The BT product parameter is B times the input signal's bit period.

The modeling of an electrical filter can also use a Gaussian filter with similar impulse response as defined in (24) or a conventional analog 5th order Bessel filter which can be easily designed using filter design toolbox in Matlab. The Matlab pseudo-codes for designing an analog 5th order Bessel filter are shown as follows:

```
[b,a]=besself(5thorder,2*pi*BTb/os_fac); %Analog filter
   [bz,az] = impinvar(b,a,1); %Digital filter
   [hf t1] = impz(bz,az,2*delay*os_fac+1,os_fac);
```

In the above pseudo-codes, the BT product parameter is defined similarly to that in the case of a Gaussian filter. Alternatively, the transfer function of an analog 5th order Bessel filter can be referred from [24].

7. Optical Receiver

The demodulation of the original message is carried out in electrical domain, thus the conversion of lightwaves to electrical signals is required. In digital optical communication, this process has been widely implemented with a PIN photodiode in a coherent or incoherent detection. The first type requires a local oscillator to coherently down-convert the modulated lightwave from optical frequency to IF frequency. The second type which has been the preferred choice for currently deployed systems is the incoherent detection which is based on square-law envelop detection of the optical signals. For incoherent detection, the recovery of clock timing is critical. In the rest of this Paper and in the simulations, ideal clock timing is assumed.

After detection, the electrical current is normally amplified with a trans-impedance amplifier before passing through an electrical filter which is normally of Bessel type. The bandwidth of the electrical filter generally varies between 0.6 and 0.8 R. At this point, electrical eye diagrams are normally observed for the assessment of signal quality. Sampling of electrically filtered received signals is next carried out. Without use of electronic equalizers, hard decision which compares the received signal level to a pre-set threshold for making the decision is implemented.

For advanced phase modulation formats such as DPSK, CPFSK or MSK, a MZDI-based balanced receiver with two photodiodes connected back-to-back is required. Excluding the distortions of waveform due to fiber dynamics and from the analytical point of view, the received electrical signals are corrupted with noise from several sources including 1) shot noise (σ^2_{shot}), 2) electronic noise σ^2_{elec} of trans-impedance amplifier, 3) dark current noise σ^2_{dark} and 4) interactions between signals and ASE noise ($\sigma^2_{signal,ASE}$) and between ASE noise itself $\sigma^2_{ASE,ASE}$ as

$$\sigma^2_{total} = \sigma^2_{shot} + \sigma^2_{elec} + \sigma^2_{dark} + \sigma^2_{signal,ASE} + \sigma^2_{ASE,ASE} \quad \textbf{(25)}$$

These noise sources are usually modeled with normal distributions whose variances representing the noise power are defined as

1) *Shot noise* is caused by the intrinsic electro-optic phenomenon of the semiconductor photodiode in which a random number of electron-hole pairs is generated with the receipt of photons causing the randomness of the induced photo-current. The shot noise is given in the following formula:

$$\sigma^2_{\mathrm{shot}} = 2 \cdot q \langle i_s \rangle B_e \quad \textbf{(26)}$$

where Be is the 3dB bandwidth of the electrical filter, <i_s> is the average signal-only photo-current after the photodiodes.

2) The electronic noise source σ^2_{elec} is injected from the trans-impedance amplifier. It is modeled with an equivalent noise current density *iNeq* over the bandwidth of the electrical filter. The unit of iNeq is $A/\sqrt{Hz}$ and the value of σ^2_{elec} is obtained as (*iNeq*) 2*Be*.

3) Value of dark current idark is normally specified with a particular photo-diode and has the unit of A/Hz. Hence, the noise power σ^2_{dark} is calculated as *idark Be.*

4) The variances of amplitude fluctuations due to the beating of signal and ASE noise and between ASE noise itself are governed by the following expressions:

$$\sigma^2_{signal,ASE} = 4 \cdot i_S i_N \frac{B_e}{B_{opt}} \quad \textbf{(27)}$$

$$\sigma^2_{ASE-ASE} = i_N^2 \frac{B_e}{B^2_{opt}} \left(2 \cdot B_{opt} - B_e\right) \quad \textbf{(28)}$$

where B_{opt} is the 3dB bandwidth of the optical filter and i_N is the noise-induced photo-current. In practice, the value of $\sigma^2_{ASE,ASE}$ is normally negligible compared to the value of $\sigma^2_{signal,ASE}$ and can be ignored without affecting the performance of the receiver.

It is worth noting that in an optically pre-amplified receiver, i.e. the optical signal is amplified at a stage before the photo-detector, $\sigma^2_{signal,ASE}$ is the dominant factor compared to other noise sources.

8. Performance Evaluation

Performance evaluation of an optical transmission system via the quality of the electrically detected signals is an essential aspect in both simulation and experiment scenarios. The key metrics reflecting the signal quality include optical signal to noise ratio (OSNR) and OSNR penalty, eye opening (EO) and eye opening penalty (EOP) where as bit error rate (BER) is the ultimate indicator for the performance of a system.

In an experimental set-up and practical optical systems, BER and the quality factor Q-factor can be obtained directly from the modern BERT test-sets and data can be exported to a portable memory for post-processing. However, it is noted that these experimental systems need to be run within at least a few hours so that the results are stable and accurate.

For the case of investigation of performance of an optical transmission system by simulation, several methods have been developed such as

1) Monte Carlo numerical method

2) Conventional method to calculate Q-factor, Q dB and hence BER based on assumption of Gaussian distribution of noise.

3) Methods based on statistical processes taking into account the distortion from the dynamic effects of the optical fibers including the ISI induced by CD, PMD and tight optical filtering.

- The first statistical technique implements the Expected Maximization theory in which the pdf of the obtained electrical detected signal is approximated as a mixture of multiple Gaussian distributions.
- The second technique is based on the Generalized Extreme Values theorem. Although this theorem is well-known in other fields such as financial forecasting, meteorology, material engineering, etc to predict the probability of occurrence of extreme values, it has not much studied to be applied in optical communications.

8.1. Monte Carlo Method

Similar to the bit error rate test (BERT) equipment commonly used in experimental transmission, the BER in a simulation of a particular system configuration can be counted. The BER is the ratio of the occurrence of errors (N_{error}) to the total number of transmitted bits N_{total} and given as:

$$BER = \frac{N_{error}}{N_{total}} \quad \textbf{(29)}$$

Monte Carlo method offers a precise picture via the BER metric for all modulation formats and receiver types. The optical system configuration under a simula-

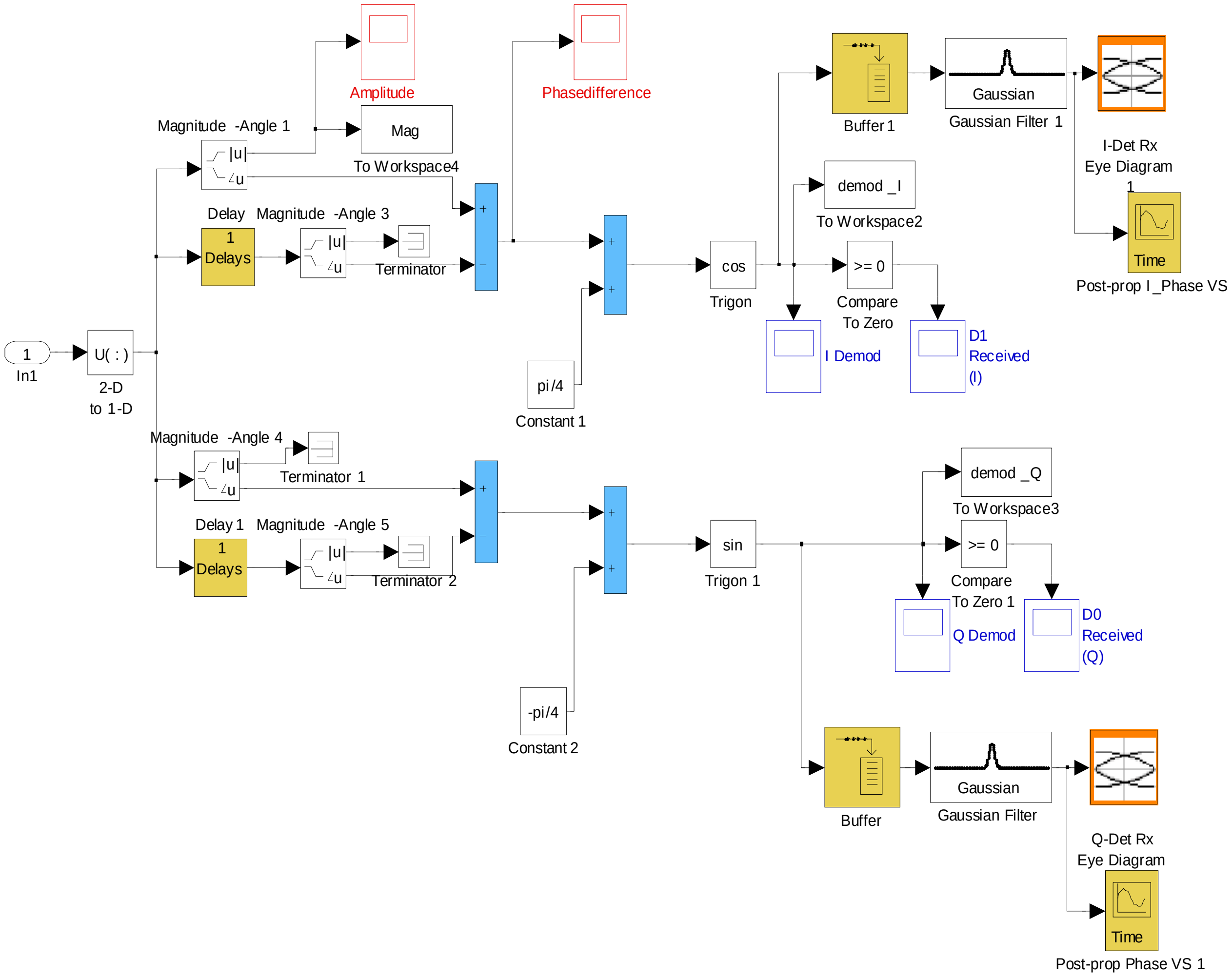


Figure 10. Simulink model of an optical balanced receiver.

tion test needs to include all the sources of impairments imposing to signal waveforms including the fiber impairments and ASE (optical)/electronic noise.

It can be seen that a sufficient number of transmitted bits for a certain BER is required and leading to exhaustive computational time. In addition, time-consuming algorithms such as FFT especially carried out in symmetrical SSFM really contribute to the long computational time. A BER of 1e-9 which is considered as 'error free' in most scientific publications requires a number of at least 1e10 bits transmitted.

However, 1e-6 even 1e-7 is feasible in Monte Carlo simulation. Furthermore, with use of forward error coding (FEC) schemes in contemporary optical systems, the reference for BERs to be obtained in simulation can be as low as 1e-3 provided no sign of error floor is shown. This is normally known as the FEC limit. The BERs obtained from the Monte Carlo method is a good benchmarking for other BER values estimated in other techniques. The time required for completion of the simulation may take several hours to reach BER of 1e-9. Thus statistical methods can be developed to determine the BER of transmission systems to save time. This is addressed in the next section.

8.2. BER and Q-Factor from Probability Distribution Functions (PDF)

This method implements a statistical process before calculating values of BER and quality Q-factor to determine the normalized probability distribution functions (PDF) of received electrical signals (for both "1" and "0" and at a particular sampling instance). The electrical signal is normally in voltage since the detected current after a photo-diode is usually amplified by a trans- impedance electrical amplifier. The PDFs can be determined statistically by using the histogram approach.

A particular voltage value as a reference for the distinction between "1" and "0" is known as the threshold voltage (V_{th}). The BER in case of transmitting bit "1" (receiving as "0" instead) is calculated from the well-known principle [25], i.e. the integral of the overlap of normalized PDF of "1"exceeding the threshold. Similar calculation for bit "0" is applied. The actual shape of the PDF is thus very critical to obtain an accurate BER. If the exact shape of the PDF is known, the BER can be calculated precisely as:

$$BER = P('1')P('0'\,|\,'1') + P('0')P('1'\,|\,'0') \qquad \textbf{(30)}$$

where: $P('1')$ is the probability that a "1" is sent; $P('0'|'1')$ is the probability of error due to receiving "0" where actually an "1" is sent; $P('0')$ is the probability that a "0" is sent; $P('1'|'0')$ is the probability of error due to receiving "1" where actually a "0" is sent; As commonly used, the probability of transmitting a "1" and "0' is equal i.e $P('1') = P('0') = 1/2$.

A popular approach in both simulation and commercial BERT test-sets is the assumption of PDF of "1" and "0" following Gaussian/normal distributions, i.e noise sources are approximated by Gaussian distributions. If the assumption is valid, high accuracy is achieved. This method enables a fast estimation of the BER by using the complementary error functions [25]:

$$BER = \frac{1}{2}\left[erfc\left(\frac{|\mu_1 - V_{th}|}{\sqrt{2}\sigma_1}\right) + erfc\left(\frac{|\mu_0 - V_{th}|}{\sqrt{2}\sigma_0}\right)\right] \qquad \textbf{(31)}$$

where μ_1 and μ_0 are the mean values for PDF of "1" and "0" respectively whereas σ_1 and σ_0 are the variance of the PDFs. The quality factor − Q-factor which can be either in linear scale or in logarithmic scale can be calculated from the obtained BER through the expression:

$$\begin{aligned} Q &= \sqrt{2}erfc^{-1}(2BER) \\ Q_{db} &= 20\log\left(\sqrt{2}erfc^{-1}(2BER)\right) \end{aligned} \qquad \textbf{(32)}$$

8.2.1. Improving Accuracy of Histogram

The common objective is to search for the proper values for number of bins and bin-width to be used in the approximation of the histogram so that the bias and the variance of the estimator can be negligible. According to [26], with a sufficiently large number of transmitted bits (N_0), a good estimate for the width (W_{bin}) of each equally spaced histogram bin is given by: $W_{bin} = \sqrt{N_0}$.

8.3. Optical Signal-to-Noise Ratio (OSNR)

The optical signal-to-noise ratio (OSNR) is a popular benchmark indicator for assessment of the performance of optical transmission systems, especially those limited by the ASE noise from the optical amplifiers – EDFAs. The OSNR is defined as the ratio of optical signal power to optical noise power. For a single EDFA with output power, *Pout*, the OSNR is given by:

$$OSNR = \frac{P_{out}}{N_{ASE}} = \frac{P_{out}}{(NF \bullet G - 1)hf\Delta f} \qquad \textbf{(33)}$$

where NF is the amplifier noise figure, G is the amplifier gain, *hf* is the photon energy, Δf is the optical measurement bandwidth.

However, OSNR does not provide good estimation to the system performance when the main degrading sources involve the dynamic propagation effects such as dispersion (including both CD and PMD) and Kerr nonlinearity effects (eg. SPM). In these cases, the degradation of the performance is mainly due to waveform distortions rather than the corruption of the ASE or electronic noise. When addressing a value of an OSNR, it is important to define the optical measurement bandwidth over which the OSNR is calculated. The signal power and noise power is obtained by integrating all the frequency components across the bandwidth leading to the value of OSNR. In practice, signal and noise power values are usually measured directly from the optical spectrum analyzer (OSA), which does the mathematics for the users and displays the resultant OSNR versus wavelength or frequency over a fixed resolution bandwidth. A value of $\Delta\lambda$ = 0.1 nm or Δf =12.5GHz is widely used as the typical value for calculation of the OSNR.

OSNR penalty is determined at a particular BER when varying value of a system parameter under test. For example, OSNR penalty at BER=1e-4 for a particular optical phase modulation format when varying length of an optical link in a long-haul transmission system configuration.

8.4. Eye Opening Penalty (EOP)

The OSNR is a time-averaged indicator for the system performance where the ratio of average power of optical carriers to noise is considered. When optical lightwaves

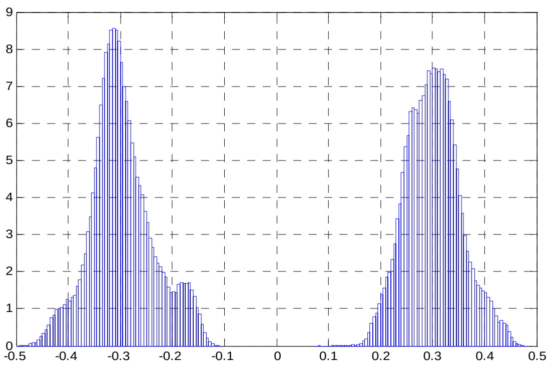


Figure 11. Demonstration of multi-peak/non-Gaussian distribution of the received electrical signal.

propagate through a dispersive and nonlinear optical fiber channel, the fiber impairments including ISI induced from CD, PMD and the spectral effects induced from nonlinearities cause the distortion of the waveforms. Another dynamic cause of the waveform distortion comes from the ISI effects as the results of optical or electrical filtering. In a conventional OOK system, bandwidth of an optical filter is normally larger than the spectral width of the signal by several times.

The eye-opening penalty (EOP) is a performance measure defined as the penalty of the "eye" caused by the distortion of the electrically detected waveforms to a reference eye-opening (EO). EO is the difference between the amplitudes of the lowest mark and the highest space.

The benchmark eye opening is usually obtained from a back-to-back measurement when the waveform is not distorted at all by any above impairments. The eye opening penalty at a particular sampling instance is normally calculated in log scale (dB unit) and given by:

$$EOP(t_{samp}) = \frac{EO_{ref}}{EO_{received}} \tag{34}$$

The EOP is useful for noise-free system evaluations as a good estimate of deterministic pulse distortion effects. The accuracy of EOP indicator depends on the sampling instance in a bit slot. Usually, the detected pulses are sampled at the instance giving the maximum eye opening. If noise is present, the calculation of the EOP become less precise because of the ambiguity of the signal levels which are corrupted by noise.

9. MATLAB Statistical Evaluation Techniques

The method using a Gaussian-based single distribution involves only the effects of noise corruption on the detected signals and ignores the dynamic distortion effects such as ISI and non-linearity. These dynamic distortions result in a multi-peak pdf as demonstrated in Figure 11, which is clearly overlooked by the conventional single distribution technique. As the result, the pdf of the electrical signal can not be approximated accurately. The addressed issues are resolved with the proposal of two new statistical methods.

Two new techniques proposed to accurately obtain the pdf of the detected electrical signal in optical communications include the mixture of multi-Gaussian distributions (MGD) by implementing the expectation maximization theory (EM) and the generalized Pareto distribution (GPD) of the generalized extreme values (GEV) theorem. These two techniques are well-known in fields of statistics, banking, finance, meteorology, etc. The implementation of required algorithms is carried out with MATLAB functions. Thus, these novel statistical methods offer a great deal of flexibility, convenience, fast-processing while maintaining the errors in obtaining the BER within small and acceptable limits.

9.1. Multi-Gaussian Distributions (MGD) via Expectation Maximization (EM)Theorem

The mixture density parameter estimation problem is probably one of the most widely used applications of the expectation maximization (EM) algorithm. It comes from the fact that most of deterministic distributions can be seen as the result of superposition of different multi distributions. Given a probability distribution function $p(x|\Theta)$ for a set of received data, $p(x|\Theta)$ can be expressed as the mixture of M different distributions:

$$p(x|\Theta) = \sum_{i=1}^{M} w_i p_i(x|\theta_i) \tag{35}$$

where the parameter are $\Theta = (w_1,...,w_M,\theta_1,...,\theta_M)$ such that $\sum_{i=1}^{M} w_i = 1$ and each p_i is a PDF by θ_i and each pdf having a weight w_i, i.e probability of that PDF.

As a particular case adopted for optical communications, the EM algorithm is implemented with a mixture of multi Gaussian distributions (MGD). This method offers great potential solutions for evaluation of performance of an optical transmission system with following reasons: 1) In a linear optical system (low input power into fiber), the conventional single Gaussian distribution fails to take into account the waveform distortion caused by either the ISI due to fiber CD and PMD dispersion, the patterning effects. Hence, the obtained BER is no longer accurate. These issues however are overcome by using the MGD method. 2) Computational time for implementing MGD is fast via the EM algorithm which has become quite popular.

The selection of Number of Gaussian distributions for MGD Fitting can be conducted as follows. The critical step affecting the accuracy of the BER calculation is the process of estimate of the number of Gaussian distributions applied in the EM algorithm for fitting the received signal pdf. This number is determined by the estimated number of peaks or valleys in the curves of 1st and 2nd derivative of the original data set. Explanation of this procedure is carried out via the well-known "Hemming Lake Pike" example as reported in [27,28]. In this problem, the data of five age-groups give the lengths of 523 pike (*Esox lucius*), sampled in 1965 from Hemming Lake, Manitoba, Canada. The components are heavily overlapped and the resultant pdf is obtained with a mixture of these 5 Gaussian distributions as shown in Figure 12(a). The figures are extracted from [29] for demonstration of the procedure.

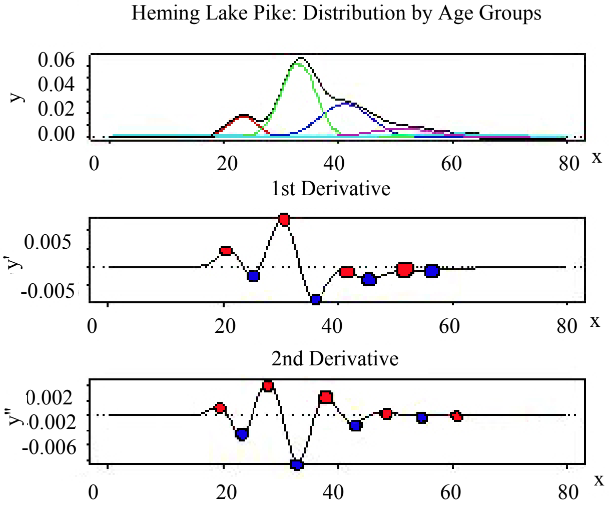


Figure 12. Five contributed Gaussian distributions.

Estimation of number of Gaussian distributions in the mixed pdf based on 1st and 2nd derivatives of the original data set (courtesy from [29]). As seen from Figure 12, the 1st derivative of the resultant pdf shows clearly 4 pairs of peaks (red) and valleys (blue), suggesting that there should be at least 4 component Gaussian distributions contributing to the original pdf. However, by taking the 2nd derivative, it is realized that there is actually up to 5 contributed Gaussian distributions as shown in Figure 12.

In summary, the steps for implementing the MGD technique to obtain the BER value is described in short as follows: 1) Obtaining the pdf from the normalized histogram of the received electrical levels; 2) Estimating the number of Gaussian distributions (N_{Gaus}) to be used for fitting the pdf of the original data set; 3) Applying EM algorithm with the mixture of N_{Gaus} Gaussian distributions and obtaining the values of mean, variance and weight for each distribution; 4) Calculating the BER value based on the integrals of the overlaps of the Gaussian distributions when the tails of these distributions cross the threshold.

9.2. Generalized Pareto Distribution (GPD)

The GEV theorem is used to estimate the distribution of a set of data of a function in which the possibility of extreme data lengthen the tail of the distribution. Due to the mechanism of estimation for the pdf of the extreme data set, GEV distributions can be classified into two classes consisting of the GEV distribution and the generalized Pareto distribution (GPD).

There has recently been only a countable number of research studies on the application of this theorem into optical communications. However, these studies only reports on the GEV distributions which only involves the effects of noise and neglect the effects of dynamic distortion factors.

Unlike the Gaussian-based techniques but rather similar to the exponential distribution, the generalized Pareto distribution is used to model the tails of distribution. This section provides an overview of the generalized Pareto distribution (GPD). The probability density function for the generalized Pareto distribution is defined as follows:

$$y = f(x|k,\sigma,\theta) = \left(\frac{1}{\sigma}\right)\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}} \qquad \textbf{(36)}$$

for $\theta < x$ when k > 0 or for $\theta < x < \frac{-\sigma}{k}$ k < 0

where k is shape parameter $k \neq 0$, σ is scale parameter and the threshold parameter θ.

Equation (36) has significant constraints given as

- When k>0: $\theta < x$ i.e there is no upper bound for x
- When k<0: $\theta < x < -\frac{\sigma}{k}$ and zero probability for the case $x > -\frac{\sigma}{k}$
- When k = 0, i.e Equation turning to:

$$y = f(x|0,\sigma,\theta) = \left(\frac{1}{\sigma}\right)e^{-\frac{(x-\theta)}{\sigma}} \quad \text{for } \theta < \text{x}$$

- If k = 0 and θ = 0, the generalized Pareto distribution is equivalent to the exponential distribution.
- If k > 0 and θ = σ, the generalized Pareto distribution is equivalent to the Pareto distribution.

The GPD has three basic forms reflecting different class of underlying distributions.

- Distributions whose tails decrease exponentially, such as the normal distribution, lead to a generalized Pareto shape parameter of zero.
- Distributions with tails decreasing as a polynomial, such as Student's t lead to a positive shape parameter.
- Distributions having finite tails, such as the beta, lead to a negative shape parameter.

GPD is widely used in fields of finance, meteorology, material engineering, etc… to for the prediction of extreme or rare events which are normally known as the exceedances. However, GPD has not yet been applied in optical communications to obtain the BER. The following reasons suggest that GPD may become a potential and a quick method for evaluation of an optical system, especially when non-linearity is the dominant degrading factor to the system performance.

1) The normal distribution has a fast roll-off, i.e. short tail. Thus, it is not a good fit to a set of data involving exeedances, i.e. rarely happening data located in the tails of the distribution. With a certain threshold value, the

generalized Pareto distribution can be used to provide a good fit to extremes of this complicated data set.

2) When nonlinearity is the dominating impairment degrading the performance of an optical system, the sampled received signals usually introduce a long tail distribution. For example, in case of DPSK optical system, the distribution of nonlinearity phase noise differs from the Gaussian counterpart due to its slow roll-off of the tail. As the result the conventional BER obtained from assumption of Gaussian-based noise is no longer valid and it often underestimates the BER.

3) A wide range of analytical techniques have recently been studied and suggested such as importance sampling, multi-canonical method, etc. Although these techniques provide solutions to obtain a precise BER, they are usually far complicated. Whereas, calculation of GPD has become a standard and available in the recent Matlab version (since Matlab 7.1). GPD therefore may provide a very quick and convenient solution for monitoring and evaluating the system performance. Necessary preliminary steps which are fast in implementation need to be carried out the find the proper threshold.

4) Evaluation of contemporary optical systems requires BER as low as 1e-15. Therefore, GPD can be seen quite suitable for optical communications.

9.2.1. Selection of Threshold for GPD Fitting

Using this statistical method, the accuracy of the obtained BER strongly depend on the threshold value (V_{thres}) used in the GPD fitting algorithm, i.e. the decision where the tail of the GPD curve starts.

There have been several suggested techniques as the guidelines aiding the decision of the threshold value for the GPD fitting. However, they are not absolute techniques and are quite complicated. In this paper, a simple technique to determine the threshold value is proposed. The technique is based on the observation that the GPD tail with exceedances normally obeying a slow exponential distribution compared to the faster decaying slope of the distribution close to the peak values. The inflection region between these two slopes gives a good estimation of the threshold value for GPD fitting. This is demonstrated in Figure 13.

Whether the selection of the V_{thres} value leads to an adequately accurate BER or not is evaluated by using the cumulative density function (cdf-Figure 14) and the quantile-quantile plot (QQ plot Figure 15). If there is a high correlation between the pdf of the tail of the original data set (with a particular V_{thres}) and pdf of the GPD, there would be a good fit between empirical cdf of the data set with the GPD-estimated cdf with focus at the most right region of the two curves. In the case of the QQ-plot, a linear trend would be observed. These guidelines are illustrated in Figure 14. In this particular case, the value of 0.163 is selected to be V_{thres}.

Furthermore, as a demonstration of improper selection of V_{thres}, the value of 0.2 is selected. Figure 16 and Figure 17 show the non-compliance of the fitted curve with the GPD which is reflected via the discrepancy in the two cdfs and the nonlinear trend of the QQ-plot.

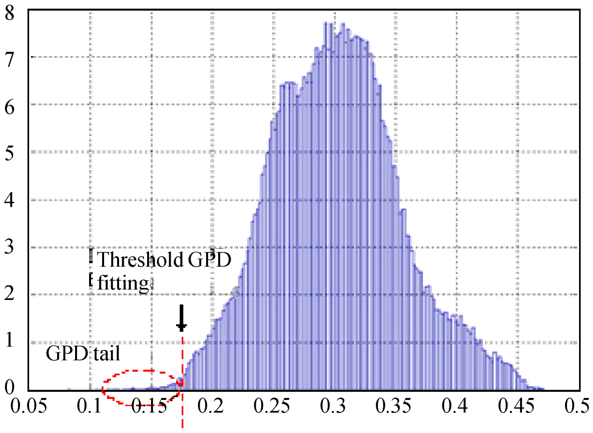


Figure 13. Selection of threshold for GPD fitting.

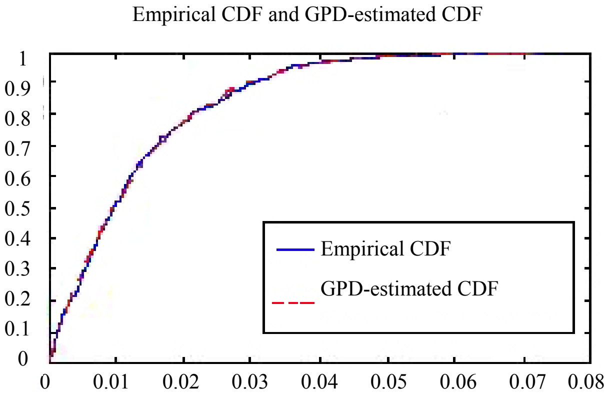


Figure 14. Comparison between fitted and empirical cumulative distribution functions.

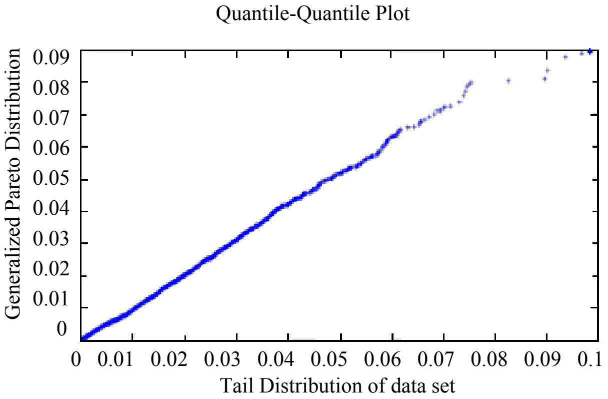


Figure 15. Quantile-quantile plot.

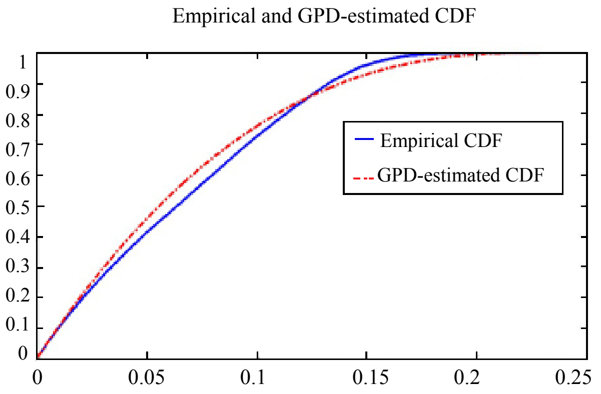


Figure 16. Comparison between fitted and empirical cumulative distribution functions.

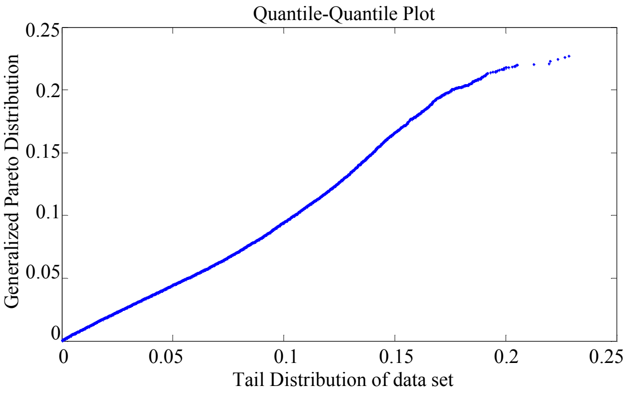


Figure 17. Quantile-quantile plot.

9.3. Validation of the Statistical Methods

A simulation test-bed of an optical DPSK transmission system over 880 km SSMF dispersion managed optical link (8 spans) is set up. Each span consists of 100 km SSMF and 10 km of DCF whose dispersion values are +17 ps/nm.km and -170 ps/nm.km at 1550 nm wavelength respectively and fully compensated i.e zero residual dispersion. The average optical input power into each span is set to be higher than the nonlinear threshold of the optical fiber. The degradation of the system performance hence is dominated by the nonlinear effects which are of much interest since it is a random process creating indeterminate errors in the long tail region of the pdf of the received electrical signals.

The BER results obtained from the novel statistical methods are compared to that from the Monte-Carlo simulation as well as from the semi-analytical method. Here, the well-known analytical expression to obtain the BER of the optical DPSK format is used, given as [30].

$$BER = \frac{1}{2} - \frac{\rho e^{-\rho}}{2} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left[I_k\left(\frac{\rho}{2}\right) + I_{k+1}\left(\frac{\rho}{2}\right) \right]^2 e^{-\frac{1}{2}(2k+1)^2 \sigma_{NLP}^2} \quad \textbf{(37)}$$

where ρ is the obtained OSNR and σ_{NLP}^2 is the variance of nonlinear phase noise.

In this case, in order to calculate the BER of a optical DPSK system involving the effect of nonlinear phase noise, the required parameters including the OSNR and the variance of nonlinear phase noise etc are obtained from the simulation numerical data which is stored and processed in Matlab. The fitting curves implemented with the MGD method for the pdf of bit 0 and bit 1 (input power of 10 dBm) as shown in Figure 18 and illustrated in Figure 19 for bit 0 and bit 1 respectively.

The selection of optimal threshold for GPD fitting follows the guideline as addressed in detail in the previous section. The BER from various evaluation methods are shown in Table 1. The input powers are controlled to be 10 dBm and 11 dBm.

Table 1 validates the adequate accuracy of the proposed novel statistical methods with the discrepancies compared to the Monte-Carlo and semi-analytical BER to be within one decade. In short, these methods offer a great deal of fast processing while maintaining the accuracy of the obtained BER within the acceptable limits.

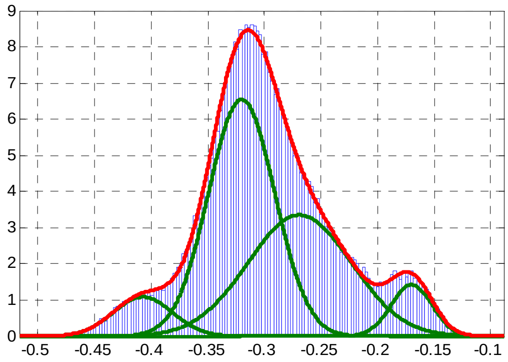


Figure 18. Demonstration of fitting curves for bit ‘0’ with MGD method.

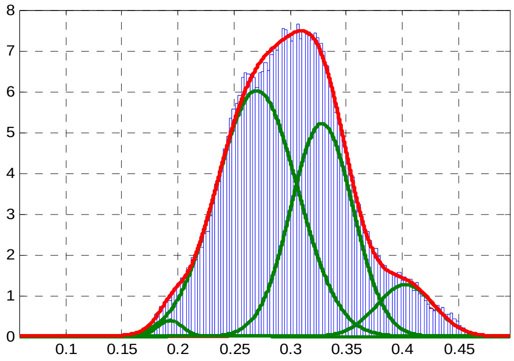


Figure 19. Demonstration of fitting curves for bit ‘0’ with MGD method.

Table 1. The BER from various evaluation methods.

Evaluation Methods / Input Power	Monte-Carlo Method	Semi-analytical Method	MGD method	GPD method
10 dBm	1.7e-5	2.58e-5	5.3e-6	3.56e-4
11 dBm		1.7e-8	2.58e-9	4.28e-8

10. Conclusions

We have demonstrated the Simulink modeling of amplitude and phase modulation formats at 40 Gb/s optical fiber transmission. A novel modified fiber propagation algorithm has been used to minimize the simulation processing time and optimize its accuracy. The principles of amplitude and phase modulation, encoding and photonic-opto-electronic balanced detection and receiving modules have been demonstrated via Simulink modules and can be corroborated with experimental receiver sensitivities.

The XPM and other fiber nonlinearity such as the Raman scattering, four wave mixing are not integrated in the Matlab Simulink models. A switching scheme between the linear only and the linear and nonlinear models is developed to enhance the computing aspects of the transmission model.

Other modulations formats such as multi-level M-DPSK, M-ASK that offer narrower effective bandwidth, simple optical receiver structures and no chirping effects would also be integrated. These systems will be reported in future works. The effects of the optical filtering components in DWDM transmission systems to demonstrate the effectiveness of the DPSK and DQPSK formats, have been measured in this paper and will be verified with simulation results in future publications. Finally, further development stages of the simulator together with simulation results will be reported in future works.

We have illustrated the modeling of various schemes of advanced modulation formats for optical transmission systems. Transmitter modules integrating lightwaves sources, electrical pre-coder and external modulators can be modeled without difficulty under MATLAB Simulink. As the popularity of MATLAB becoming a standard computing language for academic research institutions throughout the world, the models reported here would contribute to the wealth of computing tools for modeling optical fiber transmission systems and teaching undergraduates at senior level and postgraduate research scholars. The models can integrate photonic filters or other photonic components using blocksets available in Simulink. Furthermore we have used the developed models to assess the effectiveness of the models by evaluating the simulated results and experimental transmission performance of long haul advanced modulation format transmission systems.

11. References

[1] L. N. Binh, “Tutorial Part I on optical systems design,” ECE 4405, ECSE Monash University, Australia, presented at ICOCN 2002.

[2] T. Kawanishi, S. Shinada, T. Sakamoto, S. Oikawa, K. Yoshiara, and M. Izutsu, “Reciprocating optical modulator with resonant modulating electrode,” Electronics Letters, Vol. 41, No. 5, pp. 271−272, 2005.

[3] R. Krahenbuhl, J. H. Cole, R. P. Moeller, and M. M. Howerton, “High-speed optical modulator in LiNbO3 with cascaded resonant-type electrodes,” Journal of Lightwave Technology, Vol. 24, No. 5, pp. 2184−2189, 2006.

[4] I. P. Kaminow and T. Li, “Optical fiber communications,” Vol. IVA, Elsevier Science, Chapter 16, 2002.

[5] G. P. Agrawal, “Fiber-optic communications systems,” 3rd edition, John Wiley & Sons, 2001.

[6] J. B. Jeunhomme, “Single mode fiber optics,” Principles and Applications, 2nd edition: Marcel Dekker Pub, 1990.

[7] E. E. Basch, (Editor in Chief), “Optical-fiber transmission,” 1st edition, SAMS, 1987.

[8] I. P. Kaminow and T. Li, “Optical fiber communications,” Vol. IVB, Elsevier Science (USA), Chapter 5, 2002.

[9] J. P. Gordon and H. Kogelnik, “PMD fundamentals: Polarization mode dispersion in optical fibers,” PNAS, Vol. 97, No. 9, pp. 4541−4550, April 2000.

[10] Corning. Inc, “An introduction to the fundamentals of PMD in fibers,” White Paper, July 2006.

[11] A. Galtarossa and L. Palmieri, “Relationship between pulse broadening due to polarisation mode dispersion and differential group delay in long singlemode fiber,” Electronics Letters, Vol. 34, No. 5, March 1998.

[12] J. M. Fini and H. A. Haus, “Accumulation of polarization-mode dispersion in cascades of compensated optical fibers,” IEEE Photonics Technology Letters, Vol. 13, No. 2, pp. 124−126, February 2001.

[13] A. Carena, V. Curri, R. Gaudino, P. Poggiolini, and S. Benedetto, “A time-domain optical transmission system simulation package accounting for nonlinear and polarization-related effects in fiber,” IEEE Journal on Selected Areas in Communications, Vol. 15, No. 4, pp. 751−765, 1997.

[14] S. A. Jacobs, J. J. Refi, and R. E. Fangmann, “Statistical estimation of PMD coefficients for system design,” Electronics Letters, Vol. 33, No. 7, pp. 619−621, March 1997.

[15] E. A. Elbers, International Journal Electronics and Communications (AEU) 55, pp 195−304, 2001.

[16] T. Mizuochi, K. Ishida, T. Kobayashi, J. Abe, K. Kinjo, K. Motoshima, and K. Kasahara, “A comparative study of DPSK and OOK WDM transmission over transoceanic distances and their performance degradations due to nonlinear phase noise,” Journal of Lightwave Technology, Vol. 21, No. 9, pp. 1933−1943, 2003.

[17] H. Kim, “Differential phase shift keying for 10-Gb/s and 40-Gb/s systems,” in Proceedings of Advanced Modulation Formats, 2004 IEEE/LEOS Workshop on, pp. 13–14, 2004.

[18] P. J. T. Tokle., “Advanced modulation fortmas in 40 Gbit/s optical communication systems with 80 km fiber spans,” Elsevier Science, July 2003.

[19] Elbers, et al., International Journal of Electronics Communications (AEU) 55, pp 195–304, 2001.

[20] J. P. Gordon and L. F. Mollenauer, “Phase noise in photonic communications systems using linear amplifiers,” Optics Letters, Vol. 15, No. 23, pp. 1351–1353, December 1990.

[21] G. Jacobsen, “Performance of DPSK and CPFSK systems with significant post-detection filtering,” IEEE Journal of Lightwave Technology, Vol. 11, No. 10, pp. 1622–1631, 1993.

[22] A. F. Elrefaie and R. E. Wagner, “Chromatic dispersion limitations for FSK and DPSK systems with direct detec tion receivers,” IEEE Photonics Technology Letters, Vol. 3, No. 1, pp. 71–73, 1991.

[23] A. F. Elrefaie, R. E. Wagner, D. A. Atlas, and A. D. Daut, “Chromatic dispersion limitation in coherent lightwave systems”, IEEE Journal of Lightwave Technology, Vol. 6, No. 5, pp. 704–710, 1988.

[24] D. E. Johnson, J. R. Johnson, and a. H. P. Moore, “A handbook of active filters,” Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[25] J. G. Proakis, “Digital communications,” 4th edition, New York: McGraw-Hill, 2001.

[26] W. H. Tranter, K. S. Shanmugan, T. S. Rappaport, and K. L. Kosbar, “Principles of communication systems simulation with wireless applications,” New Jersey: Prentice Hall, 2004.

[27] P. D. M. Macdonald, “Analysis of length-frequency distributions,” in Age and Growth of Fish,. Ames Iowa: Iowa State University Press, pp. 371–384, 1987.

[28] P. D. M. Macdonald and T. J. Pitcher, “Age-groups from size-frequency data: A versatile and efficient method of analyzing distribution mixtures,” Journal of the Fisheries Research Board of Canada, Vol. 36, pp. 987–1001, 1979.

[29] E. F. Glynn, “Mixtures of Gaussians,” Stowers Institute for Medical Research, February 2007, http://research.stowers-institute.org/efg/R/Statistics/MixturesOfDistributions/index.htm.

[30] K. P. Ho, “Performance degradation of phase-modulated systems due to nonlinear phase noise,” IEEE Photonics Technology Letters, Vol. 15, No. 9, pp. 1213–1215, 2003.

Appendix: A Matlab program of the split-step propagation of the guided lightwave signals

```
function output = ssprop_matlabfunction_raman(input)

nt = input(1);
u0 = input(2:nt+1);
dt = input(nt+2);
dz = input(nt+3);
nz = input(nt+4);
alpha_indB = input(nt+5);
betap = input(nt+6:nt+9);
gamma = input(nt+10);
P_non_thres = input(nt+11);
maxiter = input(nt+12);
tol = input(nt+13);
%Ld = input(nt+14);
%Aeff = input(nt+15);
%Leff = input(nt+16);

tic;
%tmp = cputime;

%----------------------------------------------------------
%----------------------------------------------------------
% This function ssolves the nonlinear Schrodinger equation for
% pulse propagation in an optical fiber using the split-step
% Fourier method
%
% The following effects are included in the model: group velocity
% dispersion (GVD), higher order dispersion, loss, and self-phase
% modulation (gamma).
%
% USAGE
%
% u1 = ssprop(u0,dt,dz,nz,alpha,betap,gamma);
% u1 = ssprop(u0,dt,dz,nz,alpha,betap,gamma,maxiter);
% u1 = ssprop(u0,dt,dz,nz,alpha,betap,gamma,maxiter,tol);
%
% INPUT
%
% u0 - starting field amplitude (vector)
% dt - time step - [in ps]
% dz - propagation stepsize - [in km]
% nz - number of steps to take, ie, ztotal = dz*nz
% alpha - power loss coefficient [in dB/km], need to convert to linear to have P=P0*exp(-alpha*z)
% betap - dispersion polynomial coefs, [beta_0 ... beta_m] [in ps^(m-1)/km]
% gamma - nonlinearity coefficient [in (km^-1.W^-1)]
% maxiter - max number of iterations (default = 4)
% tol - convergence tolerance (default = 1e-5)
%
% OUTPUT
%
```

```
% u1 - field at the output
%---------------
% Convert alpha_indB to alpha in linear domain
%---------------
alpha = 1e-3*log(10)*alpha_indB/10;          % alpha (1/km) - see Agrawal p57
%---------------
%P_non_thres = 0.0000005;

ntt = length(u0);
w = 2*pi*[(0:ntt/2-1),(-ntt/2:-1)]'/(dt*nt);
%t = ((1:nt)'-(nt+1)/2)*dt;

gain = numerical_gain_hybrid(dz,nz);

for array_counter = 2:nz+1
    grad_gain(1) = gain(1)/dz;
    grad_gain(array_counter) = (gain(array_counter)-gain(array_counter-1))/dz;
end
gain_lin = log(10)*grad_gain/(10*2);

clear halfstep
  halfstep = -alpha/2;
    for ii = 0:length(betap)-1;
        halfstep = halfstep - j*betap(ii+1)*(w.^ii)/factorial(ii);
    end

     square_mat = repmat(halfstep, 1, nz+1);
     square_mat2 = repmat(gain_lin, ntt, 1);
     size(square_mat);
     size(square_mat2);
     total = square_mat + square_mat2;

 clear LinearOperator
    % Linear Operator in Split Step method
    LinearOperator = halfstep;
    halfstep = exp(total*dz/2);
 u1 = u0;
ufft = fft(u0);
% Nonlinear operator will be added if the peak power is greater than the
% Nonlinear threshold
iz = 0;
while (iz < nz) && (max((gamma*abs(u1).^2 + gamma*abs(u0).^2)) > P_non_thres)
  iz = iz+1;

  uhalf = ifft(halfstep(:,iz).*ufft);

  for ii = 1:maxiter,
    uv = uhalf .* exp((-j*(gamma)*abs(u1).^2 + (gamma)*abs(u0).^2)*dz/2);
    ufft = halfstep(:,iz).*fft(uv);
    uv = ifft(ufft);

    if (max(uv-u1)/max(u1) < tol)
      u1 = uv;
      break;
    else
      u1 = uv;
    end

  end
 % fprintf('You are using SSFM\n');
  if (ii == maxiter)

   fprintf('Failed to converge to %f in %d iterations',tol,maxiter);
  end

  u0 = u1;
 end

if (iz < nz) && (max((gamma*abs(u1).^2 + gamma*abs(u0).^2)) < P_non_thres)

%   u1 = u1.*rectwin(ntt);
    ufft = fft(u1);
    ufft = ufft.*exp(LinearOperator*(nz-iz)*dz);
    u1 = ifft(ufft);

    %fprintf('Implementing Linear Transfer Function of the Fibre Propagation');
end
```

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Geospatial Information Service Based on Ad Hoc Network

Fuling BIAN, Yun ZHANG
International School of Software, Research Center of Spatial Information and Digital Engineering,
Wuhan University, Wuhan, China
Email: *zhangyun*604*@yahoo.com.cn*

Received January 16, 2009; *revised February* 10, 2009; *accepted February* 18, 2009

Abstract

The mobile geospatial information service involves the domain of mobile communication, mobile computing, geospatial information service and other techniques. This paper focuses on the integration of spatial information and mobile communication technologies. The author proposes the architecture of mobile geospatial information service based on the Ad Hoc network. On the basis of this architecture, a system is developed, and applied in correlative fields.

Keywords: Spatial Data, Ad Hoc Network, Geospatial Information Service

1. Introduction

With the development of electronic technology, especially in the fields of computer and internet technology, mobile communication has made a rapid progress in recent years. More and more mobile phones with powerful functions have emerged in our daily life. At the same time, cell phone users are not satisfied with the only function of call service while more and more mobile services have emerged. So the proportion of mobile services rate in the communications revenue has increased year by year.

According to statistics in the resource management, socio-economic activities and public life, more than 80 percent of information involves in the geographic information with spatial localization feature. The fast development of technological and subject matter has got much attention in recent years, which are used as an important means of communication for acquiring, arranging, analyzing and managing geospatial data in productive activities and Geographic Information System. Meanwhile, the users are not satisfied with acquiring GIS Services in the use of computer equipments and Geographic Information System. Nowadays, computer network and wireless communication technology have become so widespread. It offers a new direction in the development of GIS.

2. Mobile Geographic Information Service

Geographic Information System (GIS for short) is a kind of technology for saving and analyzing spatial information, integrated with computer graphics and data. GIS is the spatial information system that has been compiled in purpose of specific application. Under the help of computer hardware, software facilities and network, it is user for collect, input, store, process, display, update and provide kinds of dynamic geographic information application, with the method of geographic models [1].

GIS has passed through two great changes: from the stage of GISystem to GIScience, then transferred to the stage of GIService [2,3]. In modern society, network technology and wireless communication technology are developing rapidly. With the development of GIS and the two technologies, a new spatial information service and application model was born, i.e. (Mobile Geographic Information Service, MGIS) [4]. MGIS offers mobile sharing with geographic information in the integration of computer, telecommunication and 3S technologies. It can help the users to acquire the maximum amount of information in a finite-range space. At the same time, it truly meets the users' needs in geography information query and decision making. It is an important content of GIS-ervice research that how to make users satisfied with geographic information conveniently geo-information for anyone and anything at anywhere and anytime [5].

Nowadays, Internet-centered information network has basically satisfied the need for information in anytime. However, Internet builds on fixed cable network failed to meet the need for acquiring the information services in mobile.

In the research of mobile geospatial information service, we should take such factors into account as:

1) The limited bandwidth and frequently disconnection in wireless networks

The disconnection and transmission delay don't need consideration in geospatial information service in traditional networks. The delay, bandwidth, signal intensity and communication protocol may be changing in the wireless networks. As a result，mobile geospatial information service must be fit for the technique status in the face of communication field in modern time.

2) Enhancement adaptability of service providing methods

Compare to cable network，wireless network has a more flexible deployment. It has many different ways in which can be connected to the network: from Infrared Ray to Bluetooth within a few-mile limit, from WLAN to satellite relay and other communication modes covering the whole world. All these factors can be considered in application system design. What's more, the cable can be used as an auxiliary method.

3) Low security in data transmission

Mobile geospatial information service applies electromagnetic wave to send or receive data through the air. It can work without cable and other transmission media. But we have to note that the process is easily perturbed by external environment because of transmission media's vulnerability. Meanwhile, the data transferred will be stolen by illegal users for uncertainty in the transmission process. Therefore, we should try to avoid long distance wireless transmission in the architecture design.

4) Limited energy and resources of mobile terminal

The mobile terminal is supplied from storage batteries. Due to the requirement of power supply, mobile equipment can not work for a long time. On one hand, the resource in mobile terminal would differ from fixed hosts in the increase of portability. On the other hand, it has a limited computational power. As a result, rational allocation of the data is necessary, especially between mobile terminal and fixed hosts. In this way, it can enhance the rates of resource utilization efficiency in querying, processing and store management.

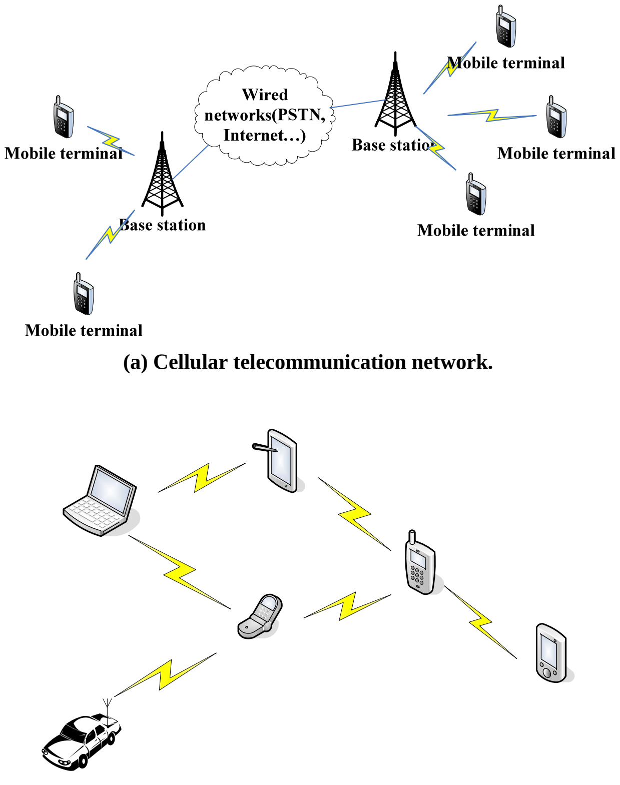
5) Multiplicity in architecture design

Traditional systems are mostly based on Client/Server mode, as indicated in Figure 1. While in mobile environment，dynamic computing resource and network joint nodes are full of uncertainty. The mobile terminal is of versatile features and configuration. According to difference between wireless networks connectivity and the type and capability of mobile terminal equipment, mobile geospatial information service should design different architecture to achieve the full potential of hardware and networks [6].

The basic network configuration of mobile geospatial information service and application environment have experienced great changes for traditional geospatial information service. Although wireless network is the similar as the extension of wired network, there is a big difference between them both in technologies and system design. So research on mobile geospatial information service must make a concrete analysis for these features.

3. Ad Hoc Network

Wireless network offers the users the data exchange and acquisition in anytime and any place. And it has the function of keeping connecting in moving state. In general, mobile communication network has a center. It only works in presupposition based network establishment. In some special occasions, such as on the battlefields，the rescue attempt after flood and earthquake, field operations and so on, traditional centered cellular networks can not meet the requirements of communication. In this way, a new network technology-Ad Hoc network has emerged, which need not rely on fixed telecommunication.



(a) Cellular telecommunication network.

(b) Ad hoc network.

Figure 1. The comparison of cellular telecommunication network and ad hoc network.

Ad Hoc network is made up of a group of wireless communication and connecting network module, without any fundamental establishment. When the cable network can not work in order, like on the battlefield and in the task of emergency relief, Ad Hoc network offers a feasible technology of communication and information storage. When two mobile terminals are in each other's covering range, they can communicate directly in Ad Hoc network [7]. But the communication coverage of mobile terminal is limited. The two hosts which are far away from each other can be linked by other nodes' forwarding.

In Ad Hoc network, the moving node changes the topology of network. Routing in ad hoc wireless networks has to cope with specifics such as limited bandwidth, high error rates, dynamic topology, resource poor devices, power constraints and hidden and exposed terminal problems. It is the essential part for Ad Hoc network that how to choose the routes to destination accurately and quickly. The typical routing algorithms are as follows [8].

DSDV (Destination-Sequenced Distance-Vector) is a Table driven routing protocol based on the classical Bellman-Ford routing algorithm. by improving freedom from loops in routing tables by using sequence numbers.

DSR (Dynamic Source Routing) generates routes only when desired. When a packet needs to be sent to a destination whose route doesn't exist, a route discovery process is initiated. If the route has been existed, a route maintenance procedure is invoked. AODV (Ad hoc On-demand Distance Vector Routing) is a pure on-demand routing protocol. It is based on DSDV protocol.

4. Geospatial Information Service Architecture Based on Ad Hoc Network

In wireless environment of geographic information service, spatial data transmission focuses on wireless network and mobile equipments in different environments.

The transmission scale of wireless local area network is relatively small. If we want to enlarge the transmission coverage, we can apply the method of adding antennas. The user end can work in the mode of Ad Hoc if a user wants to set up his own network without the help of access points. As for the problem faced by mobile geographic information service, it is necessary to apply Ad Hoc network into the architecture, as indicated in Figure 1.

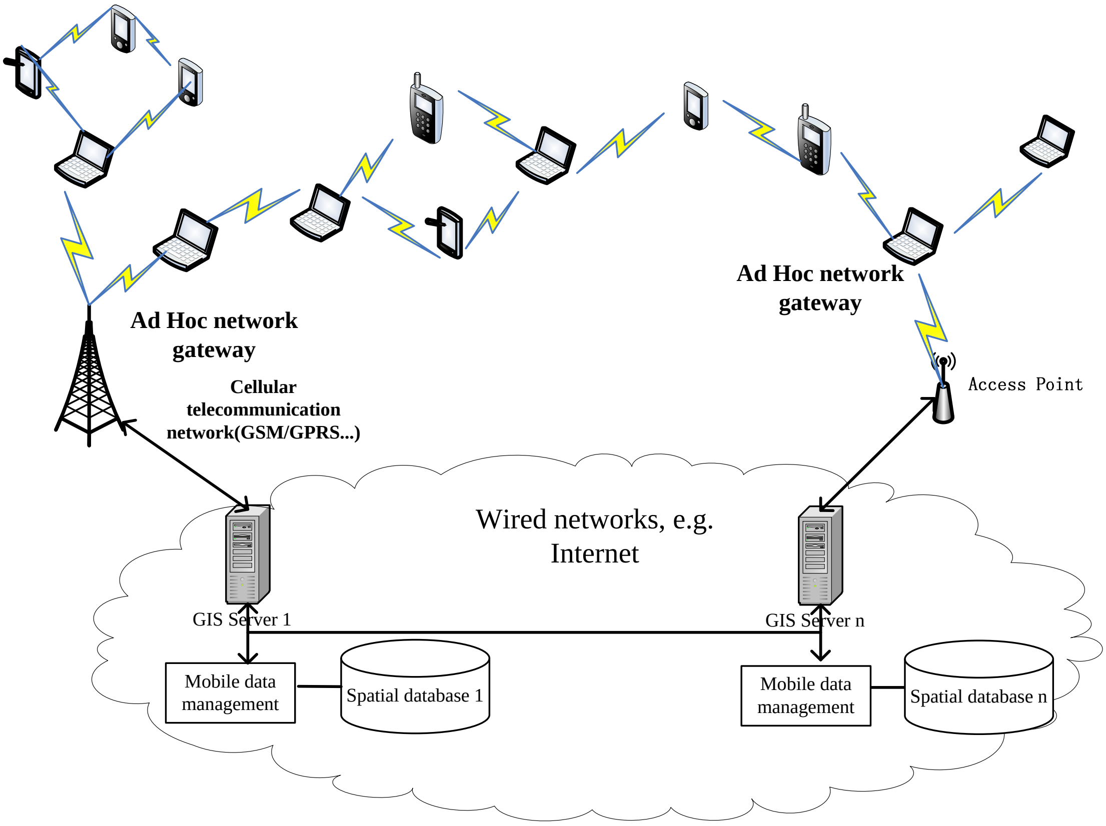


Figure 2. Geospatial information service architecture based on ad hoc network.

As indicated in Figure 1，geospatial information service is made up of mobile terminal, mobile unit, mobile supporting station, fixed host and wired networks.

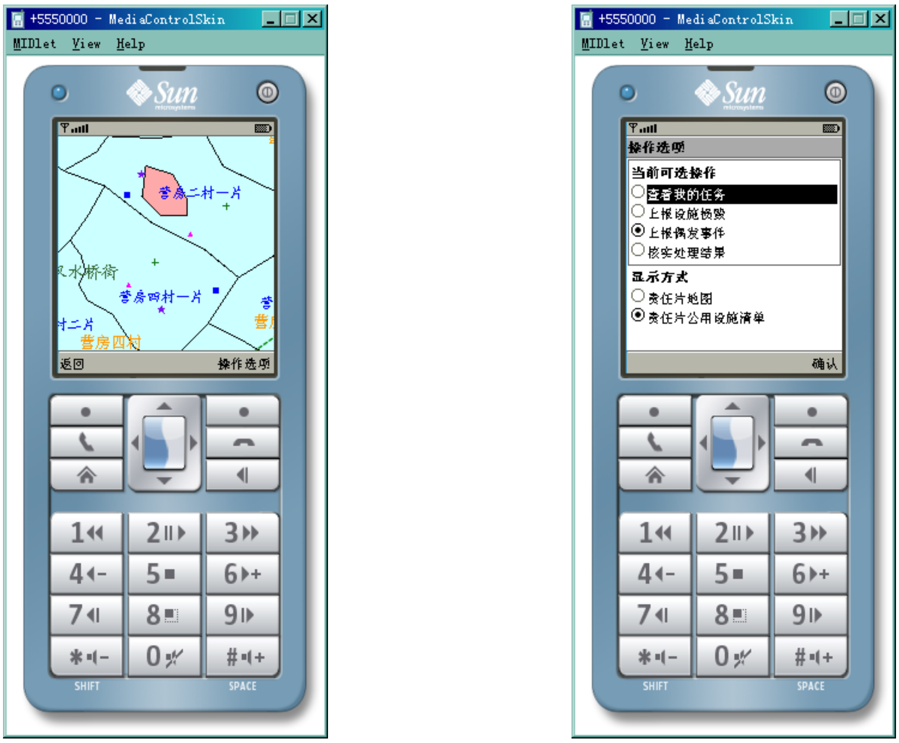
The data transmission among mobile nodes in Ad Hoc network adopts the advanced technology of distributed computing. It can break the limits of transmission distance in traditional WLAN. High speed wired network forms the connecting backbone with fixed nodes including file server, communications server, file server and database service. Wireless data transmission by mobile terminals with the Ad-Hoc routing mode is an effective method to realize wireless GIS. With the help of this technology, it can offer a network linked by many wireless node in short distance, the nodes in large scale can also be linked by multi-hop in nearby nodes to build a large Ad hoc network. In addition, traditional wireless Ad Hoc routing protocols needs to be optimized in application to mobile geospatial information service.

First the mobile terminal needs to find the Ad Hoc network gateway to connect with the cellular telecommunication network or Internet. The network gateway periodically transmits package all around to broadcast own existence. If the mobile terminal is within one-hop distance from the gateway the, the mobile terminal may find the mobile gateway through the received broadcast information. If the distance is beyond one hop, the mobile terminal will send the control package to seek the gateway, then this mobile terminal will receive from one or more than one reply from the Ad Hoc network gateway. The mobile terminal will select the gateway with least hop number to connect with the cellular telecommunication network or Internet. By using the gateway router, several Ad Hoc network may be interconnected to enable the mobile node to reach Internet.

5. The Experimental System

Based on the above architecture, the author carries out a platform of urban public administration and service information management. It is designed to gather the urban management related data, position the accurate location of emergencies, carry on the effective dispatch promptly, direct the related personnel to process the accident. The client interface on intelligent mobile phone is indicated as in Figure 3.

The success of experimental system proves that the spatial data transmission mechanism among the mobile terminals by Ad Hoc network proposed in our paper is feasible. From the angle of experiment, it is feasible to prove the rationality of this architecture. It not only can save the resources of cellular telecommunication system, but also overcome the limit of geographic information service implementation created by mobile environment.



操作选项

当前可选操作
○ 查看我的任务
○ 上报设施损毁
● 上报偶发事件
○ 核实处理结果

显示方式
○ 责任片地图
● 责任片公用设施清单

Figure 3. The client interface on intelligent mobile phone.

6. Conclusions

The emphasis of the research on mobile geospatial information service is how to organize and transmit the kinds of geographic information with less network resource and higher transmission efficiency. This paper explores the geospatial information service based on Ad Hoc network aiming to realize the geographic data sharing among mobile terminals, furtherly to increase the data safety and quality in large-scale wireless data transmission process. But there are both advantages and disadvantages on kinds of protocols in network performance. We need to further our research on how to optimize the protocols to obtain the better network performance according to the features of mobile geospatial information service.

7. References

[1] S. P. Chen, "Urbanization and urban GIS," Science Press, 2001.

[2] J. W. Shi, K. W. Kwan, J. N. Cao, *et al*., "A proactive approach for mobile GIS [C]," Proceedings of IEEE 58th Vehicular Technology Conference, Vol. 2, pp. 1000–1004, 2003.

[3] H. Wen, "Design and implementation of a prototype mobile GIS based on J2EE/J2ME," Beijing, Peking University, 2005.

[4] C. H. Peng, Y. F. Liu, L. Yan, J. Y. Liu, and J. H. Zheng, "Research on key techniques of Java-based mobile geographic information service," Computer Engineering and Applications, No. 11, 2007.

[5] D. R. Li, "The development of RS and GIS in the 21st century," Geomatics and Information Science of Wuhan University, No.2, pp. 127–131, 2003.

[6] M. Xu, "Mobile computing technology," Tsinghua University Press, September 2008.

[7] S. R. Zheng and H. T. Wang, "Ad hoc network technology," Post and Telecom Press, October 2005.

[8] K. U. R. Khan, R. U. Zaman, and A. V. Reddy, "Performance comparison of on-demand and table driven ad hoc routing protocols using NCTUns," Proceedings- UKSim 10th International Conference on Computer Modelling and Simulation, EUROSIM/UKSim2008.

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Energy Aware Clustered Based Multipath Routing in Mobile Ad Hoc Networks

M. BHEEMALINGAIAH[1], M. M. NAIDU[2], D. Sreenivasa RAO[3]
[1]*Dept. of Computer Science and Engineering J.N.T University, Hyderabad, India*
[2]*Dept. of Computer Science and Engineering, S.V.U College of Engineering, Tirupati, India*
[3]*Dept. of Electronics and Communication Engineering, J.N.T University, Hyderabad, India*
Email: *saibheem2008@gmail.com, mmnaidu@yahoo.com, dsraoece@yahoo.co.uk*

Received January 10, 2009; *revised February* 21, 2009; *accepted February* 23, 2009

Abstract

With the advance of wireless communication technologies, small-size and high-performance computing and communication devices are increasingly used in daily life. After the success of second generation mobile system, more interest was started in wireless communications. A Mobile Ad hoc Network (MANET) is a wireless network without any fixed infrastructure or centralized control; it contains mobile nodes that are connected dynamically in an arbitrary manner. The Mobile Ad hoc Networks are essentially suitable when infrastructure is not present or difficult or costly to setup or when network setup is to be done quickly within a short period, they are very attractive for tactical communication in the military and rescue missions. They are also expected to play an important role in the civilian for as convention centers, conferences, and electronic classrooms. The clustering is an important research area in mobile ad hoc networks because it improves the performance of flexibility and scalability when network size is huge with high mobility. All mobile nodes operate on battery power; hence, the power consumption becomes an important issue in Mobile Ad hoc Network. In this article we proposed an Energy Aware Clustered-Based Multipath Routing (EACMR), which forms several clusters, finds energy aware node-disjoint multiple routes from a source to destination and increases the network life time by using optimal routes.

Keywords: Clustering, CONID, MANET, Mutlipath, AODVM

1. Introduction

The history of wireless networks started in the 1970s and the interest has been growing ever since. Based on infrastructure, the wireless networks broadly classified into two types, first type infrastructure networks contains base-stations; an example of this wireless network is the cellular-phone network where a phone connects to the base-station with the best signal quality. When the phone moves out of range of a base-station, it does a "hand-off" and switches to a new base-station within reach, the second type is called as Mobile Ad hoc Network without any fixed infrastructure or centralized control; it contains mobile nodes that are connected in an arbitrary manner. It enables the users to communicate without any physical infrastructure regardless of their geographical location. All nodes in the Mobile Ad hoc Network (MANET) are mobile and can be connected dynamically in an arbitrary manner. Each node behaves as a router, takes part in discovery and maintains the routes to other nodes. The nodes are the main components of the network; these nodes can move freely at any time and can leave or join the network so the network structure changes dynamically due to mobility [1].

1.1. Clustering and Multipath Routing

In a clustering scheme the mobile nodes in a MANET are divided into different virtual groups based on certain rules. The mobile nodes may be assigned a different status such as clusterhead, clustergateway or clustermember. A clusterhead normally serves as a local coordinator for its cluster, performing intra-cluster transmission arrangement, data forwarding, and so on. A clustergateway is a non-clusterhead node with inter-cluster links, so it can access neighboring clusters and forward information between clusters. A clustermember is usually called an ordinary node [2].

A fundamental problem in the MANET is how to deliver data packets among nodes efficiently without predetermined topology or centralized control, which is the main objective of ad hoc routing protocols. Because of the dynamic nature of the network, ad hoc routing faces many unique problems not present in wired networks. Particularly in MANETs where routes become obsolete frequently because of mobility and poor wireless link quality. The Multipath routing addresses these problems by providing more than one route to a destination node. Multipath routing appears to be a promising technique for ad hoc routing protocols, the multiple paths can be useful in improving the effective bandwidth of communication, responding to congestion and heavy traffic, increasing delivery reliability and security. The traffic can be distributed among multiple routes to enhance transmission reliability, provide load balancing, and secure data transmission [3].

Most existing routing protocols for MANET build and utilize only one single route for each pair of source and destination nodes. Due to node mobility, node failures, and the dynamic characteristics of the radio channel, links in a route may become temporarily unavailable and making the route invalid. The overhead of finding alternative routes may be high and extra delay in packet delivery may be introduced. Multipath routing addresses this problem by providing more than one route to a destination node. Source and intermediate nodes can use these routes as primary and backup routes. Alternatively, source node can distribute traffic among multiple routes to enhance transmission reliability, provide load balancing, and secure data transmission.

2. Related Work

This work falls under three areas in the MANET: clustering algorithms, multipath routing protocols and energy aware routing protocols, this section reviews these areas focusing on their relationship to the proposed protocol.

Jane Y. *et al.* [2] surveyed the different clustering mechanisms and described the advantages and disadvantages of each clustering scheme. The clustering schemes broadly classified into six categories Dominating Set based clustering schemes, Low maintenance clustering schemes, Mobility aware clustering schemes, Energy efficient clustering schemes, Load balancing clustering schemes and Combined-metrics based clustering schemes. The Energy efficient clustering avoids unnecessary energy consumption or balancing energy consumption for mobile nodes in order to prolong the lifetime of mobile nodes and the network.

The on-demand routing is the most popular approach in the MANET. Instead of periodically exchanging route messages to maintain a permanent route table of the full topology, the on-demand routing protocols build routes only when a node needs to send the data packets to a destination. The standard protocols of this type are Dynamic Source Routing (DSR) [4] and the Ad hoc On-demand Distance Vector (AODV) routing [5]. However, these protocols do not support multipath.

The several multipath on-demand routing protocols were proposed, some of the standard protocols are, the Ad hoc On-demand Multipath Distance Vector (AOMDV) [6] is an extension to the AODV protocol for computing multiple loop-free and link-disjoint paths. The Split Multipath Routing (SMR) [7] is an on-demand Multipath source routing protocol can find an alternative route that is maximally disjoint from the source to the destination. The Multipath Source Routing (MSR) [8] is an extension of the DSR protocol to distribute traffic among multiple routes in a network. The Ad hoc On-demand Distance Vector Multipath Routing (AODVM) [9] is an extension to the AODV for finding multiple node disjoint paths. These protocols build multiple routes based on demand but they did not consider energy.

Several energy aware multipath on-demand routing protocols have been proposed [10−14], but these protocols are not based on clustering. The Cluster Based Multipath Dynamic Source Routing CMDSR) [15] was designed to be adaptive according to network dynamics. It uses the hierarchy to perform route discovery and distributes traffic among diverse multiple paths, but it does not consider energy in the route selection.

The Grid-Based Energy Aware Node-Disjoint Multipath Routing Algorithm (GEANDMRA) [16] considers energy aware and node-disjoint multipath, it uses grid head election algorithm to select the grid-head which is responsible for forwarding routing information and transmitting data packets. The routing is performed in a grid-by-grid manner, the network area is partitioned into non-overlapping square zones with the same size, each zone is named with a unique ID(x,y) as the conventional coordinate. At any time each node can obtain its location information from Global Position System (GPS) to know in which zone it is located. But the disadvantages of GPS are as follows [17].

- There are several factors that introduce error to GPS position calculations. A major source of error arises from the fact that radio signal speed is constant only in a vacuum. it means that distance measurements may vary as the values of the signal speed vary in the atmosphere. Water vapor and other particles in the atmosphere can slow signals down, resulting in *propagation delay*.
- Another source of errors due to *multipathfading*, which occurs when a signal bounces off a building

or terrain before reaching the receiver's antenna, also can reduce accuracy.

- Another factor affecting the precision is satellite geometry.
- The largest source of potential error is *selective availability*, an intentional degradation of L1, the civilian GPS signal. SA was originally intended to prevent a hostile force or terrorist group from exploiting the technology.
- An location fix can only be identified if a 'fix' from at least 3 satellites are available.
- In addition, distance measurements are less reliable when the receiver of satellite locks on to or closely oriented with respect to each other. Atomic clock discrepancies, receiver noise, and interruptions to ephemeris monitoring can result in minor errors.
- GPS does not work effectively if the 'line of sight' between a device and the satellites is obscured. Not only does this render GPS ineffective indoors but also if any material gets in the way of this 'line of sight'. It is fair to assume therefore that a worker in a vehicle or in a station or operating in a dense urban area with tall buildings may struggle to get a location fix from GPS.
- The device needs a dedicated GPS antenna, chipset and software, so there is a additional cost for this hardware and software.
- Power consumption in the device is significant for using GPS.

We proposed novel approach without using GPS; this approach is very suitable and applicable for the cluster based mobile ad hoc networks.

3. Energy Aware Clustered-Based Multi-path Routing (EACMR)

The HELLO messages are already in usage in on demand routing, each node maintains only the status of its neighbors. Periodically, each node broadcasts a hello message to all its neighbors to indicate its active status. Each node receives the hello messages from its neighbors and updates its neighbor information table. In the EACMR, the clusters are formed by using hello messages, this result in less overhead. The EACMR finds node-disjoint multiple routes from a source to destination and increases the network life time by using optimal routes.

3.1. Network Model

A MANET is represented by undirected graph, $G=(V,E)$ where V is the set of nodes and E is the set of bidirectional links. It is assumed that, at any time, nodes are situated randomly throughout the network area in accordance with a two-dimensional uniform random variable distribution. Each node is equipped with a single network interface card (NIC) and has a transmission radius of r. Each node has mobility, the speed is uniformly chosen between the minimum and maximum speeds. When the node reaches to its destination, it stays there for a certain pause time, after which it chooses another random destination point and repeats the process, the mobility is defined as the distance moved per unit time by a node in the network. Suppose at time t_1 the node n_i is at (x_1, y_1) and by time t_2 the node n_i has moved to (x_2, y_2), then the mobility of the node n_i denoted by

$$M_{n_i} = \frac{1}{(t_2 - t_1)}\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

All mobile node relay on battery, the energy consumption varies from 240mA at receiving mode and 280mA in the transmitting mode using 0.5V energy. Thus,when calculating the energy consumed to transmit a packet p is $E_{tx}(p)=I\times V\times t_p$. Joules are needed [18], here, I is the current, V is the voltage and t_p is the time taken to transmit the packet p.

The energy required to transmit a packet p is given by $E_{tx}(p)=280mA\times V\times t_p$.The energy is required to receive a packet p is given by $E_{rx}(p)=240mA\times V\times t_p$. The energy consumption of overhearing the data transmission may be assumed as equivalent to energy consumption of receiving of the packet. When a packet is transmitted by n_i at time t then it updates its residual battery capacity by using the following formula $R_i^t = R_i^t - E_{tx}(p)$ .when packet is received by n_i at time t then it updates its residual battery capacity by using the following formula $R_i^t = R_i^t - E_{rx}(p)$.

3.2. Cluster Formation and Maintenance

The Combined Higher Connectivity Lower ID (CONID) clustering algorithm [19] is used generates the clusters in the network. It is an extension of the lowest ID algorithm; the lowest ID algorithm does not take into account the connectivity (degree) of nodes, and therefore may produce more number of clusters than necessary. The pure connectivity based clustering algorithm modified version the lowest ID algorithm, in which ID is replaced by node degree, but it does not work properly because of numerous ties between nodes. The CONID uses node degree as the primary key and ID as the secondary key in cluster decisions. It is generalized the connectivity to count all k-hop neighbors of the given node. For k=1, the connectivity is equivalent to node degree, whenever the connectivities are the same, IDs are compared to make the decision.The clustering algorithm, refereed to as the k-CONID (k-hop connectivity ID) algorithm, works as follows.

Each node is assigned with clusterhead priority. A pair is denoted by did=(d,ID), where d is its connectivity and ID is its IP address.

Let did'= (d', ID') and did"= (d", ID"). Then did'>did' if d'> d" or d'=d" and ID' < ID". That is, a node has clusterhead priority over the other node if it has higher connectivity or in case of equal connectivity and has lower ID. The cluster formation and cluster maintenance are clearly explained in [19], after applying clustering algorithm the network is shown Figure 1. After running CONID, each clusterhead finds its all neighbor clusterheads, for example, clusterheads 18 and 10 are the neighbor clusterheads of clusterhead 2 in Figure 1.

3.3. Route Selection

The route selection is based on cost function; the main objective is to give more weight (or) cost to node with less energy to prolong its life time. Let R_i^t be the battery capacity of a node n_i at time t. Let $f_i(R_i^t)$ be the cost function of node n_i, it is also considered as node's cost or weight of node n_i at time t, the cost of node ni at time t is inversely propositional to its residual energy i.e. $f_i(R_i^t))= 1/R_i^t)$. As the battery capacity decreases, the value of cost for node n_i will increase. The main objective of route selection is to select an optimal path based on costs of clusterheads, because clusterhead normally serves as a local coordinator for its cluster, performing intra-cluster transmission arrangement, data forwarding. So clusterhead is important node within cluster and spends its energy for other nodes.

We define the following cost function to a clusterhead. Let C_i be *ith* cluster whose clusterhead at time t is denoted by CH_i. The Cost of CH_i at time t as follows

$$Cost(CH_i) = \rho_i \times \left[\frac{F_i}{R_i^t}\right] \times w_i \times N(C_i)$$

where

ρ_i: Transmit power of CH_i

F_i: Full-charge capacity of CH_i

R_i^t: Remaining battery capacity of CH_i at time t.

w_i: weight factor of CH_i, which depends upon various factors, like battery's quality, battery's capacity, life time, battery's back up, price.

$N(C_i)$:the size of cluster C_i, it is the total number of all the nodes(clustermembers, gateways and clusterhead) in C_i, it is directly proportional to the cost of the node.

Let P_j be the path from source node S to destination node D via clusterheads CH_1, CH_2,CH_3 ....CH_n at time t, it is denoted by

$$P_j = S - CH_1 - -CH_2 - -CH_3 - - - - - CH_n - D$$

On the above path, in between two clusterheads, either gateway or clustermember (non-clusterhead nodes) may exist, however they are not consider in the evaluation of cost of the path. Because main intension is to consider the total cost of path through clusterheads only, we define the cost of path P_j at time t via clusterheads is the sum of the costs of all the clusterheads on P_j, it is denoted by

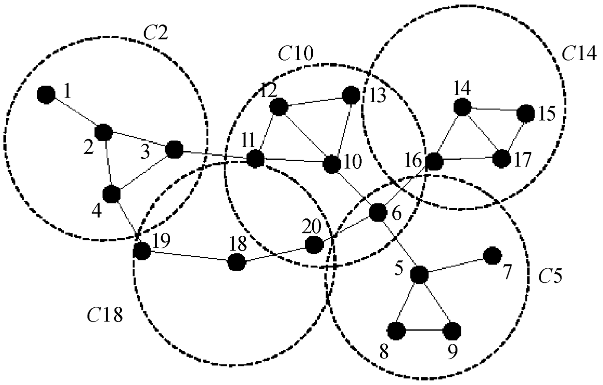


Figure 1. Network with clusters.

$$Cost(P_j) = \sum_{i=1}^{n} Cost(CH_i)$$

3.3.1 Optimal Path Selection

Let k be number of node disjoint paths from source S to destination D, an optimal path is a path whose cost is least among k paths.

The optimal path is also called primary path. Initially it is used for data transmission, if it fails, then the secondary path (whose cost is least among k-1 paths) is used for data transmission. For example in figure 2,there are three node disjoint paths from source S to destination D via clusterheads; the paths are 1–2–3–4–5, 1–7–6–5 and 1–8–9–10–5.

If source node S belongs to ith cluster whose clusterhead is CH_i then it is called as source clusterhead. Similarly, If destination D belongs to jth cluster whose clusterhead is CH_j then it is called as destination clusterhead in Figure 2. Clusterhead 1 is called as source clusterhead and clusterhead 5 is called as destination clusterhead. If

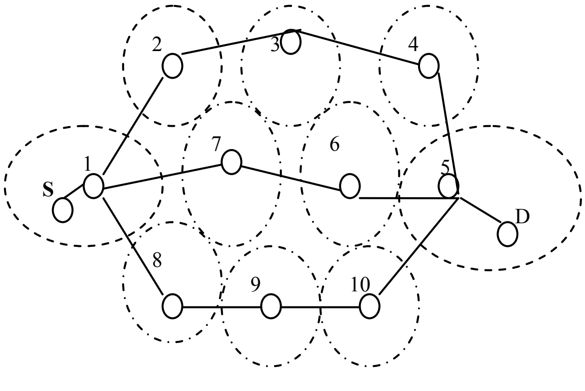


Figure 2. Node-disjoint paths through Clusterheads.

two clusterheads are neighbors then they are called as 1-hop neighbor clusterheads.In Figure 2, the clusterheads 3, 7, 8 are neighbors of the clusterhead 1.

3.4. Route Discovery

The route selection is based on the route discovery. In order to facilitate the computation of multiple node disjoint paths from the source to destination, We choose the Ad hoc On-Demand Distance Vector Multipath (AODVM) [9] protocol as a candidate protocol and make modifications to it to enable the discovery of node disjoint paths via clusterheads. The AODVM is the extension of AODV [6] to provide multiple nodes disjoint paths.

3.4.1. The Proposed Modifications

The proposed modifications are explained briefly as follows. Only clusterheads maintain routing tables and run this protocol to find nod-disjoint paths. The other nodes (Gatewaynodes and Clustermembers) don't maintain the routing tables, simply they forward the packets according to specified path.

3.4.2. Modification of Control Packets

The RREQ packet of AODVM is same as RREQ packet of EACMR. The RREP packet of AODVM is extended as RREP packet of EACMR by adding with ***Cost field***, this field carries cumulative costs of clusterheads through which it passes. The initial value of this field is zero.

3.4.3. Modification of Tables

In the AODVM, each node maintains two tables, the routing table is used to forward the data packets from source to destination where as the RREQ table is used to form the route from source to destination, the fields of both tables are shown in Figure 3 and Figure 4.

In the EACMR, the routing table is extended to include the cost field. But RREQ table is same. These two tables are maintained by clusterheads only.

3.4.4. The Proposed Modifications in Node Functioning

In the AODVM, when the source node wants to send packets to a destination, it looks up its route cache to determine if it already contains a route to the destination. If it finds that an unexpired route to the destination exists, then it uses this route to send the packet. But if the node does not have such a route, then it initiates the route discovery process by broadcasting a RREQ packet.

Dest.	Source	Last hop	Next hop

Figure 3. Fields of routing table.

Dest.	Source	Neighbor who transmitted the RREQ	hop	Expiration time

Figure 4. Fields of RREQ table.

It is modified that when the source node wants to send packets to a destination, it sends the RREQ to its clusterhead; clusterhead looks up its route cache (routing table) to determine if it already contains a route to the destination. If it finds that an unexpired route exists to the destination, then it sends the reply to the source and uses this route to forward the packets from the source node, else then it initiates the route discovery process by forwarding a RREQ packet to its all neighbor clusterheads.

In the EACMR, the propagation of RREQ packet at intermediate clusterhead follows the same rules as at intermediate node in AODVM. When RREQ packet arrives at intermediate clusterhead, The RREQ is inspected and the following two cases it is dropped.

1) If this RREQ packet was already processed by the intermediate clusterhead or

2) If TTL value of the RREQ packet becomes zero.

Otherwise, the intermediate clusterhead broadcasts the RREQ to all its neighbor clusterheads by recoding the information about the clusterhead through which RREQ was received and corresponding hop-count back to the source clusterhead into its RREQ table, in this way the RREQ is forwarded by a intermediate clusterheads. Finally several RREQs will reach to destination clusterhead through different paths from the source clusterhead.

When a RREQ packet is received to a destination clusterhead, it generates a RREP packet and sent back to its neighbor clusterhead (last hop) from which the RREQ packet has been received. When an intermediate clustehead receives the RREP packet, it checks its RREQ table to find its neighbor clusterhead (next hop in reverse path) through which shortest path back to the source clusterhead and sends the RREP packet to it by deleting corresponding entry in the RREQ table.

In order to ensure that a clusterhead does not participate in multiple paths, If it receives another RREP packet from different neighbor clusterhead, it is dropped (i.e., RREQ table is already empty), it generates a Route Discovery Error (RDER) packet and sends it back to the neighbor clusterhead through which another the RREP packet has been received. The neighbor clustered upon receiving the Route Discovery Error (RDER) packet will try to forward the RREP packet to another neighbor clusterhead through which shortest path back to the source clusterhead and sends the RREP packet to it by deleting corresponding entries in the RREQ table.

Finally several RREP packets will be received by the source clusterhead, this information is recorded into its cache based on arriving order and it sends the reply to the source .The path with minimum cost is selected as an optimal path for transmission of source's data. If clusterhead wants to transmit the data, it follows above procedure to find node disjoint paths.

3.5. Route Maintenance

The EACMR handles route maintenance in a manner similar to the AODVM. Whenever a link breakage happens in a route due to a node moving away, the previous hop node of the moved away node is responsible for sending a Route Error (RERR) message back to the source clusterhead to inform the breakage. It chooses alternative routes to maintain the connection. If there are no more redundant routes left, then it will start a new route discovery.

3.6. Congestion Control and Increasing Network Lifetime

When a congestion state occurs in a routing path, congested clusterhead sends choke packet to the source clusterhead, it distributes the incoming data packets to the other node-disjoint routing paths to avoid the congestion.

During the transmission of data packets, if battery energy of clusterhead is reached to threshold energy, then it selects a clustermember as new clusterhead whose energy is high among its clustermembers, then it sends information about cluster and it will act as clustermember.

4. Simulation

In this section, it describes the simulation, various chosen parameters for simulation and the various performance metrics. With reference to simulation model in [20], we have designed own simulator with the following modules.

Network Formation Module: This module is used to generate a random network, inputs of this module are simulation area (length x breadth), number of nodes, cell radius of each node, initial position of each node and initial energy of each node. The output of this module is a random network.

Node Mobility Module: This module sets the speed, direction and pause time of each node and allows each node to move in random direction. All the nodes in an ad hoc network are mobile. In this simulation, Random waypoint mobility model is used with pause time of each node is 10 sec and speed of each mobile node is 0 to 2m/sec.

Route Requests Event Generator Module: This module accepts the number of route requests from user, and then selects source and destination pairs randomly. Each route request follows the poison distribution process and each call duration time follows exponential distribution.

EACMR module: This is core module that incorporates several functions like Cluster formation, cluster maintenance, route discovery, route selection, route maintenance and congestion control.

Computation module: This module estimates various performance metrics like number of clusters, number of border nodes, the power consumption, residual energy, number of nodes expired (reached to threshold), overhead, throughput, end-to-end delay and other parameters.

Considering fifty nodes randomly distributed in an area of 1000m x 1000 m. It is assumed that the channel bandwidth is 2Mbps, a free space radio propagation model in which the signal power attenuates is $1/r^2$, where r is the distance between the nodes. Each node is equipped with a single network interface card and has a transmission radius of $r=14m$. The distributed coordination function of IEEE 802.11 is assumed at MAC layer.

All nodes operate in promiscuous mode, so it can overhear packets destined for others. It is assumed that the transmission power, receiving power are fixed for all the nodes and two nodes can hear each other if their distance is in the transmission range. The speeds are uniformly chosen between the minimum and maximum speeds and are set to 0m/s and 2m/s, respectively. When the node reaches its destination, it stays there for a certain pause time, after which it chooses another random destination point and repeats the process. The simulation ends after 100s. It is assumed that transmission ranges of all the nodes are equal. All nodes are assumed to have the same amount of battery capacity with full energy at the beginning of the simulation. Initial energy of each node is set to 100 Joules. The equal weight factor is chosen for all clusterheads. For the mobile scenarios, the random waypoint model is used to node mobility. In this model, a node chooses a random point in the network, and moves towards that point at a constant speed.

It is assumed that the number of route requests is denoted by λ and follows the poison process where as call holding time follows exponential distribution. When a route request occurs, two nodes are randomly selected as source and destination. The data traffic is generated by Constant Bit Rate (CBR) sessions initiated between the source and destination. Each clusterhead maintains threshold value (cut-off). The table 1 shows simulation parameters and their values.

During simulation, several performance parameters were estimated, like number of clusters (NC, number of border nodes, CH%, BP%, ASP, the total energy consumption at each clusterhead and total residual energy at each clusterhead. Along the number of clusterheads reached to the threshold, the results are shown in Table 2.

Table 1. Simulation parameters.

Sno	Parameter	Value
1	Simulation area and Network Size	1000mx1000m, 50 Nodes
2	Transmission Range	88 meters
3	Transmission Power Receiving Power	0.7 Joule/packet, 0.3 Joule/packet
4	Node Mobility Model, Pause Time and Speed	Random waypoint mobility model, 10 sec, 0 to 2m/s
5	Initial Energy, Maximum Battery Capacity	100 Joules, 100 Joules
6	Weight factor of clusterhead	1
7	Threshold Value	5 Joules
8	Route request arrival rate λ	5, 10 per 10 sec
9	Traffic type, Maximum Data Packet size	Constant Bit Rate (CBR), 512 bytes
10	Queue type and queue Size	Drop tail, 60*512 bytes
11	Total Simulation Time	100 sec

Table 2. Results.

Sno	Parameter	D=3.02	D=3.24	D=3.8
1	Number of Clusters	13	13	13
2	Elected Clusterheads	2 ,19, 6, 42, 40 33, 28, 26, 22 37, 15 ,12, 48	2, 18 ,5, 40, 32, 28, 24, 36, 50, 42, 14 ,46 ,10	2 ,19, 6, 42, 40, 32, 28, 24 36, 50, 17, 12, 46
3	Number of Border nodes	9	11	13
4	CH%	0.26	0.26	0.26
5	BP%	0.18	0.22	0.26
6	AS	3.846	3.846	3.846

The size and degree are denoted by N and D, respectively. CH% denotes the ratio of clusterhead nodes (that is, the number of clusters divided by the total number of nodes). Border nodes are nodes that belong to more than one cluster (that is, which are at distance at most k hops from at least two CHs), BP% denotes the ratio of border nodes, (that is, the number of border nodes divided by the total number of nodes). AS is the average size of a cluster (the average number of nodes in a cluster, that is, the number of nodes divided by the total number of clusters). Other results are shown in Figure 5 to Figure 7.

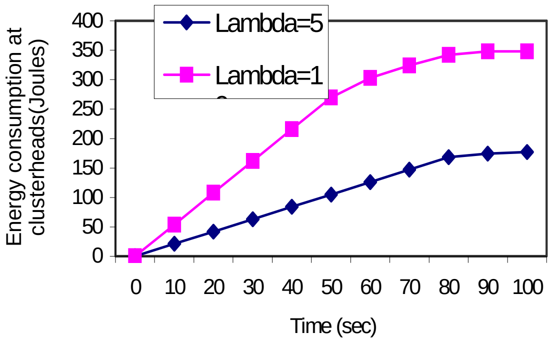


Figure 5. Energy consumption Vs time.

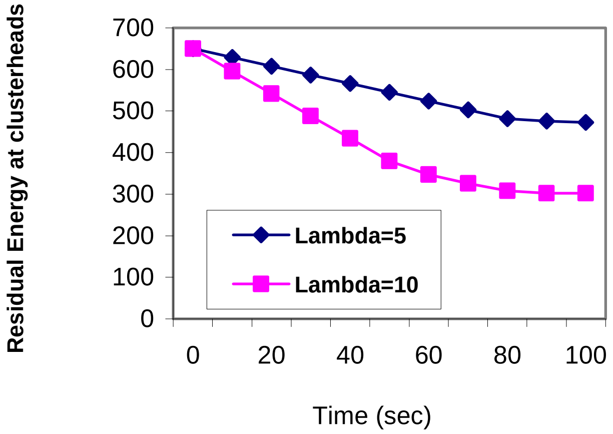


Figure 6. Residual energy Vs time.

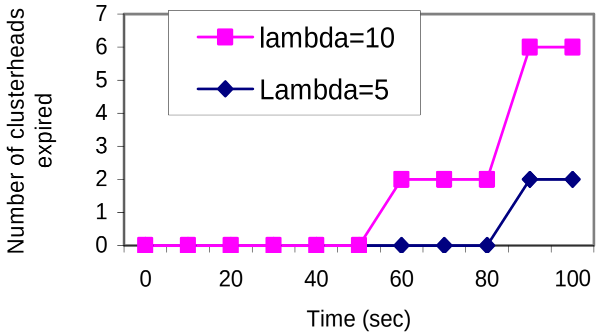


Figure 7. Number of CHs expired Vs time.

The total power consumption is directly proportional to various factors like network size, route requests arrival rate, packet arrival rate, packet size (header size and payload size), packet collision and retransmissions. Total residual energy is indirectly proportional to the power consumption. The network life depends on then node expiration which in turn depends upon energy consumption and threshold value. The node life time is indirectly proportional to the node's energy consumption and it is also directly proportional to the threshold value of the energy of each node is denoted by γ. It is also called cut-off value; it is always greater than one. During the transmission of data, each node checks whether its energy reaches to threshold or not. If its energy reaches to threshold then it will expire. The network lifetime can be defined in many ways:

- It may be defined as the time taken for K% of the nodes in a network to die
- It might be the time taken for the first node to die

- It can also be the time for all nodes in the network to die

It can also be the time for all nodes in the network to die. To maximize the lifetime of network, each clustered maintains threshold.

Figure 5 depicts the variation of total energy consumption of clusterheads versus time; Figure 6 depicts the variation of total residual energy of clusterheads versus time. Figure 7 depicts number of clusterheads expired versus time.

5. Conclusions

Proposed novel approach (EACMR) is very suitable and applicable for the cluster based mobile ad hoc networks. It is not based on GPS. It can be applied to a mobile ad hoc network that is using any clustering scheme. In this work, the CONID clustering scheme is used as background to form clusters, instead of that any clustering scheme may be used to form clusters in the network.

The EACMR is designed to find energy aware node-disjoint multiple routes from a source to a destination through clusterheads. It increases the network life time by using optimal routes, as compare to on demand multipath routing protocols, it significantly reduces the total number of route request packets using clustering technique, this result in an increased packet delivery ratio,decreasing end-to-end delays for the data packets, lower control overhead, fewer collisions of packets , decreasing power consumption. It supports the reliability by using multiple node-disjoint paths.

6. References

[1] C. S. R. Murthy and B. S. Manoj, "Ad hoc wireless networks: Architectures and protocols," Prentice Hall, 2004.

[2] J. Y. Yu and P. H. J. Chong, "A survey of clustering schemes for mobile ad hoc networks," in Proceedings of IEEE Communications Surveys, Vol. 7, pp. 32–48, 2005.

[3] Y. B. Liang, "Multipath 'fresnel zone' routing for wireless ad hoc networks," Virginia Polytechnic Institute and State University, March, 2004.

[4] J. Broch, D. Johnson, and D. Maltz, "The dynamic source routing protocol for mobile ad hoc networks," IETF Internet draft, 2004.

[5] E. P. Charles, E. M. Belding-Royer, and I. Chakeres, "Ad hoc on-demand distance vector routing," IETF Internet draft, 2003.

[6] K. M. Mahesh and R. D. Samir, "On-demand multipath distance vector routing in ad hoc networks," in Proceedings of IEEE International Conference on Network Protocols, pp. 14–23, 2001.

[7] S. J. Lee and M. Gerla, "Split multipath routing with maximally disjoint paths in ad hoc networks," in Proceedings of IEEE ICC, pp. 3201–3205, 2001.

[8] L. Wang, Y. Shu, Z. Zhao, L. Zhang, and O. Yang, "Load balancing of multipath source routing in ad hoc networks," in Proceedings of IEEE ICCC, Vol. 5, pp. 3197–3201, 2002.

[9] Z. Q. Ye, S. V. Krishnamurthy, and S. K. Tripathi, "A framework for reliable routing in mobile ad hoc networks," in Proceedings of IEEE INFOCOM, Vol. 1, pp. 270–280, 2003.

[10] P. Yuan, Y. Bai, and H. Wang, "A multipath energy-efficient routing protocol for ad hoc networks," in Proceedings of International Conference on Communications, Circuits and Systems, Vol. 3, pp. 1462–1466, 2006.

[11] L. S. Tan, L. Xie, K. T. Ko, M. Lei, and M. Zukerman, "LAMOR: Lifetime-aware multipath optimized routing algorithm for video transmission over ad hoc networks," in Proceedings of IEEE Vehicular Technology Conference, Vol. 2, pp. 623–627, 2006.

[12] S. Y. Jin, K. Kang, Y. J. Cho, and S. Y. Chae, "Power-aware multi-path routing protocol for wireless ad hoc network," in Proceedings of IEEE Wireless Communications and Networking Conference, pp. 2247–2252, 2008.

[13] P. S. Anand, A. J. Anto, V. Janani, and P. Narayanasamy, "Multipath power sensitive routing protocol for mobile ad hoc networks," in Proceedings of Wireless on Demand Network Systems, Springer, Vol. 2928, pp. 84–89, 2004.

[14] D. Y. Hwang, E. H. Kwon, and J. S. Lim, "An energy aware source routing with disjoint multipath selection for energy-efficient multihop wireless ad hoc networks," in Proceedings of International Federation for Information Processing, pp. 41–50, 2006.

[15] H. Y. An, L. Zhong, X. C. Lu, and W. Peng, "A cluster-based multipath dynamic source routing in MANET," in Proceedings of IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Vol. 3, pp. 369–376, August 2005.

[16] Z. Y. Wu, X. J. Dong, and L. Cui, "A grid-based energy aware node-disjoint multipath routing algorithm for MANETs," in Proceedings of International Conference on Natural Computation, Vol. 5, pp. 244–248, 2007.

[17] R. Bajaj, S. Rannweer, and D. P. Agrawal, "GPS: Location-tracking technology," IEEE computer, Vol. 35, pp. 92–94, April 2002.

[18] J. C. Cano and D. Kim, "Investigating performance of power-aware routing protocols for mobile ad hoc networks," in Proceedings of International Mobility and Wireless Access Workshop, pp. 80, 2002.

[19] F. G. Nocetti, J. S. Gonzalez, and I. Stojmenovic, "Connectivity based k-hop clustering in wireless networks," in Proceedings of Telecommunication Systems, Kluwer Academic Publishers, pp. 205–220, 2003.

[20] Simulation model for Maximum Battery Life Routing, http://sarwiki.informatik.hu-berlin.de.

***I. J. Communications, Network and System Sciences*, 2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

An Energy-Aware Clustering Approach for Wireless Sensor Networks

Peter K. K. Loh[1], Yi Pan[2]
[1]*School of Computer Engineering, Nanyang Technological University, Singapore*
[2]*Department of Computer Science, Georgia State University, USA*
Email: *askkloh@ntu.edu.sg, pan@cs.gsu.edu*
Received October 27, 2008; *revised February* 6, 2009; *accepted February* 11, 2009

Abstract

Energy conservation is an essential and critical requirement for a wireless sensor network with battery operated nodes intended for long term operations. Prior work has described different approaches to routing protocol designs that achieve energy efficiency in a wireless sensor network. Several of these works involve variations of mote-to-mote routing (flat routing) while some make use of leader nodes in clusters to perform routing (hierarchical routing). A key question then arises as to how the performance of an energy-aware, flat routing protocol compare with that of one based on hierarchical routing. This paper demonstrates a hierarchical routing protocol design that can conserve significant energy in its setup phase as well as during its steady state data dissemination phase. This paper describes the design of this protocol and evaluates its performance against existing energy-aware flat routing protocols. Simulation results show that it exhibits competitive performance against the flat routing protocols.

Keywords: Energy-Aware, Routing Protocol, Wireless Sensor Network, Energy Efficiency, Reliable Routing

1. Introduction

Wireless adhoc networks comprise of stationary or mobile devices that communicate over wireless channels without any fixed wired backbone infrastructure. A wireless sensor network (WSN) is a special class of adhoc networks that integrates sensing, processing and communications in small, battery-powered motes [1,2]. These motes (sensor nodes) typically collaborate on a global sensing task and deliver required data to one or more hubs. Sensor nodes that have variable-powered RF transceivers can provide greater routing performance at the cost of higher power consumption [3−5]. On the other hand, nodes that have fixed-power RF transceivers are generally cheaper but may be more prone to communication disruptions [6,7]. Despite advances in Micro- Electro Mechanical Systems (MEMS) technology, energy constraints continue to limit the operations lifetime of a WSN and new, energy-aware motes are still experimental [2,6,8,9]. Some WSNs adopt a hierarchical configuration during deployment [5,10]. The deployed network topology consists of distributed clusters of sensor nodes. Each of these clusters is managed by a cluster head (or leader sensor node) that is responsible for data aggregation within the cluster and communications between this cluster and neighbouring ones.

Within a WSN, whether clustered or non-clustered, the primary means of relaying data among nodes is via a routing protocol [5,10−19]. Hence, an essential and critical design requirement of the routing protocol is that it be energy-aware. An energy-aware routing protocol should exhibit energy efficiency and balanced energy consumption across the WSN. The first requirement ensures that the WSN can sustain operations over prolonged, unattended periods. The latter requirement ensures that sections of the WSN do not fail prematurely and disrupt operations. A routing protocol for WSNs typically comprises the three phases: set-up phase, route management phase and data dissemination phase. With clustered wireless sensor networks, the set-up phase may also incorporate the formation of clusters around each available cluster head.

Energy-aware routing protocols cannot merely deliver the message to a hub via the shortest or most energy-efficient route. Due to high usage, energy resources of nodes along these routes will be depleted faster than others and these routes will fail. Protocol design must also ensure that packet traffic is distributed relatively uniformly across the network so that energy resources of all nodes are depleted at a balanced rate. This will ensure that certain network sections/nodes will not be abruptly disconnected due to low energy resources [7,9]. These are by no means trivial requirements and pose conflicting demands on the design of energy-aware routing protocols [20]. While conventional routing protocols for wireless networks are typically concerned with throughput and network latency, energy-aware routing protocols in WSNs have to consider energy consumption, energy variance and scalability as well [2,10,12,14,21−23].

With these issues in mind, we propose an energy-aware routing protocol, Energy Clustering Protocol (ECP) that routes messages via cluster heads. Unlike other clustered configurations, ECP exploits nodes at the boundaries of the cluster (*border nodes*) to assist in the forwarding of packets as well as to reduce dependency on and energy expenditure of cluster heads. Via performance simulations against existing energy-efficient routing protocols that use energy-distance metrics, probabilistic distribution of packet traffic and MAC adaptations, we show that ECP exhibits very low energy variance as well as high energy efficiency over WSNs with increasing number of nodes. The remainder of this paper is organised as follows. Section 2 surveys three existing energy efficient routing protocols proposed for multi-hop WSNs. Section 3 highlights our motivation and the contribution of our work. Section 4 describes the detailed design of our proposed routing protocol. Simulation results are presented and discussed in Section 5. Finally, Section 6 concludes this paper followed by the references.

2. Related Work

In this section, we discuss three more recently proposed routing protocols for multi-hop WSNs. They are Energy Probabilistic Routing (EPR) [6], Gradient Based Routing (GBR) [24] and Efficient and Reliable Routing (EAR) [25]. These routing protocols are similar in the sense that they make use of neighbourhood information such as hop-count and node energy levels to relay data. They differ in their approach to distribute packet traffic.

2.1. EPR

EPR is a reactive protocol that is destination-driven. That is, the hub or sink node initiates the route request and subsequently maintains the route. EPR selects routes probabilistically based on residual energy and energy consumption, thus helping to spread energy use among all the nodes. The protocol has three phases: setup, data dissemination and route maintenance. In the setup phase, interest propagation occurs as localized flooding, in the direction of the source node, to find all routes from source to hub and their energy costs. Before sending the request, the hub sets a “Cost” field to zero. Every intermediate node forwards the request only to neighbouring nodes that are closer to the source node than itself and farther away from the hub. On receiving the request at a node, the energy cost for the neighbour that sent the request is computed and is added to the total cost of the path. Routing tables are generated during this phase. Only neighbouring nodes with paths of low cost are added to the routing table. Paths that have a very high cost are discarded. A probability is assigned to each of the neighbours in the routing table with the probability inversely proportional to the cost. In the data dissemination phase, data is relayed using information from the routing tables generated in the setup phase. Paths are chosen probabilistically according to the energy costs that were calculated earlier. This is continued till the data packet reaches the destination node. A node may therefore have multiple routes to the hub. In the route maintenance phase, localized flooding is performed intermittently from hub to source to keep all the paths alive.

2.2. GBR

The GBR protocol seeks to distribute the network traffic load evenly among all nodes to prevent overloading. The hub will broadcast an interest message that is propagated throughout the network. Each node upon receiving the interest message will record the number of hops taken by the interest message. This allows the node to know the number of hops it needs to reach the hub. The difference between the hop count of a node and that of its neighbour is the gradient of that link. Gradients are thus setup from the nodes to the hub and all messages will flow in that direction towards the hub. A node will forward a message to a neighbour with the greatest gradient. If this link is not available due to failure or disruption, the neighbour with the next highest gradient is chosen and so on. When there are multiple neighbours with links having the same gradient, one is randomly chosen. Random choice of the next hop node has a good effect of spreading traffic over time as well as achieving re-configuration to adapt to communication disruptions and distortions.

When a node detects that its energy level has dropped by 50% or more, it increases its hop count (lowering its gradient) to discourage other nodes from routing packets through it. This change in gradient is propagated as far as needed over the network to keep other gradients consistent.

2.3. EAR

In the EAR protocol, routing decisions are based on hop-count and a weighted combination metric. This second metric is a weighted combination of energy levels, distance traversed and transmission success history used to determine optimal routes during data dissemination. A "sliding-window" that keeps track of the last *N* successful transmissions via a specified RF link is used to compute transmission success history. An optimal route in EAR may not be the shortest but represents the best combination of distance, energy requirements and RF link performance. Control packets are minimized by "piggy-backing" route management information onto MAC-layer protocol packets. EAR deals well with communication disruptions and distortions in WSNs with low to moderate traffic volumes, mostly due to more-informed and therefore accurate routing decisions. However, in WSNs with a high-volume of network traffic, the proportionate increase in control packets incur an appreciable overhead affecting its performance.

3. Motivation and Contribution

The protocols in our survey use different approaches to achieve energy efficiency. A minimum cost spanning tree such as that used in EAR allows for a low total energy consumption but sacrifices node survivability by over-utilising nodes on optimal routes. Probabilistic routing over multiple alternative paths to the hub is a technique used by EPR to overcome the over-utilisation of nodes on shortest paths. Such multi-paths are built based on a weighted combination of neighbouring node distance, projected energy expenditure and node residual energy. Energy availability metrics such as those employed in GBR control routing through nodes with the highest residual energy to balance energy consumption over the network.

In all cases, the focus is primarily on balancing energy consumption during the data dissemination phase. Typically, the set-up phase in these protocols involves flooding that starts from the hub to relay location and route information throughout the WSN. This technique consumes a significant proportion of energy from all nodes in the WSN. Set-up tasks are important as they provide necessary network configuration and status information that allows subsequent successful operation of the WSN. In noisy environments and where the WSN nodes are mobile, initiating a secondary set-up phase is the most straightforward and practical way to re-configure the network and re-synchronize operations.

In our design, we adopt a node clustering approach to utilise the gains of data fusion in tandem with energy conservation. Our proposed protocol, ECP, is designed with the following advantages: energy-efficient set-up process, low and balanced energy consumption during data dissemination by utilizing cluster boundary nodes instead of solely cluster heads, and scalable performance.

4. Energy Clustering Protocol (ECP) Design

4.1. Overview

In this section, the design and operation of ECP is presented. ECP is a routing protocol that minimizes route setup energy whilst maintaining low data dissemination energy consumption. A low energy route setup cost is important in applications where the network configuration may change dynamically due to inconsistent RF links, node mobility or simply when the nodes have fallen in residual energy after a period of time. In these applications, a low route setup cost is valuable to establish new routes again that reflect the new network topology in terms of residual energy and connectivity.

Unlike LEACH [19], ECP does not assume that all network nodes are able to reach the hub directly. Nodes route data packets to cluster heads. Each cluster head then routes to its border nodes and these in turn route to border nodes and cluster head of a neighbouring cluster. In this way, data is relayed from cluster to cluster and eventually to the hub. ECP is thus able to cater to larger network deployments where motes may be scattered over significant distances. Also, the use of border nodes as routing support balances the energy consumption per cluster and obviates requirements for differentiated high transmission power cluster heads.

Another novel feature of ECP is that cluster heads are elected probabilistically. ECP elects one hop clusters in a 3-round process, each round with increasing probabilities to form its clusters. This clustering process strives to increase the number of border nodes between clusters for the conduct of inter-cluster routing. Instead of using multiple hop clusters in a multi-hierarchical setting, ECP forms one-hop clusters in a single level hierarchy in the WSN. The advantages of one hop cluster are detailed in Subsection 4.2. In ECP, nodes already in a cluster may join another cluster if through distance estimation they are detected to be nearer to the other cluster head. Thus, more energy can be conserved through this simple scheme without additional control messages. Energy is also minimised during routing since nodes are clustered based on their distance from the cluster head.

ECP does not need location information of its nodes through localisation or GPS techniques as in the BCDCP protocol [26]. The use of localisation or GPS techniques consumes additional energy in order to initialise the nodes with their geographical coordinates. ECP also does not require all nodes to send their information to the hub first for some centralized processing. Sending information to the hub for centralized processing is a common technique which allows for some useful network-wide information regarding residual energy, average energy levels or geographical location of nodes to be mapped out. Routing patterns can then be fixed through this network mapping. This method, however, is not distributed and does not scale very well with increasing network size. ECP achieves clustering and routing in a distributed manner and thus provides good scalability. The operation of ECP is divided into 3 phases: clustering, route management, and data dissemination. The following sub-sections will detail the design of these phases.

4.2. Clustering

Phase one of ECP is to cluster sensor nodes together to achieve a maximum number of border nodes and minimum number of clusters. To prevent the same border nodes from being used continuously, the clustering algorithm aims to achieve denser clusters. One-hop clustering is adopted for ECP because these clusters have been shown to be more robust and subjected to less connectivity problems and communication overheads [27]. When a node is first powered on, it will decide if it will elect itself to be a CH. The probabilities used for the 3-round clustering are 0.1, 0.4 and 1 for the first, second and third rounds of clustering, respectively. These values are determined empirically. The flowchart in Figure 1 shows how cluster formation is done.

CHs once elected wait for a random amount of time before broadcasting a PROBE message to its neighbours. The node is confirmed as a CH after the PROBE message is sent. This PROBE message announces the status of the newly formed CH to surrounding nodes. Nodes already elected to be CHs but have not yet sent the PROBE will give up their status to become cluster members. Nodes without cluster status (not cluster head or cluster members) will join the cluster as members via the PROBE message. The selection of a random range of time to wait before broadcasting a PROBE message is dependent on the density of the cluster and the maximum time for a message to be sent from one hop to the next.

Let the cluster density *d* be defined as the number of nodes which a particular sensor node is able to reach within its transmission range. Let *t* be the maximum time taken for a PROBE or REPLY message to traverse one hop. This is also the minimum time that an elected cluster head node has to wait before broadcasting the PROBE message.

Assuming a worst case scenario where all the (d+1) nodes in a potential cluster (1 potential cluster head node surrounded by *d* neighbours) may be elected as cluster heads. To prevent this scenario, the minimum range of waiting time allocated to the nodes has to be at least (d+1) *t*. That is, a node that has elected itself as cluster head waits at least *t* units or as long as (d+2)t units as shown in the right diagram of Figure 1.

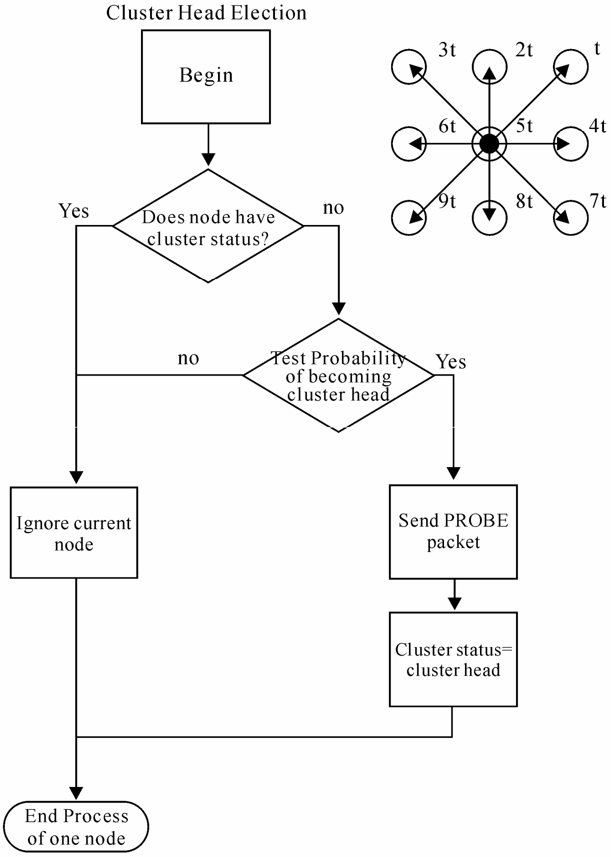


Figure 1. Cluster Formation with ECP.

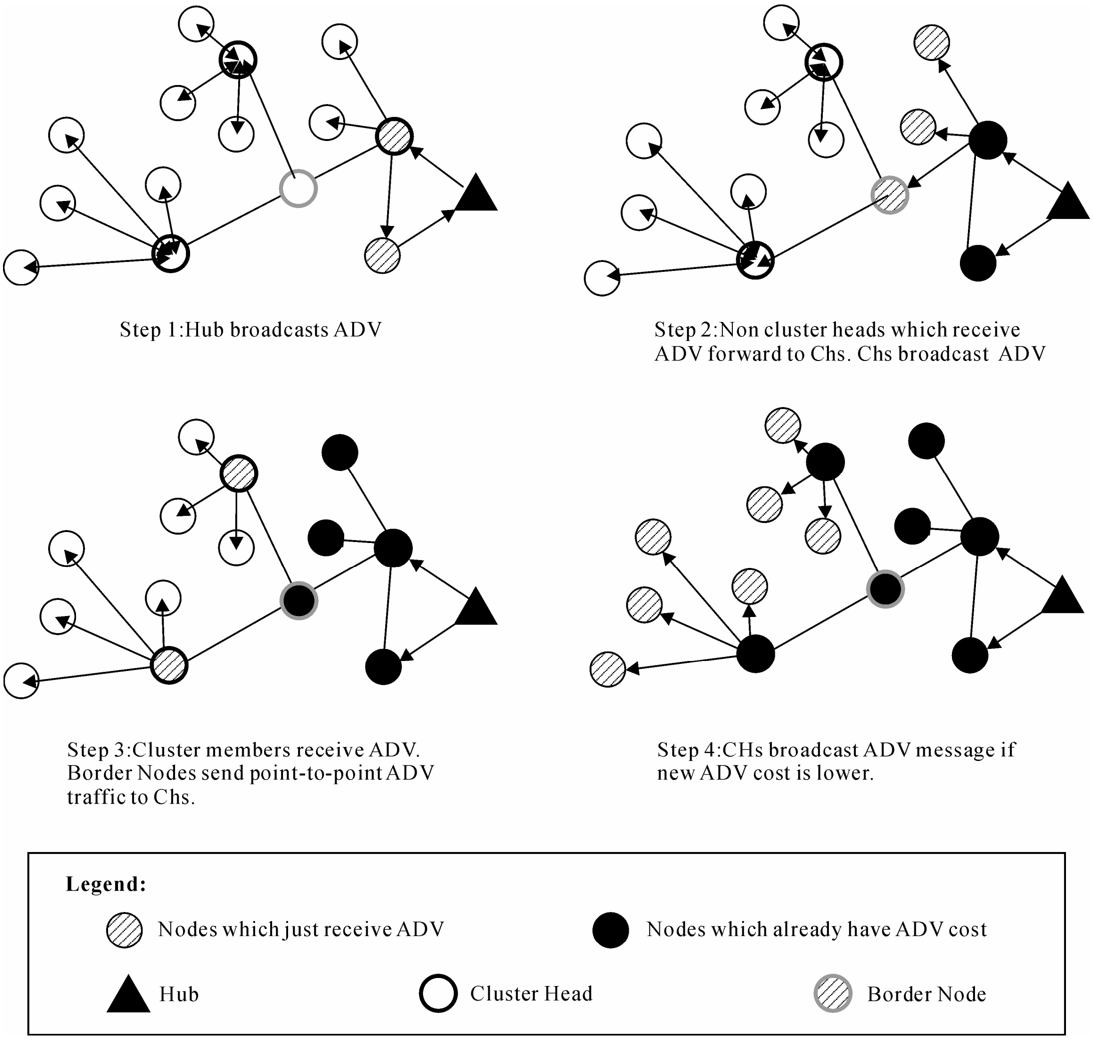
After a node elects itself as a potential cluster head and broadcasts the PROBE message, only one CH is formed and the rest of the elected CHs give up their CH candidacy to be cluster members. Upon reception of a PROBE message, nodes without a CH will reply with a REPLY message and store the address of the CH. Nodes which have already joined a CH will also reply with a REPLY message if the PROBE message is from another CH. These nodes then compare the distance between the original CH and the PROBE message from the new CH and join the CH that is nearer. Its previous CH will be regarded as a secondary CH. This ensures that no CHs will be deprived of cluster members or some CHs will have too many cluster members. All CHs thus keep a record of their cluster members using the REPLY messages.

4.3. Route Management Phase

The route management phase comprises of route propagation and route request. Route propagation avoids conventional flooding to discover an unknown network. It achieves this by using the clusters from the previous phase to forward the route messages. ECP forms a minimum energy-cost spanning tree of CHs instead of all nodes. CHs, upon receiving a routing cost, will be able to update their cluster members of the route cost by intra-cluster broadcasting. The non-border node cluster members thus play a passive role in route dissemination. This helps to save transmission cost as not all nodes will have to participate in forwarding the route messages. A route request round follows after route propagation. This is necessary because of the possible presence of isolated clusters. An isolated cluster that does not have any border nodes with another cluster misses out on route information. This phase of ECP discovers these nodes without a routing table and requests for a route.

4.3.1. Route Propagation

Route propagation begins when the hub first broadcasts an advertisement (ADV) packet with its address to its neighbours. All nodes start with an original cost to the hub of infinity. The ADV packet from the hub starts with a cost of 0. Nodes that receive the ADV packet add the cost of the ADV and the cost of transmitting from sender to receiver. If this cost is smaller than the receiver's original cost, it will add the sender's information into its route table. Otherwise, the ADV packet is ignored. Where the sender is the hub, the node will add the route to the hub and forwards future data packets direct to the hub and not to its CH. Nodes receiving the ADV packet from a non-hub node will send it to their CHs. In the case of border nodes where they have more than one cluster head (one primary CH where it sends its data packet to and secondary CHs for routing purposes), the border node sends point-to-point traffic to all its primary and secondary CHs. The CHs upon receiving the ADV will check that the cost of this ADV is lower than its original cost. If the ADV packet is of a lower cost, it will broadcast this route cost to its cluster member nodes. Border nodes upon receiving this new routing cost will thus be able to resend it to the other CHs. Figure 2 shows this.


Step 1:Hub broadcasts ADV

Step 2:Non cluster heads which receive ADV forward to Chs. Chs broadcast ADV

Step 3:Cluster members receive ADV. Border Nodes send point-to-point ADV traffic to Chs.

Step 4:CHs broadcast ADV message if new ADV cost is lower.

Legend:
▨ Nodes which just receive ADV
● Nodes which already have ADV cost
▲ Hub
○ Cluster Head
▨ Border Node

Figure 2. Route Propagation with ECP.

4.3.2. Route Request (Route propagation)

Route request starts after route propagation of ADV packets. Here, nodes that do not have a route after a timeout period potentially belong to an isolated cluster that does not have a border node with other clusters. These nodes then broadcast Route Request messages (RREQ) in an attempt to discover a neighbouring cluster that has a route to the hub. When nodes receive an RREQ packet, they will reply with an ADV packet. The sender of the RREQ packet receives this ADV and propagates the route information to its own CH. The CH then continues with the normal route propagation. CHs and non-CHs may broadcast RREQ packets and the lowest route cost ADV packet is kept. Transmission of data packets start immediately after a route to hub is received. If a lower cost ADV should arrive to the node, dynamic updating of the lower cost route is done concurrently with application sensing. The lower cost route will thus be the new route used.

4.3.3. Energy Metric

The energy metric that is used can include information about the cost of using the path, energy status of the nodes along the path or reliability of the RF links etc. ECP evaluates routes by evaluating the energy used to transmit and receive on a link. For motes with variable transmission power, the energy metric would be a function of the distance between the sender and receiver. For motes with fixed transmission power, the energy to transmit and receive on a link is the same for all nodes. As channel acquisition overhead is large, small control packets have disproportionately high energy costs. ECP minimizes the use of control packets to propagate the route information to all the nodes by propagating the route information to its CHs instead.

4.3.4. Energy Model

In the evaluation of the protocols in this study, the energy model in [19,28] is used. The energy costs of broadcast, point-to-point and non-destination traffic are different. We denote the energy lost due to channel transmission as r^2, where r is the distance between the sending and receiving nodes. Therefore, the energy expended to transmit a k-bit packet over a distance d and to receive that packet defined by:

$$\text{ETx (k, d)} = \text{Eelec} * k + \text{Eamp} * k * d^2 + b \qquad \textbf{(1)}$$

$$\text{ERx (k)} = \text{Eelec} * k + b \qquad \textbf{(2)}$$

where, ETx = Energy taken to transmit the packet
ERx = Energy taken to receive the packet
Eelec = Energy dissipation of radio transceiver circuitry
Eamp = Energy to run transmit amplifier

For simplification, we consider the radio channel to be symmetrical. The above energy model where a node consumes energy through transmitting/receiving packets may be described as a linear equation [29]. To account for energy consumption at the data link layer through device mode changes and channel acquisition cost, a fixed cost b is included that depends on the operation mode:

Broadcast traffic: In an IEEE 802.11 Broadcast, the sender listens briefly to the channel and sends the messages if the channel is clear. We define the fixed channel access cost as $b_{\text{b-send}}$ and $b_{\text{b-recv}}$:

$\text{ETx (k, d)} = \text{Eelec} * k + \text{Eamp} * k * d^2 + b_{\text{b-send}}$

$\text{ERx (k)} = \text{Eelec} * k + b_{\text{b-recv}}$

Point-to-point traffic: In IEEE 802.11, when a node sends an RTS control message identifying the receiver node, the latter responds with a CTS. Upon receiving the CTS, the the data is sent and the sender waits for an ACK from the receiver. This handshake overhead is accounted by the fixed channel access cost for sending/receiving a packet as $b$pp-send and $b$pp-recv respectively:

$\text{ETx (k, d)} = \text{Eelec} * k + \text{Eamp} * k * d2 + b_{\text{pp-send}}$

$\text{ERx (k)} = \text{Eelec} * k + b_{\text{pp-recv}}$

Non-destination traffic: Non-destination nodes in the range of either the sender or receiver overhear some or all of the packet traffic. Non-destination nodes in non-promiscuous mode can enter into a reduced energy consumption mode while data is being transmitted in the vicinity. For non-promiscuous nodes discarding traffic, Equation (2) becomes:

$$\text{Discard cost} = \text{Eelec} * k + b_{\text{discard}} \qquad \textbf{(3)}$$

Experimental values for all the b parameter in the 3 modes of operation are listed as follows:

Operation	$b_{\text{b-send}}$	$b_{\text{b-recv}}$	$b_{\text{pp-send}}$	$b_{\text{pp-recv}}$	b_{discard}
***b*(uJ)**	266	56	454	356	24

4.4. Data Dissemination

The clustering phase, route propagation and route request processes ensure that every node has a route to the hub via its own CH. Depending on the application, nodes will start generating DATA packets at periodic intervals or cluster member nodes may go into sleep mode if they are not needed. If a node has a direct route to the hub, the DATA packet will be sent direct to it. Cluster member nodes that are not border nodes will send the DATA packet to its CH for data fusion. Besides cluster members, CHs also keep a record of member nodes that are border nodes and their corresponding costs to the hub. The CHs will select the border node with the least cost to the hub and route the DATA packet to the border node. Border nodes upon receiving a DATA packet will send the packet to its record of CHs with lower energy cost than itself. Hence, DATA packets are routed from cluster to cluster till it reaches the hub. Data aggregation is also used by ECP. When DATA packets meet along the same path at a CH, the data is aggregated before transmission. Through this inter-cluster routing approach using CHs and border nodes together with data fusion and aggregation methods, the network is effectively condensed into a shortest spanning tree of CHs and their border nodes. Non-border node cluster members thus do not participate in the routing decisions. The total number of transmissions is thus reduced leading to higher energy savings and lower variance across the network.

5. Simulation

GloMoSim [30] is a discrete-event simulator designed for wireless networks. It is made up of library modules, each of which simulates a specific routing protocol in the protocol stack. Simulator settings used are shown in Table 1.

5.1. Performance Metrics

The following metrics were used to measure the performance of the routing protocols.

Packet delivery ratio (PDR): this measures the percentage of data packets generated by the nodes that are successfully routed to the hubs. It is expressed as:

$$\frac{\textit{Total number of data packets successfully delivered}}{\textit{Total number of data packets sent}} \times 100\%$$

Packet latency: this measures the average time it takes to route a data packet from the source node to the hub. It is expressed as:

$$\frac{\sum \textit{Individual data packet latency}}{\textit{Total number of data packets delivered}}$$

Energy Consumption: this measures the energy expended per delivered data packet. It is expressed as:

$$\frac{\textit{Total energy } \exp \textit{ended}}{\textit{Total number of data packets delivered}}$$

Table 1. Simulator settings.

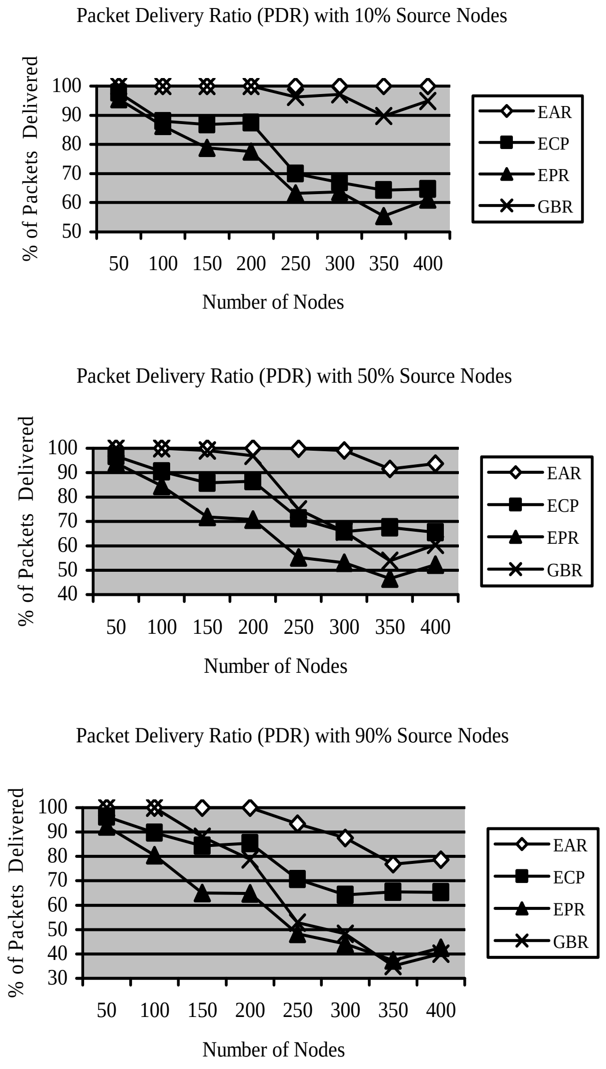
Frequency	433 MHz
Bandwidth	76800 Kbps
Radio Range	56 m
Radio Model	Signal-to-Noise (SNR) Bounded
Propagation Model	Ground Reflection (Two-Ray)
MAC Protocol	IEEE 802.11 (DCF)
Data Packet Size	24 bytes
Simulation Duration	60 minutes
Initial Node Energy	20 Joules
Total Packets per Node	120

Energy Variance: this measures the energy distribution of the network. It is expressed as:

$$\frac{\sum(\textit{Energy Consumption per Node} - \textit{Mean Energy Consumption})^2}{\textit{Total Number of Nodes} - 1}$$

5.2. Noisy Environment Tests

These tests analyze the protocols' behavioural differences in an actual operating environment. The test is conducted by generating random noise factors of between 10% and 50%. The noise factor of a node indicates the probability that packets received by the node are corrupted or lost in transmission. Results were averaged over 30 runs each with a different seed and presented in Figures 3 to 6.


Packet Delivery Ratio (PDR) with 10% Source Nodes

Packet Delivery Ratio (PDR) with 50% Source Nodes

Packet Delivery Ratio (PDR) with 90% Source Nodes

Figure 3. PDR results in noisy environment (1 Hub).

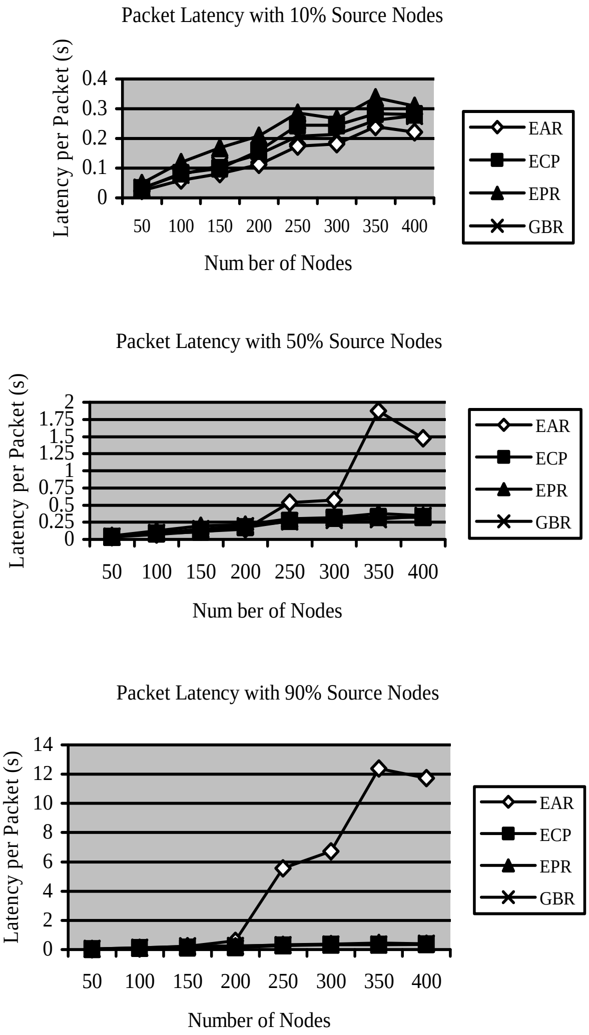


Figure 4. Latency results in noisy environment (1 Hub).

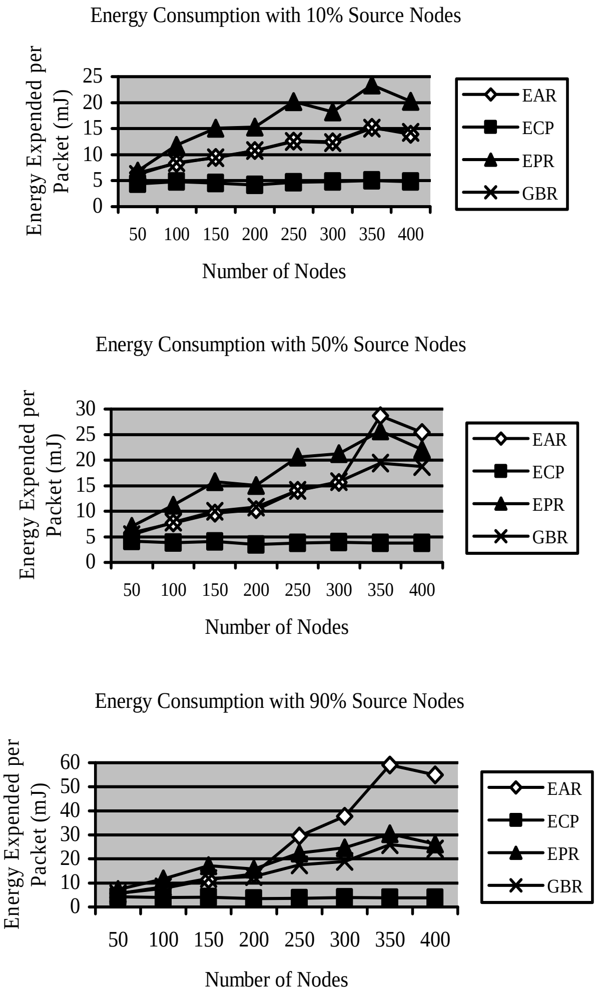


Figure 5. Energy consumption results in noisy environment (1 Hub).

5.2.1. Packet Delivery Ratio and Latency

When more active sources are included, probability of packet collisions increases and PDR of all protocols fall. ECP is able to maintain about the same throughput as packets need only be sent to CHs. The reduced transmissions help lessen the impact of noise. PDR of GBR dips the most with more source nodes because it routes via shortest paths to the hub but lacks a robust delivery mechanism. EPR performs better than GBR as it is a probabilistic protocol and does not consistently use the same routes. Packet latency of EPR is highest at 10% source nodes as it routes according to the residual energy remaining in the nodes and does not consider how long the route may take. ECP suffers from some latency overhead due to routing to the CH first before routing to the hub. GBR and EAR suffers the least latency at 10% source nodes due to shortest path routing to the hub. At 50% and 90% source nodes, EAR showed the highest latency from 200 nodes onwards. Route blacklisting occurs and data packets are routed through other less noisy RF links leading to more hops. The random back off mechanism for re-transmissions also contributes to the high latency.

5.2.2. Energy Consumption and Variance

The reason for EPR's higher energy consumption is due to the use of multi-path routing. Each node makes a localized decision to route the data packet based on probability. The node with the highest residual energy has the highest probability to be used as an intermediate node for routing to the hub. The protocol does not take a shortest path to the hub but instead aims to minimise the energy variance over the network. The energy consumption of EAR and GBR are about the same. Both protocols use optimal paths to route to the hub. The energy consumption of ECP is the lowest among the four protocols. Although additional energy is expended by routing to the CH first, the energy

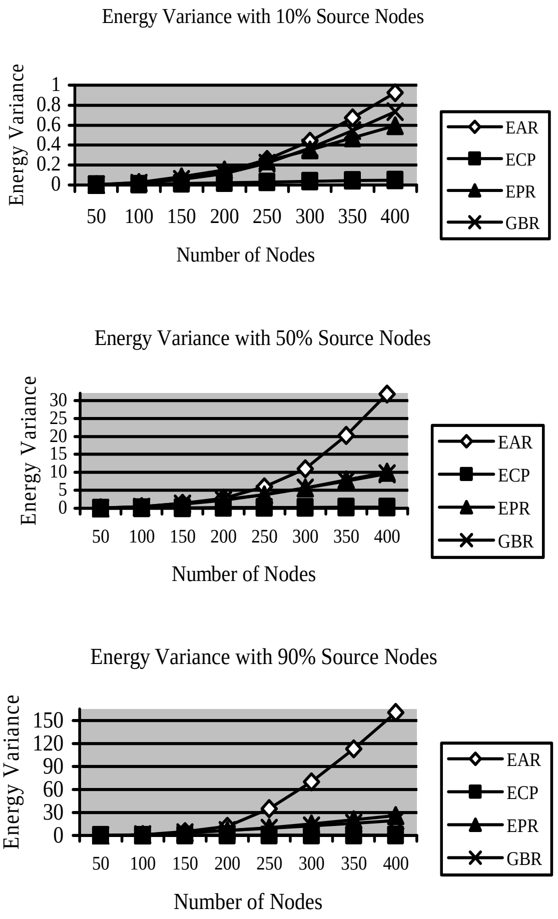


Figure 6. Energy variance results in noisy environment (1 Hub).

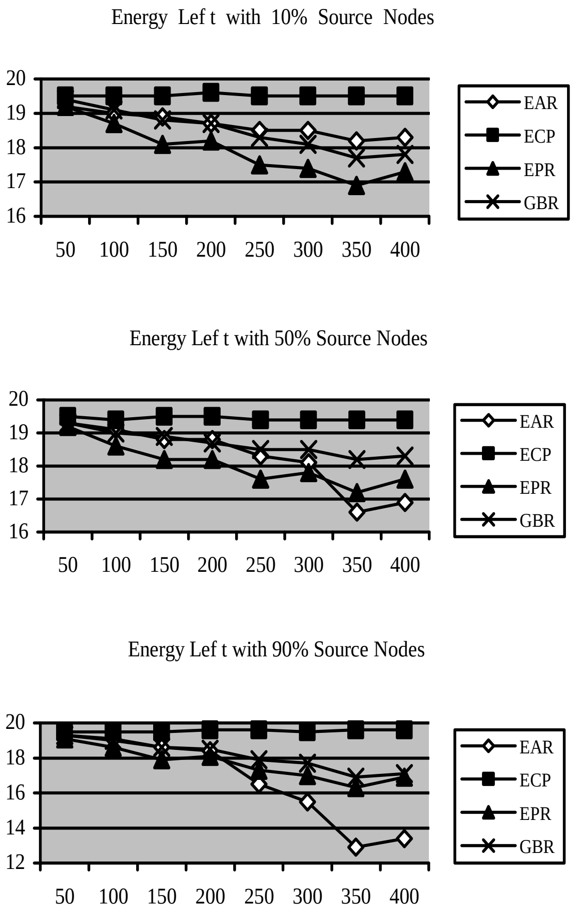


Figure 7. Average node energy left in noisy environment (1 Hub)

spent is less than the cumulative energy savings of data fusion. As the number of nodes increases, the clusters become denser. The effects of increased data packets from the higher number of nodes are mitigated by data fusion. The energy variance of EAR, EPR and GBR are shown to increase with network size. As more packets are generated, optimal paths are used continuously resulting in a higher energy variance in the network. For ECP, although more packets are generated, these packets are aggregated at the CHs thus leading to a reduced energy variance in the network.

5.2.3. Energy Remaining

Figure 7 shows the average node energy remaining after data packet transmissions have ceased in different sized networks for 10%, 50% and 90% active source nodes. In all 3 scenarios, ECP shows the highest average node energy remaining due to the use of cluster heads for packet routing. The trend is also relatively stable with increasing network size showing uniform route distribution of packets over the network. For 10% and 50% active sources, results exhibited by EAR and GBR are similar up to network sizes of 300. Beyond that, as well as with 90% active sources, the sole use of optimal paths by EAR become apparent as the node energies along these paths diminish significantly with respect to other network regions. For 10% and 50% active sources, EPR's average node energies remaining are the lowest as it uses multipath routing. However, at 90% active sources as well as with larger networks beyond 300 nodes, the multi-path routing lowers the energy variance over the network as a whole compared to protocols like EAR that use solely optimal paths.

Overall, the performance of ECP is mediocre at low network traffic levels with 10% active sources. ECP's performance only exceeds the other protocols when network traffic levels rise with 50% and 90% active sources. With these latter scenarios, the use of cluster heads to route packets to the hub minimizes the communication overheads (esp. control packets) and therefore packet losses due to collisions.

6. Conclusions

This research has demonstrated that ECP is a viable energy conserving protocol which balances energy consumption over the network. Simulation results have shown that the performance of ECP is scalable for networks as large as 400 nodes. ECP has made use of a clustering approach to reduce the number of packets sent through the network significantly, thus reducing communications costs across the network. The 3-round one-hop clustering technique of ECP lets nodes join the nearest CH without incurring excessive energy control costs leading to an energy-efficient setup. The route setup cost of ECP is shown to be lower than existing protocols, allowing new clusters to be formed inexpensively once the nodes fall in residual energy. Without assuming geographical knowledge of nodes, ECP is able route data packets reliably to the hub. Inter-cluster routing has also good scalability and is shown to be a more energy- efficient method of propagating route information to a large number of nodes compared to non-clustered WSNs. Energy efficiency of ECP outperforms the protocols of EAR, EPR and GBR without compromising packet delivery ratio, latency and energy variance. Future work could include research into improving performance in light to moderate traffic scenarios with multi-hubs, an intra-cluster protocol, reducing inter-cluster interference and the election of uniformly distributed cluster heads.

7. References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," Computer Networks, pp. 393−422, March 2002.

[2] N. S. Correal and N. Patwari, "Wireless sensor networks: Challenges and opportunities," in Proceedings of the 2001 Virginia Tech Symposium on Wireless Personal Communications, pp. 1−9, June 2001.

[3] http://www.ctr.kcl.ac.uk/iwwan2005/papers/57_not_attended.pdf.

[4] A. Manjeshwar and D. P. Agrawal, "APTEEN: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks," Proceedings of International Parallel and Distributed Processing Symposium (IPDPS'02), pp. 195−202, 2002.

[5] J. Kamimura, N. Wakamiya, and M. Murata, "Energy-efficient clustering method for data gathering in sensor networks," Proceedings of First Annual International Conference on Broadband Networks 2004, pp. 1−10, 2004.

[6] R. C. Shah and J. M. Rabaey, "Energy aware routing for low energy adhoc sensor networks," Proceedings of IEEE Wireless Communications and Network Conference, Vol. 1, pp. 350−355, March 2002.

[7] J. Chen, Y. Guan, and U. Pooch, "Customizing a geographical routing protocol for wireless sensor networks," Proceedings of International Conference on IT: Coding and Computing (ITCC,'05), pp. 586−591, 2005.

[8] R. Kannan, R. Kalidindi, S. S. Iyengar, and L. Ray, "Max-min length-energy-constrained routing in wireless sensor networks," LNCS-Lecture Notes in Computer Science, Springer-Verlag, Vol. 292, (from 1st European Workshop on Wireless Sensor Networks EWSN'2004), pp. 234−249, January 2004.

[9] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," Proceedings of ACM/IEEE International Conference on Mobile Computing and Networking, Boston, MA, USA, pp. 56–67, ACM, August 2000.

[10] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," IEEE Infocom '03, pp. 1713−1723, 2003.

[11] Q. Li, J. Aslam, and D. Rus, "Online power-aware routing in wireless ad hoc networks," IEEE/ACM International Conference on Mobile Computing and Networking (MobiCom 2001), Rome, Italy, pp. 97−107, July 2001.

[12] J. H. Chang and L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," INFOCOM, pp. 22−31, 2000.

[13] T. H. Lin, Y. S. Chen, and S. L. Lee, "PCAR: A power aware chessboard-based adaptive routing protocol for wireless sensor networks," IEEE 6th CAS Symposium on Emerging Technologies, pp. 145−148, 2004.

[14] M. Perillo and W. Heinzelman, "Dapr: A protocol for wireless sensor networks utilizing an application-based routing cost," IEEE Wireless Communications and Networking Conference (WCNC), pp. 1540−1545, 2004.

[15] M. A Youssef, M. F. Younis, and K. A. Arisha, "A constrained shortest-path energy-aware routing algorithm for wireless sensor networks," WCNC 2002-IEEE Wireless Communications and Networking Conference, No. 1, pp. 682−687, March 2002.

[16] A. Manjeshwar and D. Agrawal, "TEEN: A routing protocol for enhanced efficiency in wireless sensor networks," in Proceedings of the 15th International Parallel & Distributed Processing Symposium, IEEE Computer Society, pp. 189, 2001.

[17] http://www.cs.berkeley.edu/˜awoo/smartdust/.

[18] S. Nikoletseas, I. Chatzigiannakis, A. Antoniou, and G. Mylonas, "Energy efficient protocols for sensing multiple events in smart dust networks," Proceedings of 37th Annual Simulation Symposium, pp. 15, 2004.

[19] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks," Proceedings 33rd Hawaii International Conference on System Sciences, pp. 3005−3014, 2000.

[20] http://externe.inrs-emt.uquebec.ca/users/nuevo/glomoman.pdf.

[21] S Madiraju, C Mallanda, R Kanna, A Durresi, S. S. Iyengar, "EBRP: Energy band based routing protocol for wireless sensor networks," ISSNIP 2004, pp. 67−72.

[22] S. B. Wu and K. S. Candan, "GPER: Geographic power efficient routing in sensor networks," icnp, pp. 161−172,

12th IEEE International Conference on Network Protocols (ICNP'04), 2004.

[23] L. Li and J. Y. Halpern," Minimum energy mobile wireless networks revisited," ICC 2001, IEEE International Conference, Vol. 1, pp. 278−283, June 11−14, 2001.

[24] C. Schurgers and M. B. Srivastava, "Energy efficient routing in wireless sensor networks," Proceedings on Communications for Network-Centric Operations: Creating the Information Force, McLean, VA, 2001.

[25] K. K. Loh, W. J. Hsu, and Yi Pan, "Performance evaluation of efficient and reliable routing protocols for fixed-power sensor networks," in IEEE Transactions on Wireless Communications, 2009.

[26] S. D. Muruganathan, D. C. F. Ma, R. I. Bhasin, and A. O. Fapojuwo, "A centralized energy efficient routing protocol for wireless sensor networks," IEEE radio communications, pp. S8−S13, March 2005.

[27] G. Chen, F. Nocetti, J. Gonzalez, and I. Stojmenovic, "Connectivity based k-hop clustering in wireless networks," Proceedings of 35th Annual Hawaii International Conf on System Sciences (HICSS'02), Vol. 7, 2002.

[28] H. O. Tan and I. Korpeoglu, "Power efficient data gathering and aggregation in wireless sensor networks," SIGMOD Record, Vol. 32, No. 4, December 2003.

[29] L. M. Feeney and M. Nilsson, "Investigating the energy consumption of a wireless network interface in an ad hoc networking environment," IEEE Infocom 2001, pp. 1548−1557, 2001.

[30] http://pcl.cs.ucla.edu/projects/glomosim/.

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Hierarchical Hypercube Based Pairwise Key Establishment Scheme for Sensor Networks

Lei WANG[1,2]
[1]*College of Software, Hunan University, Changsha, Hunan, China*
[2]*Department of Computer Science, Lakehead University, Thunder Bay, Canada*
Email: *wanglei@hnu.cn*

Received November 27, 2008; *revised January* 17, 2009; *accepted February* 12, 2009

Abstract

Security schemes of pairwise key establishment, which enable sensors to communicate with each other securely, play a fundamental role in research on security issue in wireless sensor networks. A general framework for key predistribution is presented, based on the idea of KDC (Key Distribution Center) and polynomial pool schemes. By utilizing nice properties of *H*2 (Hierarchical Hypercube) model, a new security mechanism for key predistribution based on such model is also proposed. Furthermore, the working performance of tolerance resistance is seriously inspected in this paper. Theoretic analysis and experimental figures show that the algorithm addressed in this paper has better performance and provides higher possibilities for sensor to establish pairwise key, compared with previous related works.

Keywords: Pairwise Key, Sensor Networks, Key Pool, Key Predistribution, H2 Framework

1. Introduction

The security issue in wireless sensor networks has become research focus because of their tremendous application available in military as well as civilian areas. However, constrained conditions existent in such networks, such as hardware resources and energy consumption, have made security research more challenging compared with that in traditional networks.

Current research focus on such security schemes as authentication and key management issues, which are essential to provide basic secure service on sensor communications. Pairwise key establishment enables any two sensors to communicate secretly with each other. However, due to the characteristics of sensor nodes, it is not feasible to utilize traditional pairwise key establishment schemes.

Eeschnaure *et al*. [1] presented a probablitic key predistribution scheme for pairwise key establishment. This scheme picks a random pool (set) of keys *S* out of the total possible key space. For each node, *m* keys are randomly selected from the key pool *S* and stored into the node's memory so that any two sensors have a certain probability of sharing at least one common key. Chan [2] presented two key predistribution techniques: q-composite key predistribution and random pairwise keys scheme. The q-composite scheme extended the performance provided by [1], which requires at least *q* predistributed keys any two sensor should share. The random scheme randomly pickes pair of sensors and assigns each pair a unique random keys. Liu *et al.* [3] developed the idea addressed in previous works and proposed a general framework of polynomial pool-based key predistribution. Based on such a framework, they presented random subset assignment and hypercube-based assignment for key predistribution.

However, it still requires further research on key predistribution because of deficiencies existent in those previous works. Since sensor networks may have dramatic varieties of network scale, the *q-composite* scheme would fail to secure communications as a small number of nodes are compromised. The random scheme may requires each sensor to store a large number of keys, which would be contradicted with hardware constraints of sensor nodes. The random subset assignment would not ensure any two nodes to establish a key path if they do not share a common key. Though the hypercube-based assignment can make sure that there actually exist a key path, however, the possibilities of direct pairwise key establishment are not perfect, leading to large communication overhead.

In order to improve possibilities of direct pairwise key establishment, and depress communication overhead on indirect key establishment, we propose a *H*2 (Hierarchical Hypercube) framework, combined with a new key predistribution scheme. Moreover two new fault tolerance model and corresponding indirect pairwise key establishment schemes are also proposed, by applying nice properties on tolerance resistance *H*2 model has provided. The schemes has better working performance on probabilities of pairwise key establishment between any two sensors.

2. Preliminaries

2.1. Notations and Definitions

Definition1(key predistribution): Cryptographic algorithms are pre-loaded in sensors before node deployment phase.

Definition2 (pairwise key): When any two nods share a common key denoted as *E*, we call that the two nodes share a pairwise key *E*.

Definition 3 (key path): Given two nodes A_0 and A_k, which do not share a pairwise key, if there exists a path in sequence described as A_0, A_1, A_2,……, A_{k-1}, A_k and any two nodes A_i, A_j ($0 \leqslant i \leqslant k-1, 1 \leqslant j \leqslant k$) share at least one pairwise key, we call that path as a key path.

Definition 4 (*n*-dimensional hypercube interconnection network): *n*-dimensional hypercube interconnection network H_n (abbreviation as *n-cube*) is a kind of network topology that has the following characteristics: 1) It is consisted with 2^n nodes and $n \cdot 2^{n-1}$ links; 2) Each node can be coded with a different binary string with *n* bits such as $b_1b_2\ldots b_n$; 3) For any pair of nodes, there is a link between them if there is just one bit different between their corresponding binary strings.

Figure 1 illustrates the topology of a 4-dimensional hypercube interconnection network, which is consisted with 2^4=16 nodes and $4 \cdot 2^{4-1}$=32 links. And the nodes are coded from 0000 to 1111.

2.2. Related Works [1−3]

2.2.1. Polynomial-based Key Predistribution

In the scheme of polynomial-based key predistribution, the key setup server randomly generates a *t*-degree bivariate polynomial $f(x,y)=\sum_{i,j=0}^{t} a_{ij}x^iy^j$ over a finite field F_q, Notes that *q* is fairly large prime number and for any variables *x* and *y*, $f(x,y)= f(y,x)$ is always held. Then the key server computes a share of $f(x,y)$, denoted as $f(i,y)$ for each node, where *i* is assumed to be a unique ID for any sensor node. Every node is pre-loaded with its own share before node-deployment phase. Thus for any two nodes *i* and *j*, node *i* can compute the common key $f(i,j)$ by evaluating $f(i,y)$ at point *j*, and vice visa.

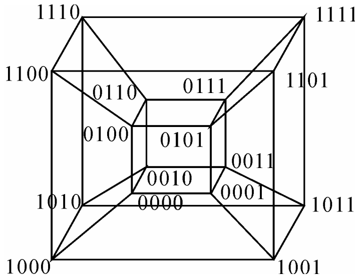


Figure 1. A 4-dimensional hypercube interconnection network.

To predistribute pairwise key with such a scheme as addressed above, node *i*'s storage overhead includes two parts: One is $(t+1)\log q$ storage space for storing a *t*-degree polynomial $f(i,y)$, the other is the storage space for its own ID information. [4] shows that this scheme has ability of *t*-collusion resistant. That is, if there exists no more than *t* compromised nodes in the network, the scheme can ensure the pairwise key is secure between any two normal nodes.

2.2.2. Polynomial Pool-based Key Predistribution

Pairwise key establishment in this scheme is processed in the following three phase: polynomial pool generation and key predistribution, direct key establishment, and path key establishment.

1) polynomial pool generation and key predistribution: This phase is mainly concerned with *t*-degree bivariate polynomial pool (*F*) generation over a finite field F_q. Then a subset $F_i \in F$ is selected and the shares of all of the polynomials in this subset are assigned to the node *i*.

2) direct key establishment: Assume that node *i* and *j* wants to establish pairwise key, if they have a common share on a same polynomial, they can establish pairwise key by utilizing the polynomial-based scheme. This phase is performed as follows: node *i* may broadcast an encryption list, α, $E_{K_v}(\alpha)$, v=1,2,…, $|F_i|$, where K_v is the share of the *vth* polynomial at point *j*. If node *j* can decrypts any one of these correctly, that means there exists a common share between the two nodes.

3) path key establishment: If there no pairwise key existent between node *i* and *j*, it's necessary to find a key path defined in Definition3. Then the two nodes transmits secret information for pairwise key generation on

this path.

2.2.3. Random Subset Assignment and Hypercube-Based Assignment

1) Random Subset Assignment: Different from polynomial scheme, the main idea of this assignment is to pick a random subset of polynomial pool, denoted as $F_i \in F$, and assign the share of this subset to node *i*.

2) Hypercube-based Assignment: Based on the concept of random subset, this assignment generates polynomial pool by utilizing hypercube model, and assign subsets to nodes according to node's ID.

3. H2 (Hierarchical Hypercube) Model

Definition 5 (*H*2 diagram): Assume that there exist 2^n nodes, the construction algorithm of *n*-dimension *H*2(*n*) is illustrated as follows:

1) Each $2^{\lceil n/2 \rceil}$ nodes are connected as a $\lceil n/2 \rceil$ dimensional hypercube, in which nodes are coded from $\underbrace{00...0}_{\lceil n/2\rceil}-\underbrace{11...1}_{\lceil n/2\rceil}$, and such kind of node code is called Inner-Hypercube-Node-Code. As a result, $2^{\lceil n/2 \rceil}$ different such kind of $\lceil n/2 \rceil$ dimensional hypercubes can be formed, where $\lfloor\ \rfloor$ represents the upper integer operation, and $\lceil\ \rceil$ means the lower integer operation.

2) The obtained $2^{\lceil n/2 \rceil}$ different such kind of $\lceil n/2 \rceil$ dimensional hypercubes are codes from $\underbrace{00...0}_{\lceil n/2\rceil}-\underbrace{11...1}_{\lceil n/2\rceil}$, and such kind of node code is called Outer-Hypercube-Node-Code. And then, the nodes in the $2^{\lceil n/2 \rceil}$ different such kind of $\lceil n/2 \rceil$ dimensional hypercubes with the same Inner-Hypercube-Node-Code are connected as a $\lceil n/2 \rceil$ dimensional hypercube, so we can obtain $2^{\lceil n/2 \rceil}$ different such kind of $\lceil n/2 \rceil$ dimensional hypercubes.

3) The graph constructed through the above two steps is called a *H*2 graph. And it is obvious that each node in the H2 graph is coded as (*r*,*h*), where *r* ($\underbrace{00...0}_{\lceil n/2\rceil} \leqslant r \leqslant \underbrace{11...1}_{\lceil n/2\rceil}$) is the node's Inner-Hypercube-Node-Code, and *h* ($\underbrace{00...0}_{\lceil n/2\rceil} \leqslant h \leqslant \underbrace{11...1}_{\lceil n/2\rceil}$) is the node's Outer- Hypercube-Node-Code.

Theorem 1: There exist 2^n in *H*2(*n*) diagram.

Proof: The conclusion is naturally held as $2^n = 2^{\lceil n/2 \rceil} * 2^{\lceil n/2 \rceil}$.

Theorem 2: The diameter of *H*2 (*n*) is *n*.

Proof: As the diameter of $\lceil n/2 \rceil$ dimension hypercube is $\lceil n/2 \rceil$, and it is naturally held for the case of $\lceil n/2 \rceil$ dimension hypercube. Thus the diameter of *H*2(*n*) is $\lceil n/2 \rceil + \lceil n/2 \rceil = n$ according to definition5.

Theorem 3: The distance of any two nodes $A(r_1,h_1)$ and $B(r_2,h_2)$ in *H*2(*n*) is expressed as $d\ (A,B)= d_h(r_1, r_2)+d_h(h_1, h_2)+1$ where d_h is *Hamming* distance.

Proof: Since the distance of any two nodes is the *Hamming* distance of their corresponding codes, it is held according to definition5.

4. Pairwise Key Establishment Scheme Based on H2 Model

As addressed above, polynomial-based and polynomoial-based schemes have some limitations. In this section we propose a new pairwise key establishment and predistribution scheme based on H2 model. The new algorithm is composed of three phases: polynomial pool generation and key predistribution, direct key establishment, and path key establishment.

4.1. Polynomial Pool Generation and Key Predistribution

Assume that there are N nodes in a wireless sensor network, where $2^{n-1} < N \leq 2^n$. A n-dimension *H*2(*n*) is then generated and we construct a polynomial pool with the following method:

1) The key setup server randomly generates $n*2^n$ bivariate *t*-degree polynomial pool over a finite fields F_q ,denoted as $F=\{ f^i_{<i_1,i_2,...,i_{\lfloor n/2\rfloor-1}>}(x,y),\ f^j_{<j_1,j_2,...,j_{\lceil n/2\rceil-1}>}(x,y) \mid 0 \leq i_1 \leq i_2 \leq ... \leq i_{\lfloor n/2\rfloor-1} \leq 1, 1 \leq i \leq \lceil n/2 \rceil,\ 0 \leq j_1 \leq j_2 \leq ... \leq j_{\lceil n/2\rceil-1} \leq 1, 1 \leq j \leq \lceil n/2 \rceil \}$.

2) The $2^{\lceil n/2\rceil-1}$ bivariate polynomials, denoted as $\{ f^j_{<j_1,j_2,...,j_{\lceil n/2\rceil-1}>}(x,y) \mid 0 \leq j_1 \leq j_2 \leq ... \leq j_{\lceil n/2\rceil} \leq 1\}$, where $1 \leq j \leq \lceil n/2 \rceil$, are assigned to the *jth* dimension of the $(i_1, i_2,..., i_{\lfloor n/2 \rfloor})$ th hypercube in *H*2(*n*).

3) The $2^{\lceil n/2\rceil-1}$ bivariate polynomials, denoted as$\{ f^i_{<i_1,i_2,...,i_{\lfloor n/2\rfloor-1}>}(x,y) \mid 0 \leq i_1 \leq i_2 \leq ... \leq i_{\lfloor n/2\rfloor-1} \leq 1\}$ where1 $\leq i \leq \lceil n/2 \rceil$, are assigned to the *ith* dimension of the $(j_1, j_2,..., j_{\lceil n/2 \rceil})$ th hypercube in *H*2(*n*).

4) For any nodes $((i_1, i_2,..., i_{\lfloor n/2 \rfloor}), (j_1, j_2,..., j_{\lceil n/2 \rceil}))$ in *H*2(*n*), the polynomial shares, denoted as $\{ f^1_{<j_2,...,j_{\lceil n/2\rceil}>}(x,y),\ f^2_{<j_1,j_3,...,j_{\lceil n/2\rceil}>}(x,y),\ldots,\ f^{\lceil n/2\rceil}_{<j_1,j_2,...,j_{\lceil n/2\rceil-1}>}(x,y)\} \cup \{ f^1_{<i_2,i_3,...,i_{\lfloor n/2\rfloor}>}(x,y),\ f^2_{<i_1,i_3,...,i_{\lfloor n/2\rfloor}>}$

(x,y),…, $f^{\lfloor n/2\rfloor}_{<i_1,i_2,\ldots,i_{\lfloor n/2\rfloor-1}>}(x,y)$}, are assigned and pre-loaded before deployment phase.

5) The server assigns a unique ID, denoted as $((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$, to every node in sequence, where $0 \le i_1 \le i_2 \le \ldots \le i_{\lfloor n/2\rfloor} \le 1$, $0 \le j_1 \le j_2 \le \ldots \le j_{\lceil n/2\rceil} \le 1$.

4.2. Direct Key Establishment

If any two nodes $A((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$ and $B((i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$ want to establish pairwise key, the node A can achieve the pairwise key with B by processing the following procedures:

Node A first computes the *Hamming* distance between B and itself, as $d_1=d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))$, $d_2=d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$. If $d_1=1$ or $d_2=1$, the node can establish the pairwise with the peer according to the conclusion of the Theorem 4.

Theorem 4: For any two nodes $A((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$ and $B((i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$, If the *Hamming* distance $d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))=1$, or $d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))=1$, then there exists certainly pairwise key between A and B.

Proof: 1) $d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))=1$: Assume that $i_t=i'_t$, where $1 \le t \le \lfloor n/2\rfloor-1$. Since $i_{\lfloor n/2\rfloor} \ne i'_{\lfloor n/2\rfloor} \Rightarrow f^{\lfloor n/2\rfloor}_{<i_1,i_2,\ldots,i_{\lfloor n/2\rfloor-1}>}(i_{\lfloor n/2\rfloor}, i'_{\lfloor n/2\rfloor}) = f^{\lfloor n/2\rfloor}_{<i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor-1}>}(i'_{\lfloor n/2\rfloor}, i_{\lfloor n/2\rfloor})$. So, There exists a pairwise key $f^{\lfloor n/2\rfloor}_{<i_1,i_2,\ldots,i_{\lfloor n/2\rfloor-1}>}(i_{\lfloor n/2\rfloor}, i'_{\lfloor n/2\rfloor})$ between A and B.

2) $d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))=1$: Imitating the step 1), it is easy to prove that there exists a pairwise key between A and B.

4.3. Indirect Key Establishment

If $d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))>1$ and $d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))>1$ then node A establish indirect pairwise key with B according to Theorem 5. In order to make it clear, we will provide a lemma before the illustration of Theorem 5.

Lemma1: For any two nodes $A((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$ and $B((i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$, assume that $d_h=k$, then there exists a k-distance path denoted as $I_0(=A)$, $I_1,\ldots,I_{k-1}$, $I_k(=B)$, where $d_h(I_i,I_j)=1$.

Proof: According to Theorem 3, d_h can be expressed as $d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))+d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))=k$. Assume that $d_h((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}))=h$, then $d_h((j_1,j_2,\ldots,j_{\lceil n/2\rceil}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))=k-h$. According to definition5, node $C((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$ and A are located in a $\lceil n/2\rceil$ -dimensional hypercube H, and node C and B are located in a $\lceil n/2\rceil$-dimensional hypercube H'. According to the properties of hypercube [5,6], there exist a path described as $I_0(=A)$, $I_1,\ldots, I_{h-1}$, $I_h(=C)$ in H, where $d_h(I_i, I_j)=1$. Similarly, another path with the same property is existed in H', denoted as $I_h(=A)$, I_{h+1}, …, I_{k-1}, $I_k(=B)$, where $d_h(I_i,I_j)=1$.

Thus there exist a integrated path in $H2$ diagram from node A to B, described as $I_0(=A)$, $I_1,\ldots,I_{k-1}$, $I_k(=B)$ where $d_h(I_i, I_j)=1$.

Theorem 5: Assume that any two nodes can communicate directly in a wireless sensor networks, and there is no compromised node in the networks, then there exist a key path for any node $A((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$ and node $B((i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$.

Proof: According to Lemma1, there exist a path for any two nodes where $d_h=k$ in $H2$ diagram. Thus the conclusion is held.

We propose the algorithm for indirect key establishment as follows. Assume the two nodes $A((i_1,i_2,\ldots,i_{\lfloor n/2\rfloor}), (j_1,j_2,\ldots,j_{\lceil n/2\rceil}))$ and node $B((i'_1,i'_2,\ldots,i'_{\lfloor n/2\rfloor}), (j'_1,j'_2,\ldots,j'_{\lceil n/2\rceil}))$ want to establish indirect pairwise key in the network, we propose the algorithm for indirect key establishment illustrated as follows.

Indirect_Key_Establishing_Algorithm(){

1) Node A computes a set L which records the dimensions in which node A and B have different sub-indexes. The set can be expressed as $L=\{(d_1,d_2,\ldots,d_k),(g_1,g_2,\ldots,g_w)\}$ where $d_1< d_2<\ldots< d_k$, $g_1<g_2<\ldots<g_w$.

2) Node A maintains a path set P with initial vale of $P=\{A\}$.

3) Assume that $U((u_1,u_2,\ldots,u_{\lfloor n/2\rfloor}), (u'_1,u'_2,\ldots,u'_{\lceil n/2\rceil}))=A$; $s=1$.

4) Node A computes intermediate nodes expressed as $V=((u_1,u_2,\ldots u_{d_s-1}, i'_{d_s}, u_{d_s+1},\ldots,u_{\lfloor n/2\rfloor}), (u'_1,u'_2,\ldots,u'_{\lceil n/2\rceil}))$. And $P=P \cup \{V\}$.

5) Assume that $U = V$.

6) If $s<k$, then $s=s+1$, and repeats the step 4, otherwise turns to step7).

7) Node A computes intermediate nodes $V=$ ($(u_1, u_2, \dots, u_{\lfloor n/2 \rfloor})$, $(u'_1, u'_2, \cdots u'_{g_s-1}, j'_{g_s}, u'_{g_s+1}, \dots, u'_{\lceil n/2 \rceil})$), and let $P=P \cup \{V\}$.

8) Let $U = V$.

9) If $s<w$, then $s=s+1$, and repeats step7); otherwise go on step10).

10) Let $P=P \cup \{B\}$.

}

According to Theorem 5, any node can compute a key path to it destination when there is no compromised node in the network. Once the path P is achieved, the two nodes can exchange secret information to generate pairwise key between themselves.

For example, the node A((001), (0101)) and the node B((100), (1100)) can establish pairwise key along the following key path: A((001), (0101)) → ((101), (0101)) → ((100), (0101)) → ((100), (1101)) → B((100), (1100)).

According to the algorithm described above, the following conclusion is naturally held.

Theorem 6: Assume that any two nodes can communicate with each other directly, and there is no compromised node in a network. If the distance between the two nodes is k, then there exists a key path with distance of k. That is, the two nodes can establish pairwise key through k-1 intermediate nodes.

4.4. Dynamic Key Path Establishment

The Indirect_Key_Establishing_Algorithm() illustrated in Subsection 4.3 can only deal with the situation that there is no compromised node in the network. However, in case of some existent compromised nodes, the algorithm would fail to find fungible intermediate node to help establish pairwise key.

We further analyze the example addressed in Subsection 4.3. When the node ((101),(0101)) is compromised, the node A and B can utilize the following path to establish pairwise key: A((001), (0101)) → ((000),(0101)) → ((100), (0101)) → ((100), (1101)) → B((100), (1100)).

When the node ((100),(1101)) is compromised, the two nodes can use the path: A((001), (0101)) → ((101), (0101)) → ((100), (0101)) → ((100), (0100)) → B((100), (1100)).

In case that the nodes ((101),(0101)),((100),(1101)) are compromised, there still exists a key path denoted as A((001),(0101)) → ((000),(0101)) → ((100),(0101)) → ((100),(0100)) → B((100),(1100)).

4.4.1. Relative Definitions of Local Weak Connectivity

Definition 6: The nodes A and B in a n-dimensional hypercube H_n are called neighbors, if that there exists only one different bit in their binary strings.

Definition 7: The node A in an m-dimensional hypercube/sub-hypercube H_m is m-disconnected, iif that all links between A and every faultless node in H_m are fault. The node A in an m-dimensional hypercube/sub- hypercube H_m is reachable, iif that A is faultless and not m-disconnected.

Definition 8 (k-dimensional local-weak-connectivity): A n-dimensional hypercube H_n is k-dimensional local-weak-connected, if all reachable nodes in each k-dimensional sub-hypercube H_k ($k \geq 1$) of H_n forms a connected graph, and the number of reachable nodes in H_k is bigger than 2^{k-1}.

Definition 9 (general local-weak-connectivity): An n-dimensional hypercube H_n is general local-weak- connected, if there exists a h-dimensional sub-hypercube H_h ($h \geq k$), which is local-weak-connected and includes H_k, as for each k-dimensional sub-hypercube H_k ($k \geq 1$) of H_n.

Figure 2 presents a 3-dimensional hypercube H_3 with two fault nodes and a 3-disconnected node. According to the above two kinds of local-weak-connectivity concepts, it is easy to prove that all reachable nodes in H_3 is global connected.

4.4.2. Global Connectivity of Local-Weak-Connected Hypercube

An n-dimensional hypercube H_n has 2^n nodes, in which each node can be represented by a binary string and has n different links. So H_n has n 2^{n-1} different links totally.

Definition 10: Any binary string $b_1\, b_2 \cdots b_{n-k}$ with given length n-k corresponds a k-dimensional sub-hypercube H_k with 2^k nodes, and the nodes in H_k can be represented by such binary strings as $b_1\, b_2 \cdots b_{n-k}$ *…*, where * can be 0 or 1.

From the construction algorithm of H_k, it is easy to know that all k-dimensional sub-hypercubes are isomor-

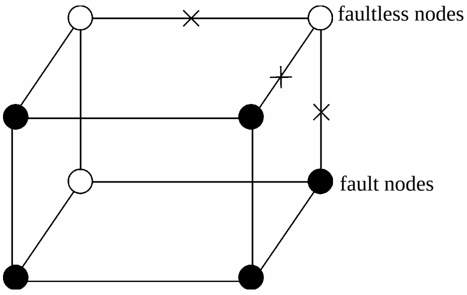


Figure 2. A 3-dimensional hypercube H3 with fault nodes and links. In which, the black dots represent fault nodes, the white dots represent faultless nodes, the lines with tokens represent fault links, and the lines without tokens represent faultless links.

phic, and H_n includes all of the k-dimensional sub-hypercubes that is isomorphic with H_k. And it is easy to prove that H_n includes 2^{n-k} -1 k-dimensional sub-hypercubes that is isomorphic with H_k and has no common nodes with H_k also.

Lemma 2: All of the reachable nodes in any two neighboring k-dimensional sub-hypercubes of the Local-Weak-Connected n-dimensional Hypercube H_n form a connected graph.

Proof: Let that H_k and H'_k are two neighboring k-dimensional sub-hypercubes, according to definition 10, we can utilize binary strings $b_1 \, b_2 \cdots b_{n-k} *\ldots*$ to represent the nodes in H_k, where * can be 0 or 1. Since H_k and H'_k are neighboring, so there exists at least one common node between H_k and $H'_k \Rightarrow$ the nodes in H'_k can be represented as $b_1 \cdots b_{t-1} * b_{t+1} \cdots b_{n-k} *\ldots* d_s *\ldots*$ and $b_1 \cdots b_{t-1} \, b_t \, b_{t+1} \cdots b_{n-k} *\ldots* d_s *\ldots* \in H_k$. $\Rightarrow b_1 \cdots b_{t-1} \, b_t \, b_{t+1} \cdots b_{n-k} *\ldots* d_s *\ldots* \in H_k \cap H'_k$. Considering that the number of reachable nodes in H'_k is bigger than $2^{k-1} \Rightarrow$ there exists at least one node A in those 2^{k-1} nodes represented by $b_1 \cdots b_{t-1} \, b_t \, b_{t+1} \cdots b_{n-k} *\cdots * d_s *\ldots*$ is reachable. Since $A \in H_n \Rightarrow$ Node A and all of the reachable nodes in H_k form a connected graph. And in addition, since $A \in H'_k \Rightarrow$ Node A and all of the reachable nodes in H'_k form a connected graph also. So, all of the reachable nodes in H_k and H'_k form a connected graph.

Theorem 7: All of the reachable nodes in n-dimensional Hypercube H_n, which satisfies the conditions of k-dimensional local-weak-connectivity, form a connected graph.

Proof: From Theorem 2, it is easy to prove that the conclusion stands.

From Theorem 7, we can obtain the following deduces easily.

Deduce 1: If n-dimensional Hypercube H_n satisfies the conditions of k-dimensional local-weak-connectivity, then all of the reachable nodes in any h-dimensional sub-hypercube $H_h(h \geq k)$ form a connected graph.

Deduce 2: If n-dimensional Hypercube H_n satisfies the conditions of k-dimensional local-weak-connectivity, then there exists at least a pair of connected reachable nodes $a_1 \cdots a_{j-1} \, 0 \, a_{j+1} \cdots a_{n-k} \, \chi_{n-k+1} \cdots \chi_n$ and $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} \, \chi'_{n-k+1} \cdots \chi'_n$ between any pair of k-dimensional sub-hypercubes of $a_1 \cdots a_{j-1} \, 0 \; a_{j+1} \cdots a_{n-k} *\ldots*$ and $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} *\ldots*$.

Proof: From deduce 1, if n-dimensional Hypercube H_n satisfies the conditions of k-dimensional local-weak-connectivity $\Rightarrow$ All of the reachable nodes in any h-dimensional sub-hypercube $H_h(h \geq k)$ form a connected graph $\Rightarrow$ All of the reachable nodes in any $(k+1)$-dimensional sub-hypercube $a_1 \cdots a_{j-1} * a_{j+1} \cdots a_{n-k} *\ldots*$ form a connected graph. And since the numbers of unreachable nodes in k-dimensional sub-hypercubes $a_1 \cdots a_{j-1} \, 0 \, a_{j+1} \cdots a_{n-k} *\ldots*$ and $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} *\ldots*$ are less than a half of the number of total nodes respectively. So, there exists reachable node $a_1 \cdots a_{j-1} \, 0 \, a_{j+1} \cdots a_{n-k} \, \chi_{n-k+1} \cdots \chi_n$ in $a_1 \cdots a_{j-1} \, 0 \, a_{j+1} \cdots a_{n-k} *\ldots*$, and there exists reachable node $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} \, \chi'_{n-k+1} \cdots \chi'_n$ in $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} *\ldots*$ certainly. And in addition, those two nodes $a_1 \cdots a_{j-1} \, 0 \, a_{j+1} \cdots a_{n-k} \, \chi_{n-k+1} \cdots \chi_n$ and $a_1 \cdots a_{j-1} \, 1 \, a_{j+1} \cdots a_{n-k} \, \chi'_{n-k+1} \cdots \chi'_n$ is connected.

Theorem 8: All of the reachable nodes in n-dimensional Hypercube H_n, which satisfies the conditions of general local-weak-connectivity, form a connected graph.

Proof: Since n-dimensional Hypercube H_n is local-weak-connected, and there exists no other sub-hypercubes that include itself in H_n. So, from definition 9, it is easy to know that H_n is n-dimensional local-weak-connected. And in addition, from definition 8, we can know that all of the reachable nodes in H_n form a connected graph.

Theorem 7 and Theorem 8 show that the hypercubes, that satisfy the conditions of the proposed two kinds of local-weak-connectivity, must be global connected.

4.4.3. K-Dimensional Local-Weak-Connectivity Based Dynamic Key Path Establishment Algorithm

KLWC-based Dynamic_Key_Path_Establishing_Algorithm(){

Input: Sensor network H_n with fault nodes and fault links (The links, whose length are bigger than the transmitting radius). And two reachable nodes A($(i_1, i_2, ..., i_{\lfloor n/2 \rfloor})$, $(j_1, j_2, ..., j_{\lceil n/2 \rceil})$) and B($(i'_1, i'_2, ..., i'_{\lfloor n/2 \rfloor})$, $(j'_1, j'_2, ..., j'_{\lceil n/2 \rceil})$) in H_n.

Output: A correct key path P from A to B in H_n.

1) Compute and determine the node T= ($(i'_1, i'_2, ..., i'_{\lfloor n/2 \rfloor})$, $(j_1, j_2, ..., j_{\lceil n/2 \rceil})$) in $H_{\lfloor n/2 \rfloor}$;

2) P=Dynamic_Key_Path_Establishment_1 (A,T);

3) P=P $\cup$ Dynamic_Key_Path_Establishment_2 (T, B);

4) If P is a correct key path from A to B, then exit, otherwise turn to step 5);

5) Compute and determine the node T=($(i_1, i_2, ..., i_{\lfloor n/2 \rfloor})$, $(j'_1, j'_2, ..., j'_{\lceil n/2 \rceil})$) in $H_{\lfloor n/2 \rfloor}$;

6) P=Dynamic_Key_Path_Establishment_2 (A,T);

7) $P=P \cup$ Dynamic_Key_Path_Establishment_1 (T, B);

8) If P is a correct key path from A to B, then exit, otherwise turn to step 9);

9) Report A, failure to establish a key path from A to B.

}

Algorithm Dynamic_Key_Path_Establishment_1(*A*,*T*):

1) Obtain the codes of nodes A and T: $A \leftarrow ((i_1,i_2,...,i_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$, $T \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$;

/* From definition 10, we can suppose that the k-dimensional sub-hypercube that includes T is $((i'_1\ i'_2 \cdots i'_{\lfloor n/2 \rfloor -k} * \ldots *), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$.*/

2) Initialize Path P: $P \leftarrow A$;

3) Initialize temporary binary string C: $C=((c_1,c_2,...,c_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil})) \leftarrow A$;

4) FOR(j=1; $j \le \lfloor n/2 \rfloor$; j++){

IF($i_j \neq i'_j$){

①According to lemma2 and deduce 1, a pair of connected reachable nodes C and D can be found through discovering in neighboring k-dimensional sub- hypercubes:

$C=((c_1,c_2,\cdots c_{j-1},c_j,c_{j+1},...,c_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$,

$D=((i'_1,i'_2,\cdots i'_{j-1},i'_j,x_{j+1},...,x_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$,

where $c_t = i'_t$ ($t \in [1, j]$);

②Join the path from $((i'_1,i'_2,\cdots i'_{j-1},i_j,i_{j+1},...,i_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ to node D into P;

③$C \leftarrow D$;

}

}

/* After the above steps, a correct key path from node $A((i_1,i_2,...,i_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ to the reachable node $((i'_1\ i'_2 \cdots i'_{\lfloor n/2 \rfloor -k}\ \chi'_{\lfloor n/2 \rfloor -k+1} \cdots \chi'_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ will be constructed. */

5) Join the path from node $((i'_1 i'_2 \cdots i'_{\lfloor n/2 \rfloor -k}\ \chi'_{\lfloor n/2 \rfloor -k+1} \cdots \chi'_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ to node T in the k-dimensional sub-hypercube $((i'_1\ i'_2 \cdots i'_{\lfloor n/2 \rfloor -k} **), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ into P. And then exit. So, a correct key path from node A to node T is discovered.

Algorithm Dynamic_Key_Path_Establishment_2(*T*, *B*):

1) Obtain the codes of nodes B and T: $T \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$; $B \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1,j'_2,...,j'_{\lceil n/2 \rceil}))$;

/* From definition 10, we can suppose that the k-dimensional sub-hypercube that includes B is $((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1\ j'_2 \cdots j'_{\lceil n/2 \rceil -k} * \ldots *))$.*/

2) Initialize Path P: $P \leftarrow T$;

3) Initialize temporary binary string C: $C=((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (c_1,c_2,...,c_{\lceil n/2 \rceil})) \leftarrow T$;

4) FOR(l=1; $l \le \lceil n/2 \rceil$; l++){

IF($j_l \neq j'_l$){

① According to lemma2 and deduce 1, a pair of connected reachable nodes C and D can be found through discovering in neighboring k-dimensional sub-hypercubes:

$C=((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (c_1,c_2,\cdots c_{l-1},c_l,c_{l+1},...,c_{\lceil n/2 \rceil}))$;

$D=((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1,j'_2,\cdots j'_{l-1},j'_l,x_{l+1},...,x_{\lceil n/2 \rceil}))$,

where $c_t = j'_t$ ($t \in [1, l]$).

② Join the path from $((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1,j'_2,\cdots j'_{l-1},j_l,j_{l+1},...,j_{\lceil n/2 \rceil}))$ to node D into P;

③ $C \leftarrow D$;

}

}

/* After the above steps, a correct key path from node $T((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j_1,j_2,...,j_{\lceil n/2 \rceil}))$ to the reachable node $((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1\ j'_2 \cdots j'_{\lceil n/2 \rceil -k}\ \chi'_{\lceil n/2 \rceil -k+1} \cdots \chi'_{\lceil n/2 \rceil}))$ will be constructed. */

5) Join the path from node $((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1\ j'_2 \cdots j'_{\lceil n/2 \rceil -k}\ \chi'_{\lceil n/2 \rceil -k+1} \cdots \chi'_{\lceil n/2 \rceil}))$ to node B in the k-dimensional sub-hypercube $((i'_1,i'_2,...,i'_{\lfloor n/2 \rfloor}), (j'_1\ j'_2 \cdots j'_{\lceil n/2 \rceil -k} **))$ into P. And then exit. So, a correct key path from node T to node B is discovered.

From the above description, we can know that the time complexity of algorithm Dynamic_Key_Path_ Establishment_1 is

$O((\lfloor n/2 \rfloor -k)2^k) + O(2^k) = O(n2^{k-1})$, and the time complexity of algorithm Dynamic_Key_Path_Establishment_2 is $O((\lceil n/2 \rceil - k)2^k) + O(2^k) = O(n2^{k-1})$, so the total time complexity of the k-Dimensional Local-Weak-Connectivity based Dynamic Key Path Establishment Algorithm is $O(n2^k)$.

Considering the percentage of the fault nodes in sensor networks, when applying the k-Dimensional Local-Weak-Connectivity based Dynamic Key Path Establishment Algorithm actually, we can set k=1,2,3. Then the total time complexity of the k-Dimensional Local-Weak-Connectivity based Dynamic Key Path Establishment Algorithm will be O（n） only. Figure 3 illustrates the relationship of dimension n and the scale of the sensor networks.

4.4.4. General Local-Weak-Connectivity Based Dynamic Key Path Establishment Algorithm

```
GLWC-based Dynamic_Key_Path_Establishing_Algorithm(){
    Input: Sensor network $H_n$ with fault nodes and fault links (The links, whose length are bigger than the transmitting radius). And two reachable nodes $A((i_1,i_2,...,i_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ and $B((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j'_1,j'_2,...,j'_{\lceil n/2\rceil}))$ in $H_n$.
    Output: A correct key path $P$ from $A$ to $B$ in $H_n$.
    1) Compute and determine the node $T=((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ in $H_{\lfloor n/2\rfloor}$;
    2)$P$=Dynamic_Key_Path_Establishment_3($A$, $T$);
    3)$P=P \cup$ Dynamic_Key_Path_Establishment_4($T$, $B$);
    4)If $P$ is a correct key path from $A$ to $B$, then exit, otherwise turn to step 5);
    5)Compute and determine the node $T=((i_1,i_2,...,i_{\lfloor n/2\rfloor}),(j'_1,j'_2,...,j'_{\lceil n/2\rceil}))$ in $H_{\lfloor n/2\rfloor}$;
    6)$P$=Dynamic_Key_Path_Establishment_4($A$, $T$);
    7)$P=P \cup$ Dynamic_Key_Path_Establishment_3($T$, $B$);
    8)If $P$ is a correct key path from $A$ to $B$, then exit, otherwise turn to step 9);
    9)Report $A$, failure to establish a key path from $A$ to $B$.
}
```

Algorithm Dynamic_Key_Path_Establishment_3(*A*, *T*):

```
1)Obtain the codes of nodes $A$ and $T$: $A \leftarrow ((i_1,i_2,...,i_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$, $T \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$;
/* From definition 10, we can suppose that the $k$-dimensional sub-hypercube that includes $T$ is $((i'_1\ i'_2 \cdots i'_{\lfloor n/2\rfloor-k} *\ldots*),(j_1,j_2,...,j_{\lceil n/2\rceil}))$.*/
2)Initialize Path $P$: $P \leftarrow A$;
3)Initialize temporary binary string $C$: $C=((c_1,c_2,...,c_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))\leftarrow A$;
4)FOR($j$=1; $j \le n$; $j$++){
    IF($i_j \neq i'_j$){
        FOR($k$=1; $k \le n$-$j$; $k$++){
            IF(According to Theorem 8, a pair of connected reachable nodes $C$ and $D$ can be found through discovering in neighboring $k$-dimensional sub-hypercubes:
            $C=((c_1,c_2,...c_{j-1},c_j,c_{j+1},...,c_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$,
            $D=((i'_1,i'_2,...i'_{j-1},i'_j,x_{j+1},...,x_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$, where $c_t=i'_t$ ($t\in$ [1,$j$]);
                ①Join the path from $((i'_1,i'_2,...i'_{j-1},i_j,i_{j+1},...,i_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ to node $((i'_1,i'_2,...i'_{j-1},i'_j,x_{j+1},...,x_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ into $P$;
                ②$C \leftarrow ((i'_1,i'_2,...i'_{j-1},i'_j,x_{j+1},...,x_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$;
                ③Break;
            }
        }
        IF($k > n$-$j$){
            WHILE($k \le n$){
                IF(In the $k$-dimensional hypercube $((c_1,c_2,...c_{\lfloor n/2\rfloor-k}*...*),(j_1,j_2,...,j_{\lceil n/2\rceil}))$, there exists no faultless key path from node $C$ to node $T$) $k$++;
            }
        }
        IF( $k > n$) exit. Then $H_{\lfloor n/2\rfloor}$ is not general local-weak-connected, and we cannot find a correct key path from node $A$ to $T$.
        ELSE Join the path from $C$ to $T$ in the $k$-dimensional hypercube $((c_1,c_2,...c_{\lfloor n/2\rfloor-k}*...*),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ into $P$;
    }
}
5) Exit. And a correct key path from $A$ to $T$ is discovered.
```

Algorithm Dynamic_Key_Path_Establishment_4(*T*, *B*):

```
1) Obtain the codes of nodes $T$ and $B$: $T \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j_1,j_2,...,j_{\lceil n/2\rceil}))$ and $B \leftarrow ((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j'_1,j'_2,...,j'_{\lceil n/2\rceil}))$;
/* From definition 10, we can suppose that the $k$-dimensional sub-hypercube that includes $B$ is $((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(j'_1\ j'_2 \cdots j'_{\lceil n/2\rceil-k} *\ldots*))$.*/
2) Initialize Path $P$: $P \leftarrow T$;
3) Initialize temporary binary string $C$: $C=((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(c_1,c_2,...,c_{\lceil n/2\rceil})) \leftarrow T$;
4) FOR($l$=1; $l \le n$; $l$++){
    IF( $j_l \neq j'_l$ ){
        FOR($k$=1; $k \le n$-$j$; $k$++){
            IF(According to Theorem 8, a pair of connected reachable nodes $C$ and $D$ can be found through discovering in neighboring $k$-dimensional sub-hypercubes:
            $C=((i'_1,i'_2,...,i'_{\lfloor n/2\rfloor}),(c_1,c_2,...c_{j-1},c_j,c_{j+1},...,c_{\lceil n/2\rceil}))$,
        $D$=( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ ,
        $(j'_1, j'_2, \dots j'_{l-1}, j'_l, x_{l+1}, \dots, x_{\lceil n/2 \rceil})$ ),
        where $c_t = i'_t$ ($t \in [1,j]$);
    ①Join the path from ( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ , $(j'_1, j'_2, \dots j'_{l-1}, j_l, j_{l+1}, \dots, j_{\lceil n/2 \rceil})$ ) to node ( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ , $(j'_1, j'_2, \dots j'_{l-1}, j'_l, x_{l+1}, \dots, x_{\lceil n/2 \rceil})$ ) into $P$;
    ②$C \leftarrow$ ( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ , $(j'_1, j'_2, \dots j'_{l-1}, j'_l, x_{l+1}, \dots, x_{\lceil n/2 \rceil})$ );
    ③Break;
      }
    }
    IF($k > n-j$){
    WHILE($k \le n$){
    IF(In the $k$-dimensional hypercube ( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ , $(c_1, c_2, \dots c_{\lceil n/2 \rceil - k} * \dots *)$ ), there exists no faultless key path from node $C$ to node $B$) $k$++;
      }
    }
    IF($k > n$) exit. Then $H_{\lceil n/2 \rceil}$ is not general local-weak-connected, and we cannot find a correct key path from node $T$ to $B$.
    ELSE Join the path from $C$ to $B$ in the $k$-dimensional hypercube ( $(i'_1, i'_2, \dots, i'_{\lfloor n/2 \rfloor})$ , $(c_1, c_2, \dots c_{\lceil n/2 \rceil - k} * \dots *)$ ) into $P$;
      }
    }
    5) Exit. And a correct key path from $T$ to $B$ is discovered.
```

From the above description, we can know that the time complexity of algorithm Dynamic_Key_Path_ Establishment_3 is O ($\lfloor n/2 \rfloor 2^{k_{\min}}$) + O ($2^{k_{\min}}$) = O ($\lfloor n/2 \rfloor 2^{k_{\min}}$), where k_{min} is the smallest integer that satisfies the condition of k-dimensional local-weak-connectivity. And the time complexity of algorithm Dynamic_Key_Path_Establishment_4 is O ($\lceil n/2 \rceil 2^{k_{\min}}$) + O ($2^{k_{\min}}$) = O ($\lceil n/2 \rceil 2^{k_{\min}}$), so the total time complexity of the general Local-Weak- Connectivity based Dynamic Key Path Establishment Algorithm is O ($\lceil n/2 \rceil 2^{k_{\min}}$).

Considering the percentage of the fault nodes in sensor networks, when applying the general Local- Weak-Connectivity based Dynamic Key Path Establishment Algorithm actually, we can set k=1,2,3. Then the total time complexity of the General Local-Weak- Connectivity based Dynamic Key Path Establishment Algorithm will be O(n) only.

5. Analysis

5.1. Feasibilities of the Algorithm

Theorem 9: In our algorithm, the possibility of direct key establishment for any two nodes can be expressed as $P_{H2} \approx (2^{\lfloor n/2 \rfloor} + 2^{\lceil n/2 \rceil})/(N\text{-}1)$.

Proof: As the algorithm has assigned any node, denoted as ($(i_1, i_2, \dots, i_{\lfloor n/2 \rfloor})$, $(j_1, j_2, \dots, j_{\lceil n/2 \rceil})$), shares of polynomials expressed as F_A = { $f^1_{<j_2, \dots, j_{\lceil n/2 \rceil}>}(j_1, y)$, $f^2_{<j_1, j_3, \dots, j_{\lceil n/2 \rceil}>}(j_2, y), \dots,$ $f^{\lceil n/2 \rceil}_{<j_1, j_2, \dots, j_{\lceil n/2 \rceil - 1}>}(j_{\lceil n/2 \rceil}, y)$} $\cup$ { $f^1_{<i_2, \dots, i_{\lfloor n/2 \rfloor}>}(i_1, y)$, $f^2_{<i_1, i_3, \dots, i_{\lfloor n/2 \rfloor}>}(i_2, y), \dots,$ $f^{\lfloor n/2 \rfloor}_{<i_1, i_2, \dots, i_{\lfloor n/2 \rfloor - 1}>}(i_{\lfloor n/2 \rfloor}, y)$}. It's clear that there are $2^{\lfloor n/2 \rfloor} + 2^{\lceil n/2 \rceil}$ nodes which can establish direct pairwise key with the node A. Thus $P_{H2} \approx (2^{\lfloor n/2 \rfloor} + 2^{\lceil n/2 \rceil})/(N\text{-}1)$ as the network scale is within the area $2^{n-1} < N \le 2^n$.

Suppose that a sensor network has N=10000 sensor nodes, then n=14. The possibility of direct key establish is about $P_{H2} \approx$ 2.56% according to the conclusion drawn by Theorem 6. However, the possibility decreases to $P_H \approx$ 0.14% if the algorithm addressed in [3] is used.

Theroem10: Assume that the possibility of direct key establishment in H2-based scheme is defined as P_{H2}, while the possibility in hypercube is denoted as P_H, then $P_{H2} >> P_H$.

Proof: Suppose the number of a network is within the area of $2^{n-1} < N \le 2^n$, and $P_H \approx \frac{n}{N-1}$ as addressed in [3]. Thus $\lim_{n \to \infty} \frac{P_H}{P_{H2}} = \lim_{n \to \infty} \frac{n}{2^{\lfloor n/2 \rfloor} + 2^{\lceil n/2 \rceil}}$ =0.

5.2. Overhead Analysis

Node's Storage Overhead

1) Any node is required to store t-degree bivariate polynomials whose number is n over the finite fields q, which occupies $n(t+1)\log q$ bits.

2) In order to keep the security of the Keys, for any bivariate polynomial $f(x,y)$, node A is required to store the ID information of the compromised nodes that can establish direct key with A by using $f(x,y)$. Since the degree of $f(x,y)$ is t, then $f(x,y)$ will be divulged when there are more than t nodes are compromised. So, for any bivariate polynomial $f(x,y)$, node A needs only to store the ID information of n compromised nodes that can establish direct key with A by using $f(x,y)$. In addition, since the node's ID is a vector of n bits, and from Theorem 4, we can know that node A needs only to store one bit for each compromised node to determine the whole ID information of the compromised node. So, the total storage cost is *nt bits*.

3) Also the node's own ID information occupies about *n bits* storage space, as it is expressed as $((i_1, i_2, ..., i_{\lfloor n/2 \rfloor}), (j_1, j_2, ..., j_{\lceil n/2 \rceil}))$.

All of the storage overhead address above sum up to $n(t+1)\log q+nt+n= n(t+1)\log 2q$ *bits*.

Theorem 11: The *H*2-based and the hypercube-based schemes have the same storage overhead.

Proof: According to the analysis on storage overhead addressed in Subsection 5.4 in [3], the result is certainly held.

Communication Overhead

In a sensor network, sending a unicast message between two arbitrary nodes may involve the overhead of establishing a route. In case of no compromised node existent in the network, any one node can communicate with the others directly. Assume that the overhead for a hop is defined as 1, then for two arbitrary nodes whose Hamming distance is *L*, the minimum communication overhead is *L*. We further inspect average communication overhead on *H*2-based path key establishment.

Suppose there are two nodes $A((i_1, i_2, ..., i_{\lfloor n/2 \rfloor}), (j_1, j_2, ..., j_{\lceil n/2 \rceil}))$ and $B((i'_1, i'_2, ..., i'_{\lfloor n/2 \rfloor}), (j'_1, j'_2, ..., j'_{\lceil n/2 \rceil}))$ In the formal part of node's code, the probability of $i_e = i'_e$, $e \in \{1, \ldots, \lceil n/2 \rceil\}$ is 1/2; Similarly, the probability of $j_e = j'_e$, $e \in \{1, \ldots, \lceil n/2 \rceil\}$ is also 1/2 in the latter code part. Thus the probability for the two nodes to have *i* different sub-index in the formal part is expressed as *P*[*i different sub-indexs in former part*]= $\frac{1}{2^{\lfloor n/2 \rfloor}} \frac{(\lfloor n/2 \rfloor)!}{i!(\lfloor n/2 \rfloor - i)!}$. In the latter part, we also have:

P[*j different sub-indexs in later part*]= $\frac{1}{2^{\lceil n/2 \rceil}} \frac{(\lceil n/2 \rceil)!}{j!(\lceil n/2 \rceil - j)!}$.

Thus the average communication overhead can be summarized as:

$$L = \sum_{i=1}^{\lfloor n/2 \rfloor} (i-1) \times P[\text{i different sub-indexs in former part}]$$
$$+ \sum_{j=1}^{\lceil n/2 \rceil} (j-1) \times P[\text{j different sub-indexs in former part}].$$

Theroem12: The average communication overhead in the *H*2-based scheme is less than that in the hypercube-based scheme.

Proof: According to the analysis on communication overhead addressed in Subsection 5.4 in [3], the result is certainly held.

Figure 5 shows that the comparison on communication overhead between the *H*2-based scheme and the hypercube-based scheme.

5.3. Security Analysis

Here we put focus on two types of attacks against *H*2-based scheme: 1) An adversary may compromise pairwise key between any two nodes or prevent them to establish a pairwise key. 2) The adversary may focus its power to attack against the whole network, for purpose of lowering the probability of pairwise key establishment, or in creasing communication cost.

5.3.1. Attacks against Pairwise Key between Two Nodes

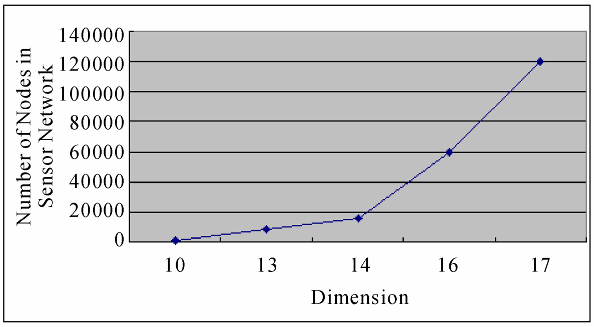


Figure 3. The relationship of dimension *n* and the scale of the sensor networks.

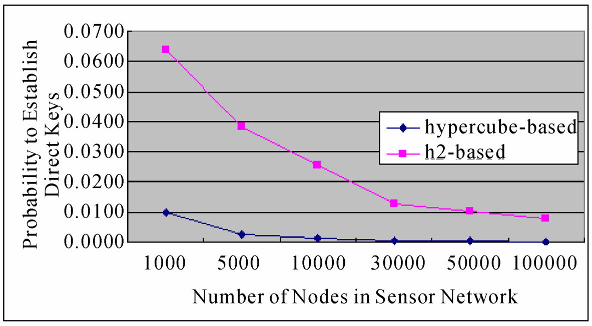


Figure 4. The comparison of probability to establish direct key between H2-based and Hypercube-based algorithms.

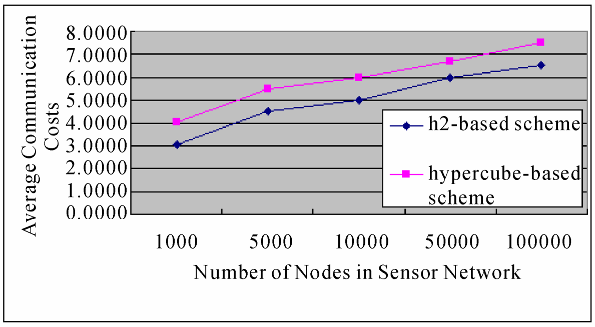


Figure 5. The comparison on average communication overhead between the H2-based and the Hypercube-based schemes.

1) Suppose an adversary launches an attack against two particular nodes, in order to filch their pairwise key. In case that those two nodes are not compromised:

① If the node *u and v* can establish direct pairwise key, the only means to compromise the key is to resolve the polynomial $f(x,y)$, which is shared by the two nodes. As the degree of this polynomial is adversary t, the adversary is required to compromise at least t+1 compromised nodes with the same share of $f(x,y)$.

② If the node u and v need to establish indirect pairwise key, the adversary is required to compromise an intermediate node, or filch the common share of the bivariate polynomial $f(x,y)$ between the two nodes. However, even if the adversary succeeds to achieve the pairwirse key, the nodes u and v can also select alternatives to re-establish key path.

2) Suppose the adversary launch attacks to prevent pairwise key establishment against two particular nodes, denoted as u and v, which are assumed are not compromised. Then the adversary is required to compromise n bivariate polynomials of the node u or v. Notes that to those polynomials are t-degree, which means that if such attacks succeeded, at least $n(t+1)$ nodes should have been compromised.

As addressed above and analysis presented in Subsection 5.5.1 in [3], if an adversary launches attacks against nodes, the security of the *H*2-based scheme if equivalent with that of the hypercube-based scheme. That is, we have the following theorem.

Theorem 13: The security of the *H*2-based scheme if equivalent with that of the hypercube-based scheme.

5.3.2. Attacks against the Whole Network

Suppose that an adversary has known the distribution state of polynomials for each node, he would launch attacks against the whole network systematically by compromising polynomials one by one. Assume that the adversary has compromised l bivariate polynomials, which means that at most $l2^{\lceil n/2 \rceil}$ nodes have been pre-loaded one of those compromised polynomials. However, the rest of the regular nodes, denoted as $N\text{-}\,l2^{\lceil n/2 \rceil}$ do not contain compromised polynomial shares. That means $N\text{-}\,l2^{\lceil n/2 \rceil}$ nodes can still work properly. Notice that those regular nodes should avoid to use compromised shares to establish pairwise key.

Clearly, the number of nodes influenced by adversaries in the *H*2-based scheme is more than that in the hypercube-based scheme. However, on the condition that the adversary fails to compromise all of the polynomials, the effected nodes can select other regular nodes to establish pairwise key with others.

In addition, it has proved that the probability of direct key establishment in the *H*2-based scheme is much higher than that in the hypercube-based scheme. Thus in the process of direct key establishment among non-compromised nodes, the degree of the influence cause by adversaries on the *H*2-based scheme is less than that on the hypercube-based scheme. That means the the *H*2-based scheme has ability to secure communications among nodes effectively in sensor networks.

5.3.3. Security Performance

Based on the nice properties of fault tolerance in *H*2-baed scheme, a source node can re-establish pairwise key with the destination by selecting alternative key path.

As addressed in Subsection 4.1, the polynomial pool has $n*2^n$ bivariate t-degree polynomials, that is, $|F|=n*2^n$; As every node contains n different polynomial shares, given a particular share of a bivariate polynomial f, the probability for each node to contain such a share is $n/|F|$. Assume that the number of nodes in a network is $2^{n-1}<N \le 2^n$, and the number of supposed compromised node is N_C, the probability for those compromised nodes to contain i shares of f is

$$P_i = \frac{N_c!}{(N_c-i)!i!}\left(\frac{n}{|F|}\right)^i\left(1-\frac{n}{|F|}\right)^{N_c-i}$$

As the adversary needs to compromise at least t+1 nodes to filch f, the probability of being compromised for f is $P_c=1-\sum_{i=0}^{t} P_i$.

According to Theorem 6, the compromised probability of direct key establishment for any two non-compromised nodes is expressed as $P_{link}= P_c \times P_{H2}$, in case that a particular polynomial f is compromised.

Figure 6 shows the fraction of compromised direct keys between non-comrpomised nodes as a function of the number of compromised keys for H2 and hypercube-based schemes where N=30000 and t=2.

Figure 6 shows that based on the assumption of same network scale and the proportion of compromised nodes, *H*2-based scheme provides higher probability than hypercube-based scheme for direct key establishment between any two non-compromised nodes. *H*2-based scheme would not fail to establish direct key until the proportion increases to 40%, while for *Hypercube*-based scheme, accepted proportion is about 30%.

We further inspect the probability of compromised indirect key. As addressed in Theorem 6, the probability of direct key establishment for any two nodes is $P_{H2} \approx (2^{\lfloor n/2 \rfloor}+2^{\lceil n/2 \rceil})/(N\text{-}1)$, the probability of indirect key establishment can be expressed as 1-$P_{H.}$ Thus the probability of compromised indirect key is estimated as $(1\text{-}P_{H2})[1\text{-}(1\text{-}\frac{N_C}{N})\times(1\text{-}P_c)^2]$.

Figure 7 shows that the probability of compromised indirect key between any two non-compromised nodes is a function of the fraction of compromised nodes where N=30000 and t =2.

Figure 7 shows that based on the same conditions of network scale and fraction of compromised nodes, *H*2-based scheme has better performance than hypercube-based scheme on indirect key establishment. The figure also shows that *H*2-based scheme would not fail to establish indirect key until the fraction of compromised nodes rises up to 60%. However, the fraction is only about 40% for *Hypercube*-based scheme.

Here we consider overall security performance of the two schemes. We define the probability of compromised pairwise key (direct or indirect key) is

$$P_{key}=P_{H2}\times P_c+(1-P_{H2})[1-(1-\frac{N_c}{N})\times(1-P_c)^2].$$

Figure 8 shows that the probability of compromised pairwise key is a function of the fraction of compromised nodes where N=30000 and t=2 for the two schemes.

From Figure 8, we can know that the probability of the pairwise key between any two non-compromised nodes when the *H*2-based scheme is applied, is lower than that when the *Hypercube*-based scheme is applied, supposing that the scale and percentage of compromised nodes of the sensor networks are the same.

So, from the above description, it is obvious that the security performance of the *H*2-based scheme is better than that of *Hypercube*-based scheme.

5.4. The Probability of Pairwise Key Re-estab-lishment

A source node has to re-establish key path to the destination once some intermediate nodes have been compromised. According to the previous presented two kinds of dynamic key path establishing algorithms, it is easy to know that the algorithms can find a new alternative key path certainly, when k =1,2 or 3, as long as the distribution of the compromised nodes in the whole sensor network satisfy the conditions of 1,2 or 3- dimensional local-weak-connectivity. Next, lets analyze the probability of pairwise key re- establishment when the distrivution of the compromised node do not satisfy the conditions of 1,2 and 3-dimensional local- weak-connectivities.

According to the pairwise key establishment scheme addressed above, each node in a network is able to communicate $2^{\lfloor n/2\rfloor}+2^{\lceil n/2\rceil}$ nodes to establish direct pairwise key. Assume that the fraction of compromised nodes is p, then the number of non-compromised nodes among $2^{\lfloor n/2\rfloor}+2^{\lceil n/2\rceil}$ is $(1-p)*(2^{\lfloor n/2\rfloor}+2^{\lceil n/2\rceil})$. On the condition that a key path is available among those non-compromised nodes, it's certainly possible for a source node and the destination to establish indirect pairwise key. So, when the distrivution of the compromised node do not satisfy the conditions of 1,2 and 3-dimensional local-weak-

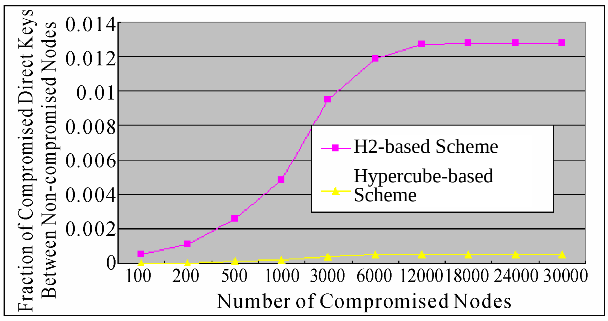


Figure 6. The relation between the fraction of compromised direct keys and the number of compromised nodes in H2- based scheme and Hypercube-based scheme.

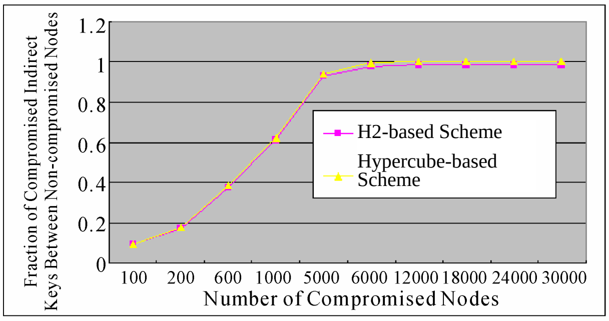


Figure 7. The comparison on the fraction of compromised indirect keys-number of compromised nodes relation between the H2-based scheme and the Hypercube-based scheme.

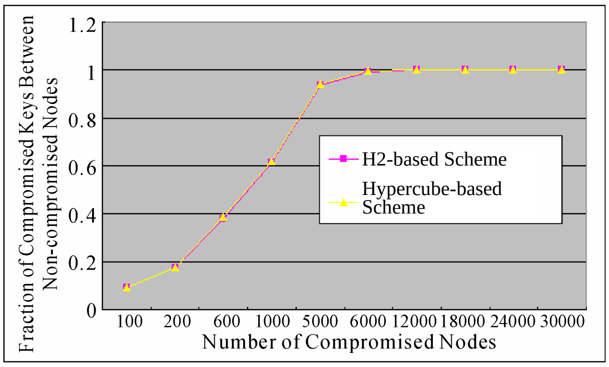


Figure 8. The comparison on the fraction of compromised keys- number of compromised nodes relation between the H2-based scheme and the Hypercube-based scheme.

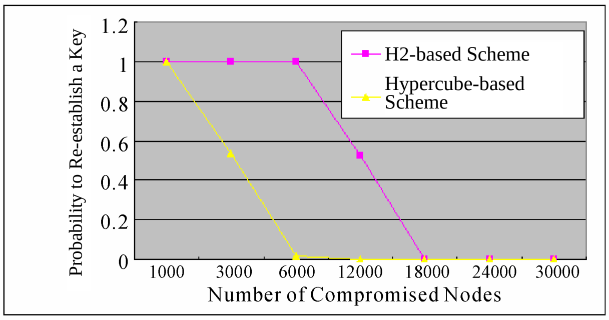


Figure 9. The relation between the probability of re-established keys and the number of compromised nodes in the H2-based and Hypercube-based schemes.

connectivities, the probability of pairwise key re-establishment can be estimated as:

$$P_{re}\text{=}1\text{-}[1-(1-p)^2(1-P_c)^2]^{(1-p)*(2^{\lfloor n/2\rfloor}+2^{\lceil n/2\rceil})}$$

Assume that *N*=30000 and *t*=2, Figure 9 shows that the probability of pairwise key re-establishment is a function of number of compromised nodes in *H*2 and hypercube-based schemes. It also shows that the probability of pairwise key re-establishment in *H*2 scheme is higher than that in hypercube-based scheme for any two non-compromised nodes.

6. Conclusions

An *H*2-based key predistribution scheme is proposed. Compared with polynomial pool-based scheme, it can improve working performance on probability of direct key establishment without additional storage requirement. Moreover, experimental figures show that our algorithm has lower communication cost and more secure than previous related works.

7. Acknowledgment

This work was supported in part by the China Scholarship Council (CSC) in form of a scholarship that funds the author, Lei Wang, studying at Lakehead University as a visiting researcher in Canada.

8. References

[1] L. Eeschnaure and V. D. Gligor, “A key-management scheme for distributed sensor networks,” in proceedings of the 9th ACM Conference on Computer and Communication Security, pp. 41−47, 2002.

[2] H. Chan, A. Oerrig, and D. Song, “Random key predistribution schemes for sensor networks,” in IEEE Syposium on Research in Security and Privacy, pp. 197−213, 2003.

[3] D. G. Liu, P. Ning, and R. F. Li, “Establishing pairwise keys in distributed sensor networks,” ACM Journal Name, Vol. 20, pp. 1−35, 2004.

[4] C. Blundo, A. Desantis, S. Kutten, *et al*., “Perfectly secure key distribution for dynamic conferences,” in Advances in Cryptology-CRYPTO’92, LNCS, 740, pp. 471−486, 1993.

[5] L. Wang and Y. P. Lin, “Maximum safety-path matrices based fault-tolerant routing for hypercube multi-computers,” Journal of Software, Vol. 15, No. 7, pp. 994−1004, 2004.

[6] L. Wang and Y. P. Lin, “A fault-tolerant routing strategy based on maximum safety-path vectors for hypercube multi-computers,” Journal of China Institute of Communications, Vol. 16, No. 4, pp. 130−137, 2004.

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Adaptive Processing Gain Data Services in Cellular CDMA in Presence of Soft Handoff with Truncated ARQ

Dipta DAS (Chaudhuri)*, Sumit KUNDU+
**Department of Electronics & Communication Engg, Dr. B. C. Roy Engg College, Durgapur, India*
+Department of Electronics & Communication Engg, National Institute of Technology, Durgapur, India
Email: *dipta_chaudhuri07@rediffmail.com, sumit.kundu@nitdgp.ac.in*

Received October 27, 2008; *revised December* 8, 2008; *accepted February* 11, 2009

ABSTRACT

An adaptive data transmission scheme based on variable spreading gain (VSG) is studied in cellular CDMA network in presence of soft handoff (HO). The processing gain is varied according to traffic intensity meeting a requirement on data bit error rate (BER). The overall performance improvement due to processing gain adaptation and soft HO is evaluated and compared with a fixed rate system. The influence of soft HO parameters on rate adaptation and throughput and delay performance of data is indicated. Further truncated automatic repeat request (T-ARQ) is used in link layer to improve the performance of delay sensitive services. The joint impact of VSG based transmission in presence of soft handoff at physical layer and T-ARQ at link layer is evaluated. A variable packet size scheme is also studied to meet a constraint on packet loss.

Keywords: Adaptive Systems, Variable Spreading Gain, Soft Handoff, ARQ

1. Introduction

Multimedia services are becoming increasingly important in wireless networks. The demand for high rate packet data transmission and quality of services (QoS) in wireless networks is growing at a rapid pace. Code Division Multiple Access (CDMA) is very promising to meet the demand for high data rate and quality of service (QoS) in wireless networks. The cellular capacity of CDMA system is limited in the uplink by maximum tolerable interference at the base station (BS). The cellular capacity indicates that there is a practical number of admissible users which should not be exceeded in order to ensure QoS (quality of service) of the admitted users. However for a fixed data rate system, there will be redundant margin on system capacity when traffic level is much lower than the maximum allowed number. Adaptive transmission schemes utilize the system resource more efficiently over a fixed rate scheme where the data transmission rate is controlled depending on generated interference or channel traffic intensity [1−3]. Packet data transmission is gaining importance in CDMA networks. Variable processing gain (VSG) and multi-codes (MC) are two interesting approaches for increasing transmission rate of data in CDMA [4].

CDMA uses soft handoff (HO) where the handoff mobile near a cell boundary transmits to and receives from two or more BS-s simultaneously [5]. Soft HO provides a seamless connectivity, reduces “ping-pong” effect as present in hard HO, lowers probability of lost calls and eases power control [5,6]. It extends the coverage and increases the reverse link capacity [6,7] by reducing overall interference.

Several research papers have analyzed adaptive transmissions based on VSG [1−4] without considering the effects of soft HO. Since soft HO affects the generated interference, it is expected to have significant impact on spreading gain selection and successful transmission of packet data, which is considered in the present paper.

However performance of data services is limited by interference and channel fading. Adaptive modulation and coding (AMC), adaptive antenna array providing space diversity and several receiver algorithms at physical layer [8,9] are used to enhance the throughput performance of packet data oriented systems. Alternately channel fading can be mitigated by automatic repeat request (ARQ) protocol at the data link layer which ensures persistent retransmission of packets associated with a particular message until it is received correctly. ARQ is effective in improving system throughput relative to only forward error correction (FEC) [9]. Further ARQ can be combined with FEC in a hybrid ARQ scheme. To minimize delay and buffer sizes in practice, truncated ARQ has been adapted to limit the number of retransmissions. Further the transmission of video/image requires the delay to be bounded i.e. a packet is to be dropped if it is not received correctly after a finite number of retransmissions [8].

Several research papers have studied the combined effects of physical layer issues with the link layer issues like ARQ [8,9]. However the issue of soft handoff at physical layer is not considered in [8,9].

In the present paper we consider VSG based data transmission meeting a constraint on upper limit of BER in presence of soft handoff. A simulation study is carried out to evaluate performance of packet data in terms of throughput and delay considering the joint effects of VSG and soft handoff. Two cases of retransmissions namely infinite ARQ and truncated ARQ at link layer are also considered. First the performance of data has been simulated in VSG with soft HO considering infinite ARQ. Next T-ARQ is considered at link layer for real time services along with soft HO at physical layer. Joint effects of soft HO and truncated ARQ on throughput and delay performance of a packetized data are evaluated for an imperfect power control CDMA. The performance in each case is compared with fixed rate. Effects of soft HO parameters on spreading gain adaptation and data performance in terms of throughput, delay and packet loss associated with truncation are indicated. Further a variable size packet transmission is also considered to meet a constraint on packet loss rate. Thus it has been possible to satisfy the BER constraint, delay constraint and the packet level QoS such as packet loss rate as well.

Sections 2 and 3 briefly describes the cellular scenario and our simulation model. Results and discussions are presented in Section 4. Finally we conclude in Section 5.

2. System Model

A cluster of three sectored cells with uniformly distributed mobile data users (MS) and equal number of MS-s (N_d) per sector are considered. All data users transmit at the same rate using a single code. For fixed rate system the user transmits on single code at a fixed rate R_b while in adaptive system the transmission rate $R_b(l)$ is variable depending on traffic load (l) which is Poisson distributed with mean (λ). The processing gain (pg) of all codes are equal; where $pg(l)$ =W/ $R_b(l)$; W is spread bandwidth. Processing gain $pg(l)$ (hence data rate $R_b(l)$) is selected depending on traffic load (l) satisfying a BER criterion. A “continuously active” data traffic model as in [10] is considered where each user generates a sequence of fixed length packets. A new packet is generated as soon as the preceding packet is either delivered successfully or dropped due to truncation in ARQ. The soft HO region is defined based on the distance from the base station (BS) as in Figure 1. An MS located outside the handoff boundary R_h is considered to be under soft HO with three neighboring BS-s. Each sector is divided into two regions, soft HO regions (B, C, D) and non-HO region (A, E, F) of cell #0,1and 2 respectively in Figure 1. BS$_0$, BS$_1$ and BS$_2$ are the BS-s of cell #0, 1 and 2 respectively. The propagation radio channel is modeled as in [7]. The link gain for a location

$$(r,\theta) \text{ is } \; G_i(r,\theta) = d_i(r,\theta)^{-\alpha_p} 10^{\xi_S/10} \tag{1}$$

where $d_i(r,\theta)$ is the distance between the MS and BS_i , α_p is the path loss exponent and $10^{\xi_s/10}$ is the log-normal component with ξ_s normally distributed with 0 mean and variance σ_s^2. The shadow fading at i-th BS is [7]

$$\xi_{s-i} = a\zeta + b\zeta_i \quad \text{with} \quad a^2 + b^2 = 1 \tag{2}$$

ζ and ζ_i are independent Gaussian random variables with zero mean and variance σ_s^2 . Out-cell interference consists of interference due to MS-s from region (E,C,G,H) of cell #1 and (D,F,I,J) of cell #2. MS-s in furthest sectors (G,H,I,J) are assumed to be power controlled by respective BS-s. The reference user is located in non-HO region of reference sector i.e. in region ‘A’. Total in-cell interference in cell # 0 is

$$I_{in} = I_1 + I_2 \tag{3}$$

where I_1 is due to all MS-s in A and those in B connected to BS$_0$, I_2 is due to MS-s in B but connected to BS$_1$ and BS$_2$. The out-cell interference is

$$I_{out} = 2(I_E + I_{c1} + I_{c2} + I_{co} + I_G + I_H) \tag{4}$$

I_E is the interference due to MS-s in E and connected to BS$_1$. Similarly I_{c1} and I_{c2} are due to MS-s in region C and power controlled by BS$_1$ and BS$_2$ respectively. I_{co} is due to MS-s in C and controlled by BS$_0$. I_G and I_H are the interference due to MS-s in G and H. MS-s in these farthest sectors are assumed to be power controlled by respective BS i.e. BS$_1$. A multiplication factor of two is used in Equation (4) to include contribution of cell #2. The actual received power from desired user is $U = S_R e^S$, where S is a Gaussian r.v. with mean 0 and variance $\sigma_S^2 = \sigma_e^2$. The BER (P_e) for data user is simulated as described in later section in the above soft HO environment considering direct sequence spreading and BPSK data modulation having spread b.w of W. The maximum allowed bit rate of data users $R_b^*(l)$ for traffic intensity is adjusted such that ($P_e \le \beta$), the corresponding processing gain is selected as $pg^*(l)$. The retransmission probability P_r is given as [11]

$$P_r = 1 - (1 - P_e)^{L_p r_c} \qquad \textbf{(5)}$$

where L_p is the length of the packet in bits and r_C is the FEC code rate. For continuously active data users, the average packet delay is the same as the packet transfer time T_p as there is no waiting delay in the queue. The time required for transmitting a packet of length L_p by a data user transmitting at a rate of $R_b(l)$ is :

$$T_p = \frac{L_p}{R_b(l)} = \frac{L_p \quad pg^*(l)}{R_c} \qquad \textbf{(6)}$$

where R_c is the chip rate. We assume that acknowledgement from the receiver is instantaneous and perfectly reliable. In case of truncated ARQ the maximum number of retransmissions in ARQ has to be bounded since only finite delay and buffer sizes can be afforded in practice. If $T_{\max}$ is the maximum allowed packet delay, the maximum number of retransmissions is given as $N_{\max} = \lfloor T_{\max} / T_p \rfloor$. Thus with truncated ARQ if a packet is not received correctly after $N_{\max}$ retransmissions, it is dropped and declared as a packet loss. The average delay with truncation

$$D = \frac{L_p . pg^*(l)}{R_c} \left\{ \frac{1 - P_r^{(N_t+1)}}{1 - P_r} \right\} \qquad \textbf{(7)}$$

The average throughput is defined as the average number of information bits successfully transferred per sec and is given as

$$G = \frac{L_p . r_c}{D} = \frac{r_c . R_c (1 - P_r)}{pg^*(l) \; (1 - P_r^{(N_t+1)})} \qquad \textbf{(8)}$$

Some services may require maintaining packet loss (packet QoS) below a prescribed limit (δ). A variable packet length scheme is used where the packet size is adjusted under different traffic and soft HO conditions so as to maintain packet loss below a desired limit (δ). The length of the packet (L_p^*) is selected satisfying $P_{loss} \le \delta$:

$$\text{i.e. } \{1 - (1 - P_e)^{L_p^* r_c}\}^{N_t^* + 1} \le \delta \,, \quad N_t^* = \lfloor T_{\max} / T_p^* \rfloor \qquad \textbf{(9)}$$

$$\text{where } T_p^* = L_p^* / R_b \qquad \textbf{(10)}$$

The packet size L_p^* is found by simultaneously satisfying (9) and (10).

However with infinite retransmission i.e. infinite ARQ, there is no packet loss. In such situation the average delay [11]

$$\text{D} = \frac{T_i}{(1 - P_r)} = \frac{L_p \; pg^*(l)}{R_c (1 - P_r)} \qquad \textbf{(11)}$$

and the average throughput (G) given as

$$G = \frac{L_p r_c}{D} = \frac{r_c R_c (1 - P_r)}{pg^*(l)} \qquad \textbf{(12)}$$

In the next section we present our simulation model for both infinite and truncated ARQ cases.

3. Simulation Model

The simulation is developed in MATLAB using the following parameters: PR_h indicates the degree of soft HO, shadowing correlation (a^2), pce σ_e, traffic intensity λ. he soft HO region boundary R_h given as $R_h = R_o \sqrt{1 - PR_h}$ where R_o is the radius of the cell, normalized to unity and hexagonal cell is approximated by a circular one with radius R_o. Users are assumed to be uniformly distributed.

3.1. Generation of Users Location and Interference

1) The number of users (N_d) is generated by generating a Poisson distributed r.v with mean λ.

2) Locations (r, θ) of all (N_d) users are generated and users are divided into non-HO (N_h) and soft HO (N_s) region based on their location. Assuming the desired user in non-HO region, let the remaining interfering users in non-HO are (N_h -1). Number of users in soft HO region: $N_s = N_d - N_h$.

3) For each of those in soft HO region (N_s), the link gains corresponding to each of three BS-s involved in soft HO are generated as

$G_i(r, \theta) = r_i^{-\alpha_p} e^{\xi_i}$, i =0,1, 2. where ξ_i is a Gaussian r.v with mean 0 and variance $b^2 \sigma_s^{\,2}$, r_i is the distance from i-th BS. The user is power controlled by the BS for which the link gain is maximum i.e. it is power controlled by BS_i if G_i is maximum; $i = 0,1,2$.

4) The interference received at reference BS

$$I= S_R \exp(r_n)\ (\frac{G_0}{G_i}) \qquad \textbf{(13)}$$

if connected to BS$_i$, where $i = 0, 1, 2$. Here r_n is a normal r.v. with 0 mean and standard deviation σ_e. S_R is the required received power at the respective BS which is normalized to unity in the simulation since SIR is unaffected by assigning S_R =1.

5) Next interference due to (N_h-1) MS-s in non_HO region (A) of reference cell, each power controlled by BS$_0$ is considered as

$$I_2 = S_R \sum_{i=1}^{N_h-1} e^{r_{n,i}} \qquad \textbf{(14)}$$

6) Now the interference due to MS-s in adjacent sectors i.e. (region E,C,D and F) of cell#1 and #2 are found in similar manner. The number of MS-s in E and F are $(N_d - N_s)$ each. Let $I_3 = I_E + I_C$ and $I_4 = I_D + I_F$

7) Interference from MS-s in G,H, I and J regions are generated following step 4. Let $I_5 = I_G + I_H$ and $I_6 = I_I + I_J$.

8)Total interference I=$\sum_{k=1}^{6} I_k$ **(15)**

9) Signal from desired user

$$U = S_d e^x\,, SIR = U / I \qquad \textbf{(16)}$$

x is Gaussian with mean ‘0’ and variance σ_e^2

3.2. BER Simulation

A Gaussian noise sample n_g with ‘0’ mean and variance $\sigma_g^2 = 1./(2\,pg.\,SIR)$ is added to each bit of a transmitted sequence and received bits are compared with the transmitted bits. Here SIR is found following steps A(1) to A(8) for a given pg .

3.3. Selection of Processing Gain

1. An initial low value of R_b is chosen and BER is simulated as described above.

2. R_b is incremented in steps, the highest value of R_b^* for which $BER \le \delta$, is chosen and corresponding processing gain $pg^* = \left\lceil W\!/\!R_b^* \right\rceil$ is selected.

3.4. Packet Loss and Variable Packet Length

1) For truncated ARQ case:

A sample of Gaussian noise as in (B) is added to each transmitted bits of a packet of $L(= L_p r_c)$ information bits. The received L bits of a packet are checked with their corresponding transmitted bits to assess packet error. If the received packet is incorrect, the same packet (i.e. the same bit pattern) is re-transmitted N_t times, where N_t = 1,2,… $N_{\max}$. A packet loss occurs if it is not received correctly after N_t re-transmissions. Average delay (D) is estimated as: $((N_p + retx\_count)./N_p)T_p$ where T_p is as in (6), N_p : number of transmitted packets, $retx\_count$: total retransmissions of N_p packets (each packet is retransmitted maximum up to $N_{\max}$ time.

The throughput is: $G = L_p r_c / D$ **(17)**

An initial small packet size is chosen and it is incremented in steps (Δ=2) till packet loss just exceeds δ . The highest packet length for which $P_{loss} < \delta$ is L_p^*. Now throughput, delay and packet loss are estimated by simulation following steps as mentioned above where pg is chosen as $pg^*(l)$ and L_p is chosen as L_p^* .

2) For infinite ARQ case:

If the received packet is incorrect, the same packet is retransmitted until the packet is finally received correctly instead of limiting the retransmission to $N_{\max}$ as in case of T-ARQ. Then delay and throughput in this case are estimated in the same manner as described above. Total number of erroneous packet is counted out of a large number of transmitted packets to estimate the packet error rate (PER).

4. Results and Discussions

The parameters assumed in simulation are listed in Table 1 as shown. For the case of truncation we assume $T_{\max}$ =350 msec and $N_t = \lceil N_{\max}/2 \rceil$ where $N_{\max}$ is maximum retransmission corresponding to $T_{\max}$.

Table 1. Parameters for simulation.

Spread BW (W)	Chip rate R_c	Fixed PG	Fixed L_p	SIR threshold γ_{th}	pce σ_e (dB)
5.0 MHz	5.0 Mcps	312	1024	7 dB	1 and 2
Path loss α_p	σ_s (dB)	r_c	Shadowing correlation (a^2)	Degree of soft HO PR_h	Max allowed packet loss
4	6	0.5	0 and 0.3	0.3, 0.7	5%
Target BER 10^{-3}	$T_{\max}$ 350msec				

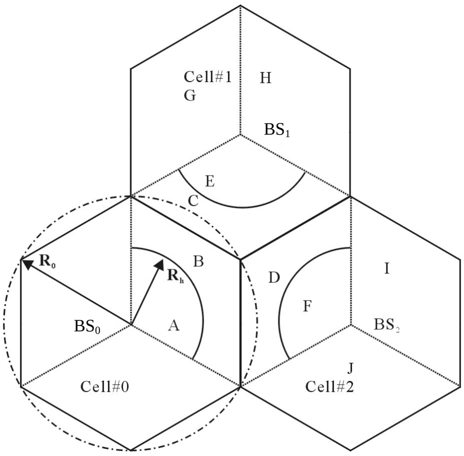


Figure 1. Cellular Layout for soft HO. A, E, F are non HO region. B, C, D are soft HO region. Cell # 0 is reference cell.

Figure 2 shows the effects of soft HO on BER performance of fixed rate and adaptive transmission. In adaptive transmission, the BER is always maintained as $P_e \leq 10^{-3}$ for the entire range of traffic by adjusting the pg i.e. transmission rate is reduced with increase in traffic as in curves (iv,v). While in fixed rate (PG=312), the BER increases with increase in traffic intensity due to increased interference as in curves (i, ii and iii). Higher degree of soft HO reduces BER in fixed pg case as seen in curves (i) and (ii) due to reduced level of interference. In case of adaptive transmission, as the transmission rate is adjusted (varying the pg) keeping BER fixed, a higher degree of soft HO will lead to increase in data rate i.e. lower processing gain allocation for same BER constraint of $P_e \leq 10^{-3}$. Two cases of adaptive pg with different levels of soft HO have been shown in curves (iv, v) where $\text{BER} \leq 10^{-3}$ in both the cases. However the data rate (allocated pg) for these cases will be significantly different as seen in next fig.

Figure 3 shows the effects of soft HO on allocation of processing gains in VSG based adaptive transmissions. The processing gains (pg) are chosen satisfying BER $\leq 10^{-3}$ under different conditions of soft HO. As the traffic intensity (λ) increases, data rate is reduced by increasing pg so as to reduce the interference level for maintaining BER constraint. Higher degree of soft HO reduces BER for fixed data rate. As BER is kept fixed, higher degree of soft HO will allow higher data rate or lower value of allocated pg in curves (ii and iii). Thus it is seen that as PR_h increases from 0.3 to 0.7, allocated pg for $\text{BER} \leq 10^{-3}$ reduces from 339 to 260 at λ=8. Higher shadowing correlation (a^2) as well as lower pce also reduces BER for fixed PG. Thus in a similar manner as in case of PR_h, this will also increase data rate (or reduce processing gain) for a target BER as seen in curves (i,iii) and curves (iii,iv).

The throughput performance with T-ARQ is depicted in Figure 4. Adaptive pg based transmission is seen to achieve a higher throughput as compared to a fixed one for most range of traffic in curves (ii,iii). This is because BER is maintained below a limit in adaptive case whereas it increases with traffic in case of fixed pg. Further improvement in throughput is achieved with higher degree of soft handoff in curves (i,ii). However as seen in Figure 4 vide curves (ii,iii), over a range of traffic say up to λ =7, adaptive pg and fixed pg yields close throughput performance. This is because our arbitrary chosen value of fixed pg i.e. PG=312 is close to adaptive pg-s found (satisfying $P_e \leq 10^{-3}$) over this traffic range. For example at λ =7, allocated adaptive pg =304.

Effects of soft HO parameters on packet delay are depicted in Figure 5. Truncation always maintains packet delay below a certain level, here 0.5 times of $T_{\max}$ = 350 msec, as number of retransmission $N_t = \lceil N_{\max}/2 \rceil$. Adaptive transmission yields less delay as compared to fixed case for most range of traffic. Adaptive transmission always maintains a fixed BER hence a fixed level of PER (packet error rate) while BER increases with traffic in case of fixed pg. However around $\lambda = 7$ the performance of fixed and adaptive are close as our chosen fixed pg of 312 becomes close to found value of adaptive pg. Higher degree of soft handoff is found to lower delay further as seen in curves (ii,iii).

Figure 6 shows the packet loss associated with T-ARQ vs traffic intensity for fixed and adaptive pg based transmissions. Adaptive pg always maintains a fixed level of packet error. Thus adaptive pg yields lower packet loss as compared to fixed rate for moderate traffic in curves (i,ii). The packet loss reaches a floor for adaptive case while it increases with higher traffic in case of fixed rate. Further the packet loss is kept below $\delta = 5\%$ using variable packet size as seen in curve (iv). Higher soft handoff reduces packet loss as in curves (ii, iii).

Figure 7 shows the variation of packet size with traffic intensity. The packet size is varied in order to ensure that packet loss is always $\leq 5\%$. Higher traffic intensity requires smaller size of packet. Further higher degree of soft handoff (PR_h) can transmit larger packet as seen in curves (i,iii). Higher shadowing correlation also allows larger packet while meeting the loss constraint as in curves (i,ii).

Table 2 shows the maximum number of retransmissions allowed vs traffic under two different soft handoff conditions. As the traffic intensity increases, the allo-

Table 2. Maximum number of retransmissions vs traffic (mean arrival rate) for adaptive pg based transmissions.

λ	5	6	7	8	9	10	11	12
(1) $N_{\max}$	6	5	4	4	3	3	2	2
(2) $N_{\max}$	8	7	6	5	4	4	4	3

(1) $a^2=0.3, PR_h=0.3, pce=2dB$ $T_{\max}$=350 msec

(2) $a^2=0.3, PR_h=0.7, pce=2dB$ $T_{\max}$=350 msec

cated data rate is reduced (i.e. allocated pg increases). Thus the packet transfer time is increased which in turn reduces the maximum number of retransmissions for a prescribed maximum allowed delay. Higher soft HO allows more number of retransmissions.

Figure 8 shows the effects of truncated ARQ on delay. We have chosen truncation in retransmission as $N_t = \lceil N_{\max}/2 \rceil$ to ensure a maximum delay of 0.5 times of $T_{\max}$=350msec i.e. maximum allowed delay of 175 msec in this case. It is seen that in case of infinite retransmission, the delay increases rapidly with traffic intensity. T-ARQ always yields lower delay as compared to infinite ARQ (curves ii,iii). Further in case of T-ARQ, though delay increases with traffic, it is always maintained below a chosen desired limit of 175 msec (i.e. 0.5

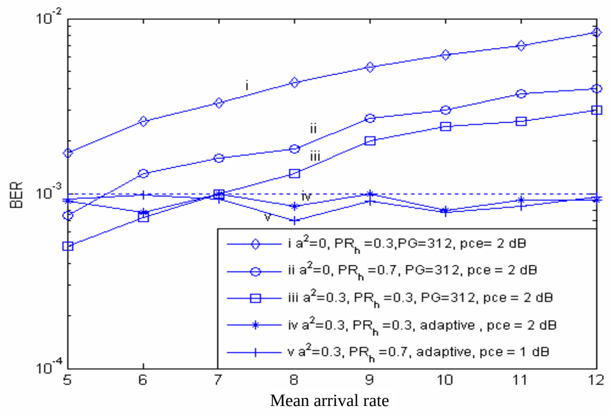


Figure 2. BER vs mean arrival rate (λ users/sec) per sector.

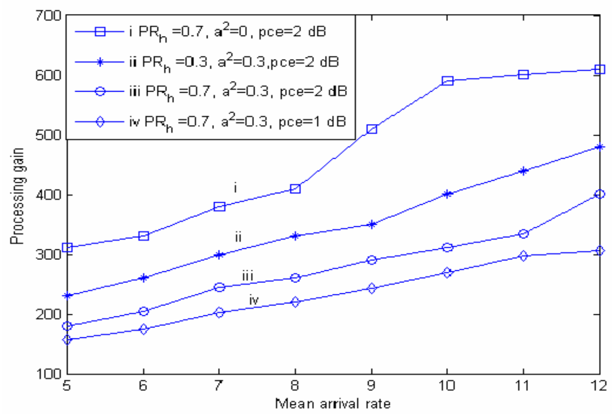


Figure 3. Processing gain vs mean arrival rate (λ users/sec) per sector.

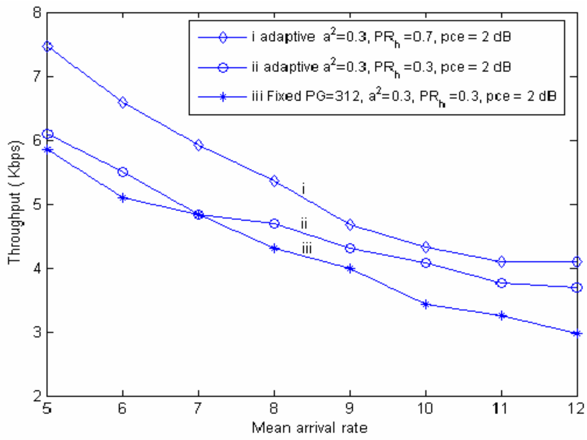


Figure 4. Throughput (Kbps) vs mean arrival rate.

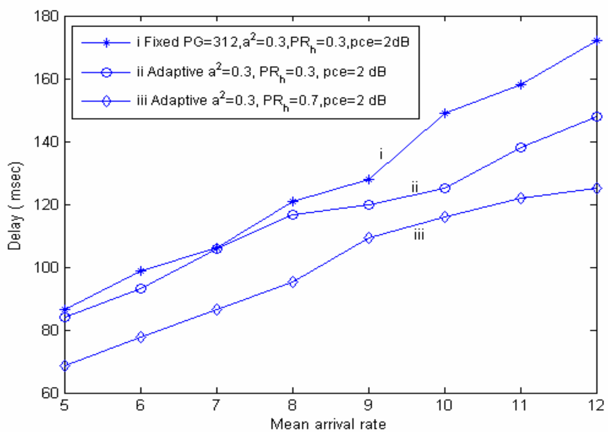


Figure 5. Delay (msec) vs mean arrival rate.

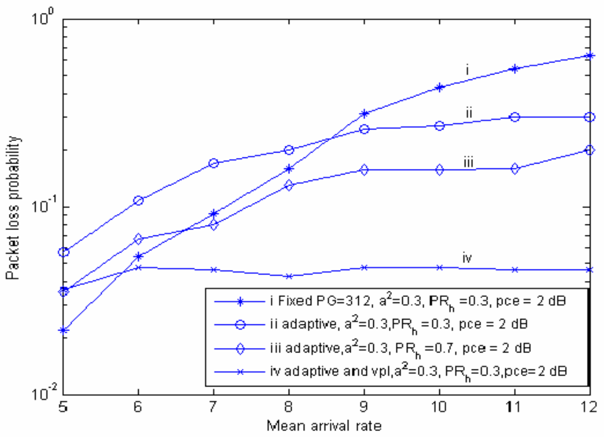


Figure 6. Packet loss probability vs mean arrival rate.

$T_{\max}$). With infinite retransmission the delay exceeds desired limit of 175msec for λ =10 onwards in case of a^2=0.3 and λ =6 onwards for a^2=0 in curves (i,ii). Thus T-ARQ satisfies delay constraint at the cost of packet loss. Further higher shadowing correlation lowers the delay in curves (i,ii).

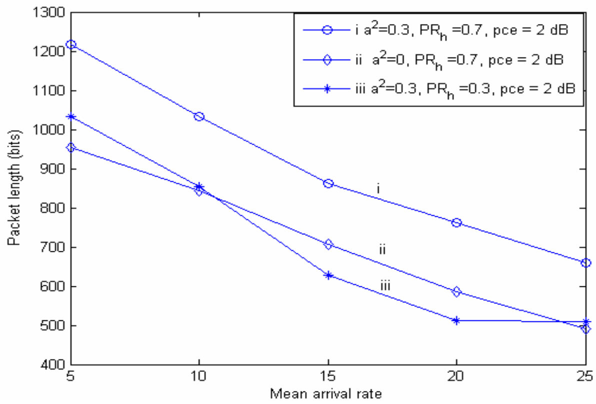


Figure 7. Effects of soft HO and shadowing correlation on packet length for adaptive pg based Transmissions.

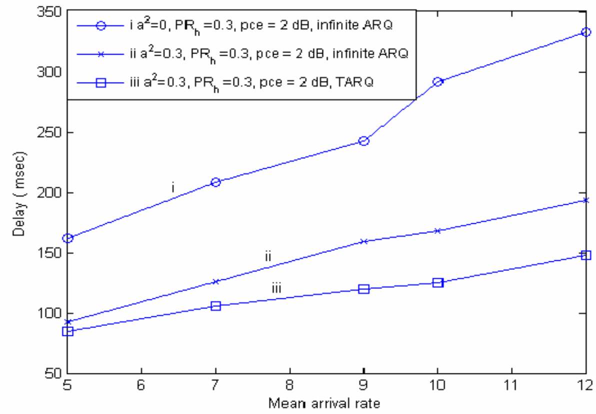


Figure 8. Effects of truncation in ARQ on delay.

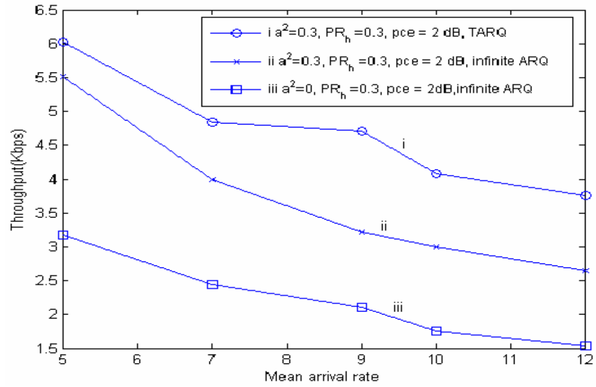


Figure 9. Effects of truncation in ARQ on throughput.

Figure 9 shows the effects of truncated ARQ and shadowing correlation on throughput. In present work throughput indicates the average number of information bits successfully transferred per sec. Since a packet is delivered successfully with in a given time limit in truncation (or dropped if not successful), truncation yields higher throughput as compared to infinite retransmissions as in curves (i,ii). Also higher shadowing correlation improves throughput as in curves (ii,iii).

5. Conclusions

Performance of packet data is evaluated in presence of soft handoff under adaptive transmission based on variable processing gain and is compared with a fixed rate system. The processing gain is varied according to traffic intensity so as to ensure an upper limit on channel BER. Truncated ARQ is used to ensure a prescribed maximum limit on packet delay. Soft handoff parameters, shadowing correlation, pce and T-ARQ are found to have significant impact on processing gain allocation and data performance. Adaptive transmission outperforms fixed transmission in terms of throughput, delay and packet loss. Adaptive transmission enhances throughput by 19% and reduces packet loss by 16% for a traffic intensity of $\lambda = 10$ users/sec under a given soft handoff scenario. Higher degree of soft handoff and higher shadowing correlation improve the situation further. An increase in degree of soft handoff by 133% further enhances throughput by 6% and reduces packet loss by 12 % for a traffic intensity of $\lambda = 10$ users/sec. Using a variable packet length transmission the packet loss could be maintained below a limit. Thus data BER, delay constraint and packet QoS such as packet loss are simultaneously satisfied using variable processing gain, variable packet length and T-ARQ. Higher degree of soft handoff and higher shadowing correlation allow transmission of larger packet size. An increase in degree of soft handoff by 133% allows 20% increase in packet size for a traffic of $\lambda = 10$ users/sec satisfying a packet loss constraint.

6. References

[1] K. Choi, S. Kim, Shin, and K. Cheun, "Adaptive processing gain CDMA networks over poison traffic channel," IEEE Communications Letters, Vol. 6, No. 7, pp. 273−75, July 2002.

[2] K. Choi, Y. Chae, and J. Park, "Throughput-delay performance of interference level adaptive transmission in voice/data integrated CDMA network with variable spreading gain," IEE Proceedings on Communications, Vol. 151, No. 3, pp. 217−220, June 2004.

[3] L. L. Yang and L. Hanzo, “Adaptive rate DS-CDMA systems using variable spreading factors,” IEEE Transactions on Vehicular Technology, Vol. 53, No. 1, pp. 72–81, January 2004.

[4] S. Kumar and S. Nanda, “High data rate packet communication for cellular network using CDMA: Algorithms and performance,” IEEE Journal on Selected Areas in Communications, Vol. 17, No. 3, pp. 472–491, March 1999.

[5] D. Wong and T. Lim, “Soft handoff in CDMA mobile system,” IEEE Personal Communications, pp. 6–17, December 1997.

[6] H. Jiang and C. H. Davis, “Coverage expansion and capacity improvement from soft handoff for cellular CDMA,” IEEE Transactions on Wireless Communications, Vol. 4, No. 5, pp. 2163–2171, September 2005.

[7] J. Y. Kim and G. L. Stuber, “CDMA soft HO analysis in the presence of power control error and shadowing correlation,” IEEE Trans on wireless Communications, Vol. 1, No. 2, pp. 245–255, April 2002.

[8] Q. Liu, S. Zhou, and G. B. Giannakis, “Cross layer combining of adaptive modulation and coding with truncated ARQ over wireless links,” IEEE Transactions on Wireless Communications, Vol. 3, No. 5, pp. 1746–1755, September 2004.

[9] B. Lu, X. Wang, and J. Zhang, “Throughput of CDMA data networks with multi-user detection, ARQ and packet combining,” IEEE Transactions on Wireless Communications, Vol. 3, No. 5, pp. 1576–1589, September 2004.

[10] J. Kim and M. Honig ,“Resource allocation for multiple class of DS-CDMA traffic,” IEEE Transactions on Vehicular Technology, Vo. l49, No. 2, pp. 506–518, March 2000.

[11] S. Kundu and S. Chakrabarti, “Performance of high rate data in wideband CDMA with correlated interferers,” in GESTS International Transactions on Communication & Signal Processing, Vol. 7, No. 1, pp. 53–64, June 2006.

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

Ubiquitous Media with UPnP and RFID-Based User Interfaces

Gerrit KALKBRENNER
Technische Universität Dortmund, Embedded Systems, Otto Hahn Str. 16, *Dortmund, Germany*
E-mail: *gerrit.kalkbrenner@udo.edu*

Received November 3, 2008; *revised March* 1, 2009; *accepted March* 3, 2009

Abstract

The evolution of systems and networks, including PDA, handhelds, mobile phones, WLAN, and Bluetooth provides us new scenarios for media presentation. Because of the growing number of such (personal) devices in the Personal Area Network of the User it is necessary to set up a system, in which the user doesn’t lose control over the media and their corresponding presentation devices. Digitalisation will lead us to a split of into content (music, video), storage (e.g. compact discs, server), and user interfaces (receiver). Media will no longer be stored in shelves at home, but in storage spaces located somewhere in the network. New user interfaces like PDA and mobile phones will replace the panel field of an old fashion CD-Player and amplifier. A protocol is required for synchronisation and controlling this media scenario, which is UPnP (Universal Plug and Play).

This paper describes a scenario based on UPnP and its implementation provided by the author.

Keywords: Ubiquitous Media, UPnP, RFID

1. Introduction, Ubiquitous Media Scenario

In a near future media (Music, Video) will be bought in a spontaneous way, probably by staying in front of a music advertisement table in the Metro or somewhere on the street [1,2]. Using our PDA or mobile phone, we just push a button [3]. With location-based service [4] the system knows our place and-based on that information-the system knows the music title on the advertisement table. The music, bought during this procedure, will be transferred via Internet/DSL to our home server or to a central server. This server replaces data stores like CD's or DVD's. Using this server the content will be published in a way that we can access it in a ubiquitous manner. The corresponding cover/booklet will be sent later by letter. The cover is still important, as the human is a haptic oriented being, which likes it to hold something in his hands. Beside this, the cover/booklet contain some important metadata regarding the music/video.

To play the music in a former time we used a CD player, an amplifier and a set of speaker. Those components also serve as a user interface to the media, beside their main electrical function. Therefore, the panel of the CD player builds the user interface for music presentation control. Using different equipment, we get additional user interfaces for the same content. In order to simplify user interfaces other devices like mobile phones and PDA will come into the scenario. [5,6] By pressing a button on my personal PDA, music will be presented on the local speaker, regardless of the given equipment. By moving from the living room to the kitchen, the music will follow with me to the speaker system, which is e.g. installed in the kitchen.

Within this scenario, media content can be distinguished in to 4 components: the **1) media data,** (pcm-files, avi-files) they are stored on a CD, DVD, USB stick, or on a server hard disc. On the other side we have **2) public meta data**, which gives information about the title, author and other, the **3) personal meta data**, which describes the actual storage places, my rights on the media (see DRM) and other, and finally some **4) user interfaces**, which allows the control of the presentation of music and video. Beside this, there is the presentation equipment (speaker, projector) and the presentation event itself).

The central question during the design of a ubiquitous media system is a protocol, which allows the set up, and the dynamic control, e.g. the usage of different user interfaces within a ubiquitous media system.

In this context, the standard UPnP (Universal Plug and Play) [7] got some relevance. This paper gives a practical perspective of a working scenario of ubiquitous media using UPnP.

2. Problems

Due to the emerging technology development of wireless networks [8] and hardware for multimedia presentations, applications will change rapidly [9]. Because of the growing number of such (personal) devices in the personal area Network (PAN) of the user, it is necessary to develop a system, in which the user does not lose control over his devices [10]. Digitalisation will lead us to a splitting of content and storage devices (e.g. compact discs, hard disks, flash memory). It will no longer be stored in shelves at home, but in storage spaces located anywhere in the network [11]. Therefore, the storage space available to the user will increase. At the same time, the overview to one's own content will decrease. Upcoming media systems have to grant transparency to the user. The user wants to feel that he knows where his content is located and that he can reach it easily every time, everywhere. Otherwise, users would feel overwhelmed and uncomfortable with those devices. The present situation on the multimedia-market is very much like that.

To avoid these effects a new media system is required, in which the user is the focus of the development. Even though he can reach a much larger amount of content within his system and uses more different devices, he feels comfortable with this system and he has full control over it. This can be achieved by integrating new interaction techniques (as speech-control and gesture recognition in combination with improved graphic user interfaces) in one universal interface.

For unknown reasons, the concepts of **Context Awareness** [12,13] and **Ubiquitous Computing** [14,15] are not yet explored for their use in media systems [16]. Both concepts provide us a media system, in which the user is confronted neither with difficult user's manuals nor with changing interfaces.

The actual development of computer indicates some tendencies: devices will decrease in size and the amount of devices available for each user will increase. With wireless networks user will have access to more content. In this paper we develop a concept, within a user can use his media in a comfortable way. Main concepts are abstraction, transparency, and awareness. We can abstract from the physical storage just by simulating them on a known location. First, we will discuss some under laying concepts.

3. Foundation Concepts, Ubiquitous Media

The markets for music, video, and games are growing together. Where before we had different markets for each type of media, in the near feature there will be only one, but with much more participants. The market for computer games now days have a larger volume than the market for movies. This is technically enabled by the increasing power of available systems, which today offers more CPU-power and memory and provide multimedia capabilities. In addition, the size of devices is reduced [17].

An implementation of ubiquitous media requires a concept of media transparency. It allows us to access media regardless from its format and its storage place. Media transparency is the main concept for ubiquitous media, so we should take a closer look of it details.

For media transparency, we need:

- Access Transparency-user can access media, regardless from the presentation device, underlying network, and media server.
- Location Transparency-user can access media regardless from its storage place.
- Format Transparency-media content is coded in different formats (mp3, PCM). With format transparency, an abstraction of the format can be achieved. Media player are mainly multi format capable. Non-fitting formats are converted on the fly by other system components.
- DRM Transparency-witch rights a user has on media content is only clearly visible during the buying process. Later the user wants to access media according the rights and he is not willing to engage with DRM aspects.

The user has to be placed into the middle of a ubiquitous media scenario. Around him, we build up a media system, which:

1) can be used in a simple way in every usage situation,

2) is working in the background in a unobtrusive way

3) adapts to the necessities of a user

In the next chapters, we will build up a ubiquitous media system.

4. Requirement Analysis

Based on the ubiquitous media idea we take a deeper look about the usage scenario and its changes. Based on this we set up a concept of media transparency. Based on this a concept for the dynamic integration of new user interfaces is required. It provides a comfortable access to future media systems; even they provide increased functionality, more presentation systems and a larger amount of media content. The quality of user interfaces can be increased significantly by considering context awareness. It enables the media system to react on the situation, in which the user currently is involved. In this way, the system can be controlled in an intuitive way.

Critical is the protection of personal information. In order of building a transparent and adaptive system, it needs to collect and store information about the user.

5. System Design

The System can be successfully built when separating some aspects, which traditionally are combined tightly. These aspects are 1) user interfaces and system control, 2) media data storage and access, and 3) system management.

By separating and hiding some of these aspects, a system can be build, which partially disappears from the sight of the user. He can concentrate again on his main aim, the consummation of media content.

The system will be build up by the following components:

5.1. Media Server

The assignment of the media server will be the storage and delivery of media content. There might be several servers in a system, which holds all content data: mp3-, pcm-, avi-, vob-, and other media files. Special server might be capable to convert media formats into each other. A server might be build up by regular pc hardware, network attached hard drives (NAS), CD/DVD changer, or hard disk based recorder/receiver. Renderer/ presentation systems will access media content via networks (e.g. Ethernet, WLAN, and Bluetooth) and selected protocols (e.g. NFS, SMA, and RTP).

5.2. Renderer (presentation environment)

The main task of a rendering system is the access of content from media server and its preparation for presentation. Audio content might be presented via a speaker system, video content via a television set or by a projection system (beamer). Renderer can be build up from regular PC hardware or by special equipment, which fit into the apartment and living ambient. There are special components available, which are build up by a LINUX system and supports the sound and vision platform. These Systems are highly extendable.

5.3. Control Points

User interfaces, which are used to control the ubiquitous media system, mainly are built as control points. A control point enables the user to adjust main system parameter like selected media, play, pause, stop, presentation device, and other. Control points can be built up from regular PC hardware, remote control, PDA, mobile phones or other small devices. However, a control point does not have necessarily a user interface. The phone during a phone call itself might set the music to pause. A control point reacts to user and other events, and in result, it controls some of the system parameter. Figure 1 shows the system components.

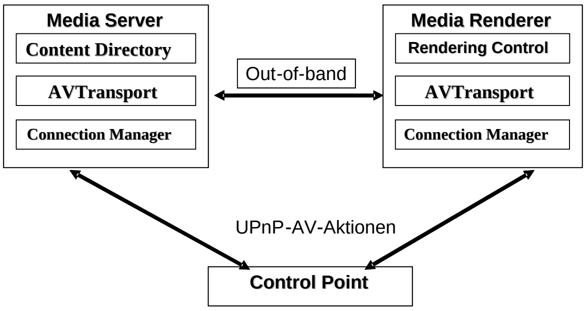


Figure 1. Selecting and controlling media presentation.

6. System Implementation

Main components might be implemented in Java, in order of portability. Protocols are implemented using UPnP (Universal Plug and Play) which is based on SOAP, XML and HTTP.

UPnP enables control points to detect server and renderer in a dynamic way. Commands like “music pause”− originated on a PDA-are transferred via UPnP to the renderer, which on the other hand notifies the server via UPnP protocol to stop the streaming of content.

6.1. Detection of System Components

Control points have to detect server and renderer components, in order to check their facilities and to transmit commands to them.

Based on networks like Bluetooth or WLAN a IP connection will be established. The aim of a DHCP server is to assign IP addresses. As part of the UPnP protocol, a detection request is sent by broadcast. Renderer and server will send back their profiles and tables of offered functions.

UPnP offers an abstraction from IP address to service classes and names, so control points can deal with them on a high-level abstraction.

6.2 Choosing Media

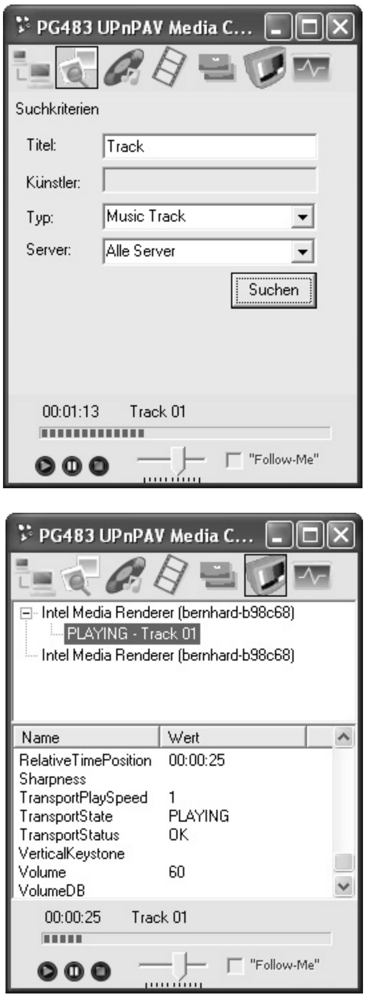
The user wants to select media from the offer of his server (single system or a collection). Also bought music (please compare the purchasing scenario from the beginning of this paper) will be placed on this server. A search request is sent to al corresponding server. From the result the user can selects music or videos. Figure 2 shows the implementation of a control point on a PDA. Also mobile phones can be used for this purpose.

6.3. Playing Media

By selecting a music or video title, the request is sent to the nearest renderer. The renderer requests the media data directly from the server and performs the presentation.

6.4. Considering the Location of Persons (location awareness)

An important aspect in the ubiquitous media scenario is the ubiquitous access of media. If the user moves to a different room, e.g. the music may follow him. Several implementations are conceivable. One solution is a passive infrared sensor system in each room. The problem with this solution consists in a restriction of privacy [12]. Further more the system may fail in case of slow movement, more persons and animals.



Titel: Track
Künstler:
Typ: Music Track
Server: Alle Server

00:01:13 Track 01

□ "Follow-Me"

Intel Media Renderer (bernhard-b98c68)
PLAYING - Track 01
Intel Media Renderer (bernhard-b98c68)

Name	Wert
RelativeTimePosition	00:00:25
Sharpness	
TransportPlaySpeed	1
TransportState	PLAYING
TransportStatus	OK
VerticalKeystone	
Volume	60
VolumeDB	

00:00:25 Track 01

□ "Follow-Me"

Figure 2. Selecting and controlling media presentation.

We decided to take the position of the currently used Bluetooth access point into consideration, in order to calculate the position of the PDA-and its user. Bluetooth cells are small, so the location precision is sufficient. Using the UPnP architecture an access point can be seen as a server, which provides position data. An automated control point with a minimal user interface tracks this data and-in case of user movement-it instructs the current renderer to stop. Then it directs a secondary renderer e.g. in the kitchen to continue the playback at the interrupted playback position.

6.5. Scriptable Control Point (Stub)

The system can also react on other events. In order to keep the implementation simple we developed a scriptable control point. With this, it is easy to implement automated tasks. New Interfaces might be integrated in a simple way.

The system may measure the loudness in the room. A ubiquitous media system should notice this and might adapt the speaker volume.

6.6. Context awareness (telephone calls)

A ubiquitous media system should adapt to several events. One type of event are telephone calls. During a phone call, a television transmission should be set to time shift mode. Music and video presentations should be set to pause.

Our implementation uses voice over IP telephony in order to simplify the implementation. We implemented a combination of asterisk VoIP client [18] and the scriptable control point, which stops the music playback. After terminating the call, the playback continues.

6.7. Magic DVD Cover

The classical DVD cover has still some importance, as human are orientated in a way that they like to take things with her hand. That's why we like to increase the importance of the DVD cover. We are doing this by adding an RFID tag. By taking a DVD from the shelf and placing it in front of the renderer, this DVD will be discovered by an integrated RFID receiver, which indicates the ubiquitous media system to play the corresponding movie.

6.8. CD Cover with Touch Sensitive Elements

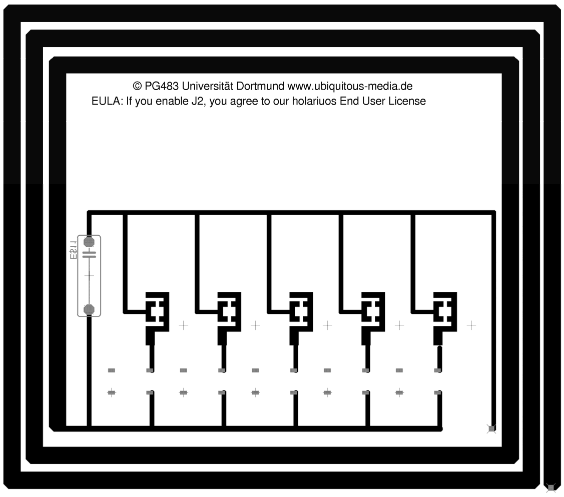

© PG483 Universität Dortmund www.ubiquitous-media.de
EULA: If you enable J2, you agree to our holariuos End User License

Figure 3. CD-cover integrated switchable RFID-TAG.

In order of introducing new user interfaces, we developed a CD cover with sensitive elements. By touching e.g. the third track on the content list of the CD booklet, the ubiquitous media system will play the according track.

However, it would expect a bit too much if the user has to replace batteries even in his CD booklet. So we have to choose a technique which works without batteries. Therefore, we implemented RFID-Tags in the CD cover. To distinguish between the different music tracks we developed switched RFID tags, which are activated individually by pressing a selected region on the booklet. Figure 3 shows the antenna and fife RFID tags together with the corresponding switches.

A RFID receiver might receive the commands from the RFID tag. This RFID receiver works also as a control point, as it tells the renderer to play a certain music file.

7. Conclusions

Within this implementation, we were able to introduce some basic concepts for ubiquitous media. We implemented access transparency, so the user can access media, regardless from the presentation device, underlying network, and media server.

Secondary we implemented location transparency. The user can access media, regardless from its storage place. With this, it is possible to build a "follow me" mode for music. If the user moves to a different room, the music will pause and continue in the other room (location awareness).

And third, we implemented format transparency. The system itself is responsible for content and coding of different formats (mp3, pcm).

We did not implemented DRM transparency, as DRM works against the ubiquitous media concept. The goal of DRM developer is to fix content to a storage medium or a player, in order of controlling the presentation. Further work is required in order to fulfill DRM transparency by integrating different media player systems.

With our magic DVD cover we came back to haptic based user interfaces. Our CD cover with touch sensitive elements leads us to intuitive user interface. By introducing context awareness a phone call can set the television transmission to time shift.

The protocol UPnP (Universal Plug and Play) gives a good foundation for the implementation of a flexible ubiquitous media system.

8. References

[1] J. A. Vince and R. Earnshaw, “Digital media: The Future,” London (Springer), 2000.

[2] F. Schöner, “Multimedia revolution der musik-und medienwirtschaft,” Reinhard Flender, Elmar Lampson, Musik im Internet, Berlin (Kulturverlag Kadmos), pp. S83−S110, 2001.

[3] S. Drews, “Ubiquitous media, vision des digital home der zukunft und anforderungen hinsichtlich seiner realisierung,” Magisterarbeit, Technische Universität Berlin.

[4] N. Dyer and J. Bowskill, “Ubiquitous communications and media-steps toward a wearable learning tool,” J. A. Vince, R. Earnshaw (Eds.), Digital Media: The Future, London (Springer), pp. S61− S74. 2000.

[5] G. Kalkbrenner, “Mobile management of local infrastructure,” Softcom, 2002.

[6] G. Kalkbrenner and F. Nebojsa, “Campus mobil-mobile services for campus and student needs,” Softcom, 2002.

[7] http://www.upnp.org/.

[8] Wireless World Research Forum, Book of Visions, Visions of the Wireless World, 2001.

[9] B.−L. Tim, “The World Wide Web: Past, present and future,” o.O.

[10] L. Barkhuus and A. Dey , “Is context-aware computing taking control away from the user? Three levels of interactivity examined,” Proceedings of UbiComp ’03, pp. 150−156, Springer, 2003.

[11] W. Buxton, “Living in augmented reality-ubiquitous media and reactive environments,” K. Finn, A. Sellen, S. Wilber, (Eds.), Video, Mediated Communication, Hillsdale N. J. (Erlbaum), pp. 363−384, 1997.

[12] B. N. Schilit, J. Hong, and M. Gruteser, “Wireless location privacy protection computer,” Vol. 36, No. 12, pp. 135−137, December 2003.

[13] Special Issue on ContextAware. Computing, Personal and Ubiquitous Computing, Vol. 5, Springer Verlag, 2001.

[14] M. Weiser, “The testbed devices of the infrastructure for ubiqitous computing project,” 1994. http://ubiq.com/hypertext/weiser/testbeddevices.html.

[15] Abowd, Gregory: Ubiquitous Computing International Conference, Proceedings, Berlin, Heidelberg, New York, Barcelona 2001.

[16] L. Lessig, “The future of ideas - The fate of the commons in a connected world,” New York (Vintage Books), 2002.

[17] F. Mattern (Ed.), “Total vernetzt Szenarien einer informatisierten Welt,” Berlin, Heidelberg, New York, u. a. (Springer), 2003.

[18] http://www.asterisk.org/.

Call for Papers

International Journal of Communications, Network and System Sciences (IJCNS)

ISSN 1913-3715 (Print) ISSN 1913-3723 (Online)

http://www.scirp.org/journal/ijcns/

IJCNS is an international refereed journal dedicated to the latest advancement of communications and network technologies. The goal of this journal is to keep a record of the state-of-the-art research and promote the research work in these fast moving areas.

Editors-in-Chief

Prof. Huaibei Zhou	Advanced Research Center for Sci. & Tech., Wuhan University, China
Prof. Tom Hou	Department of Electrical and Computer Engineering, Virginia Tech., USA

Subject Coverage

This journal invites original research and review papers that address the following issues in wireless communications and networks. Topics of interest include, but are not limited to:

- MIMO and OFDM technologies
- UWB technologies
- Wave propagation and antenna design
- Signal processing and channel modeling
- Coding, detection and modulation
- 3G and 4G technologies
- Sensor networks
- Ad Hoc and mesh networks
- Network protocol, QoS and congestion control
- Efficient MAC and resource management protocols
- Simulation and optimization tools
- Network security

We are also interested in:

- Short reports—Discussion corner of the journal:

  2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data.
- Book reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

http://www.scirp.org/journal/ijcns ijcns@scirp.org

I. J. Communications, Network and System Sciences, **2009, 2, 91-168**
Published Online May 2009 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 2 **May 2009**

ISSN: 1913-3715

9771913371005 06