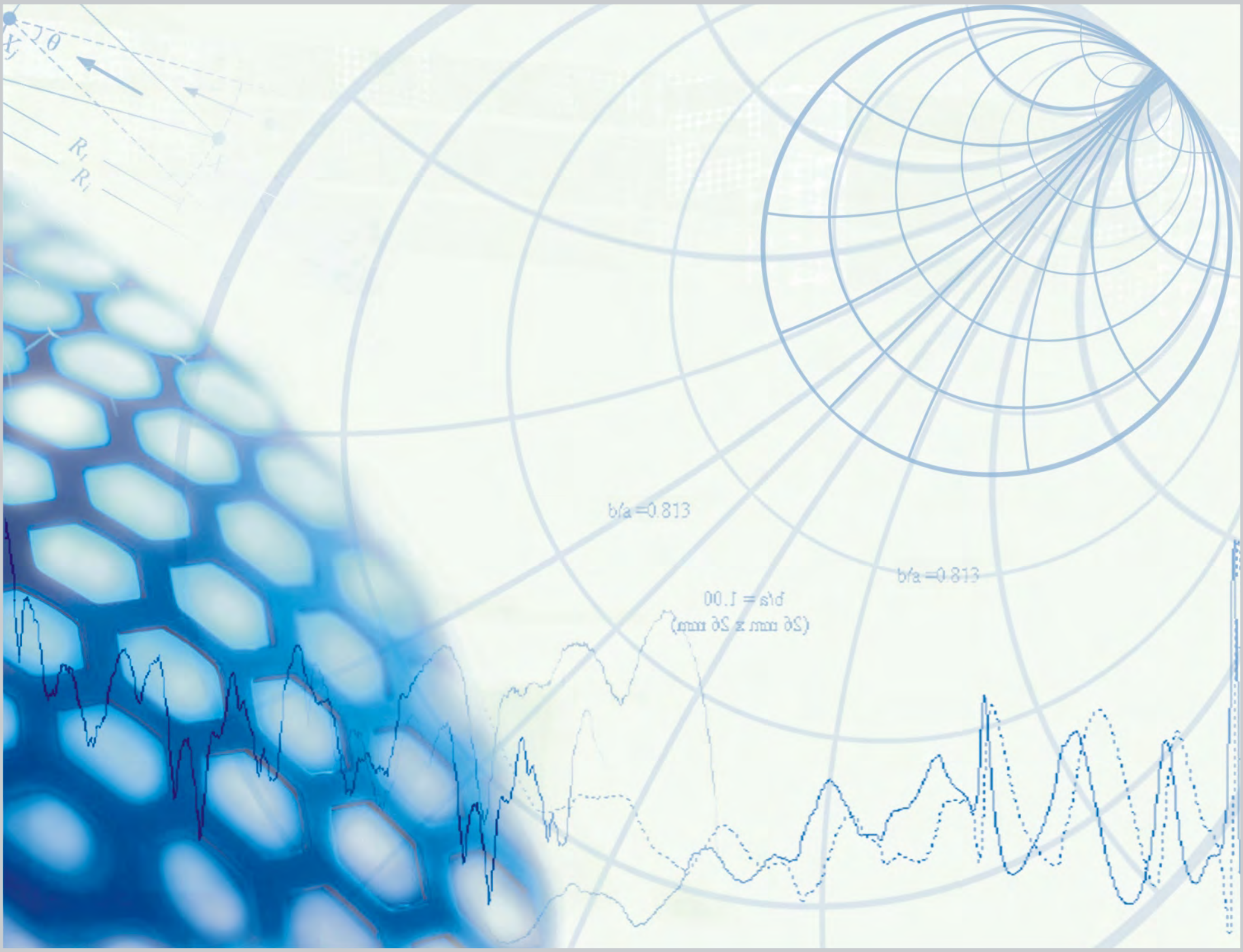
International Journal of

Communications, Network and System Sciences

ISSN: 1913-3715

Vol. 1, No. 4, November 2008



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, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 1 **November 2008**

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 1 2008

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

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:

E-mail: ijcns@scirp.org

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Collision-Tolerant Transmission with Narrow-Beam Antennas

Hong-Ning DAI[1], Kam-Wing NG[1], Min-You WU[2], Bo LI[3]
[1]*Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR*
[2]*Computer Science and Engineering Department, Shanghai Jiao Tong University, Shanghai, P.R.China*
[3]*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, P.R.China*
Email: {hndai,kwng}@cse.cuhk.edu.hk, wu-my@sjtu.edu.cn, bill.leebo@gmail.com
Received April 29, 2008; revised September 1, 2008; accepted September 23, 2008.

Abstract

The application of directional antennas in wireless ad hoc networks brings numerous benefits, such as increased spatial reuse and mitigated interference. Most MAC protocols with directional antennas are based on the RTS/CTS mechanism which works well in wireless ad hoc networks using omni-directional antennas. However, RTS/CTS frames cannot mitigate the interference completely. Besides, they also contribute a lot to the performance overhead. This paper studies the problem from a new perspective. We have found that the transmission success probability under directional transmission and directional reception is quite high when the antenna beamwidth is quite narrow. Motivated by the analytical results, we design a lightweight MAC protocol without RTS/CTS frames. The evaluation results demonstrate that this new protocol performs better than MAC protocols based on the RTS/CTS mechanism. The results also show that a collision-tolerant transmission is feasible under the narrow beam configuration.

Keywords: Wireless Networks, Directional Antennas, Medium Access Control

1. Introduction

The application of *directional* antennas to wireless ad hoc networks has received enormous interest in recent years. Directional antennas can greatly improve network performance by increasing network connectivity, expanding transmission range, enhancing spatial reuse and reducing interference. Recent studies such as [1–10] focus on designing new MAC layer protocols to improve network performance.

Most of these MAC schemes with directional antennas are based on a four-way handshaking scheme, known as request-to-send/clear-to-send (RTS/CTS). The RTS/CTS mechanism has been proposed to resolve the hidden terminal problem in wireless networks using *omni-directional* antennas which can broadcast RTS/CTS frames to inform neighboring nodes of the oncoming transmission. Those nodes that have received the RTS/CTS frames can defer their transmission to avoid collisions. However, using RTS/CTS cannot eliminate hidden terminals completely even in wireless networks with omni-directional antennas [11]. Furthermore, Choudhury et al. [5] have found that using directional antennas causes new interference such as new hidden terminals and the deafness problem, which cannot be solved by using the RTS/CTS mechanism. Essentially, directional antennas can radiate or receive signals more effectively in one direction, which can cause much less interference than omni-directional antennas. So, does the RTS/CTS mechanism still work well with directional antennas?

Many novel mechanisms have been proposed to eliminate the new hidden terminal problem and the deafness problem, which are caused by directional antennas. Although Korakis et al. [3] propose a Circular-DMAC scheme to combat the new hidden terminal problem and the deafness problem, transmitting multiple RTS/CTS frames for each data transmission severely degrades the performance. Other schemes, such as Tone-based DMAC [7] and BTDMAC [12] can alleviate the impacts of the hidden terminal and deafness problems by sending tones over another channel or over the data channel after data transmission. However, these bulky and complicated schemes also bring additional cost and performance penalty.

How to use directional antennas in wireless networks more effectively? We address this problem from another viewpoint. When the beamwidth of a directional antenna is lessened (a narrower beamwidth), the interference caused by the antenna will also be reduced. We have found that when the beamwidth is quite narrow and the network is not so dense, the collision probability is quite low. It is the purpose of this paper to study the performance of wireless networks using narrow-beam antennas. In particular, we are interested in the following problems:

- What will happen when the beamwidth of the directional antennas is lessened? What is the impact of other factors on the success transmission probability, such as the node density?
- How effective is the RTS/CTS mechanism in wireless networks using directional antennas? If RTS/CTS is turned off, will the network throughput degrade significantly?

In the next section, we briefly survey the related work in the literature. Section 3 describes the models used in this paper and analyzes the success transmission probability for directional transmission and directional reception. In Section 4, we present a lightweight MAC protocol without the RTS/CTS mechanism and compare its performance with a representative MAC protocol using the RTS/CTS mechanism. Section 5 offers some implications of our results. Finally, we summarize our paper in Section 6.

2. Related Work

Many studies [1–10] focus on designing new MAC protocols with directional antennas. Most of them are based on the IEEE 802.11MAC [13], which typically uses RTS/CTS to reduce interference in wireless networks. Although the RTS/CTS mechanism works well in wireless networks equipped with omni-directional antennas, it cannot mitigate interference completely [11]. Besides, using RTS/CTS in wireless networks with directional antennas is not as effective as we expected. For example, ref. [5] shows that RTS/CTS cannot completely mitigate new interfering nodes caused by directional antennas.

To address the new hidden terminal problem and the deafness problem, many researchers propose more complex schemes, such as Circular-DMAC [3], Tone-based DMAC [7] and BT-DMAC [12]. Although they can mitigate the impacts of hidden terminals and deafness, they also bring additional cost on network performance. For example, Circular-DMAC needs a sender to transmit multiple RTS frames before each data transmission, which greatly degrades the network performance. Tone-based DMAC and BT-DMAC also need to send out-of-band tone signals to reduce interference.

Other studies [14–17] concentrate on capacity analysis and performance evaluation on wireless ad hoc networks using directional antennas. Yi et al. [15] have found that using directional antenna in arbitrary networks achieves a capacity gain of $2\pi/\sqrt{\alpha\beta}$ when both transmission and reception are directional. Here, α and β are transmitter and receiver antenna beamwidths, respectively. Under random networks, the throughput improvement factor is $4\pi^2/(\alpha\beta)$ for directional transmission directional reception. Ref. [14] studies the asymptotic bounds on the amount of capacity gains that directional antennas can acquire. Wang et al. [16] model and analyze multiple directional transmission and reception modes coupled with omni-directional or directional receptions. Carvalho and Garcia-Luna-Aceves [17] propose a realistic analytical model which considers the binary exponential back-off operation of IEEE 802.11.

In this paper, we try to find the relationship between the interference, the beamwidth of directional antennas and the density of nodes.

3. Analytical Models

In this section, we analyze the successful transmission probability with directional antennas. The successful transmission probability does not only depend on the activity of the interfering nodes but also on the transmission/reception mode of directional antennas. First, we present the antenna model in Section 3.1. Section 3.2 discusses the interference range for directional transmission. Finally, we analyze the successful transmission probability under the directional transmission and directional reception mode.

3.1. Antenna Model

The radiation pattern of a direction antenna is often depicted as the gain values in each direction in space. We can project the radiation pattern of an antenna to an azimuthal or elevation plane. The projection of the pattern typically has a main *lobe* (beam) of the peak gain and side-/back-lobes of smaller gains.

Since modeling a real antenna with precise values for main and side-/back-lobes is difficult, we use an approximate

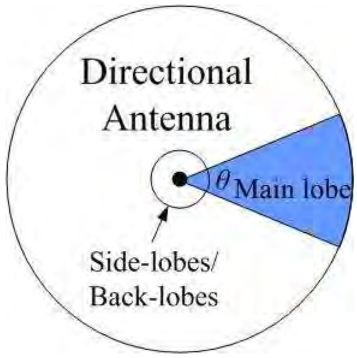


Figure 1. The antenna model.

antenna pattern [18]. In an azimuthal plane, the main lobe of antenna can be depicted as a sector with angle θ, which is denoted as the main beamwidth of the antenna. The side-/back-lobes are aggregated to a circle, as shown in Figure 1. The narrower the main beamwidth of the antenna is, the smaller the side-/back-lobes are. Take the above antenna model as the example. The gain of the main beam is more than 100 times of the gain of side-lobes when the main beamwidth is less than 40° [18]. Thus, the side-/back-lobes can be ignored when the main beam is quite narrow.

Our proposed model assumes that a directional antenna gain G^d is within a specific angle θ, where θ is the beamwidth of the antenna. The gain outside the beamwidth is assumed to be zero. At any time, the antenna beam can only be pointed to a certain direction, as shown in Figure 1, in which the antenna is pointing to the right. Thus, the probability that the beam is switched to cover each direction is $\theta/(2\pi)$. The antenna gain pattern is given by:

$$g(\theta)=\begin{cases} G^d & \text{if angle within } \theta \\ 0 & \text{otherwise} \end{cases}$$

3.2. Interference Range

Compared with omni-directional antennas, directional antennas have different transmission properties. Directional antennas can radiate or receive ratio signals more effectively in a certain direction than other directions. Thus, directional antennas have different transmission range and interference region, compared with omni-directional antennas. In this subsection, we investigate the interference region of directional antennas and the relationship between the transmission range and the interference range.

When a signal is propagated from the transmitter to the receiver, whether it is correctly accepted by the receiver is mainly determined by the receiving power of the signal at the receiver end. In open space, if the transmitting power is fixed, the receiving power is mostly decided by the path loss along the distance between the transmitter and the receiver. Under this condition, multi-path and shadowing effects can be ignored since they are so trivial compared with the large path loss. Therefore, in this paper, we assume that the signal propagation follows the *two-way ground* model which is typically used in open space.

According to [19], under the assumption of the two-way ground model, the receiving power of a signal at the receiver can be calculated by the following equation.

$$P_r(d)=P_tG_tG_r\frac{h_t^2h_r^2}{d^4} \quad \textbf{(1)}$$

where $P_r(d)$ is the receiving power at the receiver which is far from the transmitter with the distance d, P_t is the

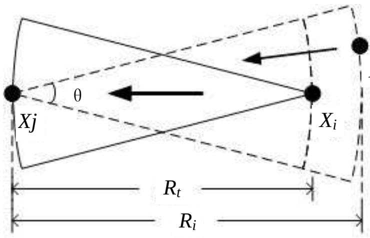


Figure 2. The interference model.

transmitter power, G_t and G_r are the transmitter antenna gain and the receiver antenna gain, respectively, and h_t and h_r are the antenna height of the transmitter and the antenna height of the receiver, respectively.

Consider a large-scale wireless ad hoc network with n static nodes. Without loss of generality, the distribution of the nodes follows a Poisson distribution with a parameter ρ over the 2-D plane. The probability $p(i,S)$ of finding i nodes in an area of S is given by:

$$p(i,S)=\frac{(\rho S)^i}{i!}e^{-\rho S} \quad \textbf{(2)}$$

We also assume that every node has an identical antenna and transmits with a fixed power. Thus, each node has the same transmitting range R_t and the same interference range R_i. In the scenario shown in Figure 2, suppose that node X_i transmits to node X_j over a channel. The receiver X_j locates exactly within the transmitting range R_t of the transmitter X_i.

The successful reception of the signal is mainly decided by the *signal-to-interference-plus-noise ratio* (SINR), which is often required to be greater than a threshold. When their transmission is on-going, an interfering node X_k at the interference range R_i away from the receiver starts the transmission toward the receiver at the same time. So, it will have an interfering signal with the strength P_i at the receiver X_j. Since the thermal noise is negligible compared with interference signals, similar to [11], we do not count it in our model as well. Thus, we have $SINR=\frac{P_r}{P_t}=\frac{R_i^4}{R_t^4}\geq\sigma$, where σ is the SINR threshold. In practice, σ is usually set to 10. So, we get the interference range $R_i=\sqrt[4]{\sigma}R_t$.

3.3. Directional Transmission and Directional Reception

A directional antenna has two modes: an omni-directional mode with a gain G^o and a directional mode with a gain G^d. Since antennas in the directional mode can radiate or receive radio waves more effectively in some directions than in others, the directional gain G^d is generally greater than the omni-directional gain G^o. The transmitter or the receiver equipped with a directional antenna can choose any one of the two modes to transmit or receive frames. Hence, there are four combinations for the transmission and reception modes of directional antennas: 1) Omni-directional Transmission and Omni-directional Reception (OTOR); 2) Directional Transmission and Omni-directional Reception (DTOR); 3) Omni-directional Transmission and Directional Reception (OTDR); 4) Directional Transmission and Directional Reception (DTDR).

According to Equation (1), the larger the antenna gains both the transmitter and the receiver have, the higher the receiving power the receiver has. Besides, the transmission range between the transmitter and the receiver will be extended if the antenna gains of them are increased. Thus, when both the receiver and the transmitter use the directional mode, the communication range between them is maximized. On the other hand, the receiver is only susceptible to the interfering signals from its receiving direction when it is using the directional mode. So, DTDR also has the smallest interference area compared with the other three modes. Hence, DTDR is a preferred method to utilize directional antennas. In this paper we only discuss the transmission under the DTDR mode.

Let us consider the scenario shown in Figure 2. When node X_i begins to transmit with node X_j, this packet is successfully received by node X_j if no node within the sector region covered by X_j's antenna beam transmits toward X_j. First, we need to calculate the probability that no node can interfere with node X_j. Since the placement of nodes follows the 2-D Poisson distribution with the density ρ, there are $\rho\pi R_i^2 \cdot \frac{\theta}{2\pi}$ nodes within the sector region covered by X_j's antenna beam. The area of this region is denoted by S. Among these nodes, the interfering node X_k can cause interference with node X_j only when it has a frame to send and its antenna beam is pointed to node X_j. We assume that a node begins to transmit with a probability p.Then, the probability that node X_k can interfere with node X_j is $p \cdot \frac{\theta}{2\pi}$. There, the probability P that no nodes within region can cause collisions with node X_j is given by:

$$P = \sum_{i=0}^{\infty}(1-p\frac{\theta}{2\pi})^i \cdot \frac{(\rho S)^i}{i!}e^{-\rho S} = e^{-p\frac{\theta}{2\pi}\rho S} = e^{-p(\frac{\theta}{2\pi})^2 \rho\pi R_i^2} \qquad \textbf{(3)}$$

To simplify the calculation, we use $N = \rho\pi R_t^2$, which denotes the average number of nodes within a node's transmission range. Since $R_i = \sqrt[4]{\sigma}R_t$, we have $\rho\pi R_i^2 = \sqrt{\sigma}N$. Replacing the corresponding part in Equation (3), we have:

$$P = e^{-p(\frac{\theta}{2\pi})^2\sqrt{\sigma}N} \qquad \textbf{(4)}$$

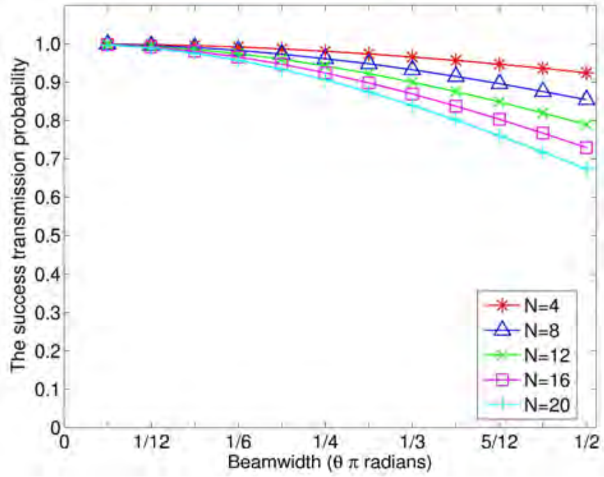


Figure 3. The probability of a successful transmission.

When $p = 0.1$ and $\sigma = 10$, we set different $N = 4,8,12,16,20$ respectively and then we get the results in Figure 3. Figure 3 shows that the successful transmission probability is high when the beamwidth is narrow. For example, when θ is less than $\frac{\pi}{6}$, the success probability is always above 95%. One possible reason is that using directional mode at the receiver end can greatly reduce the collision probability.

Results under a narrow beamwidth ($\theta \le \frac{\pi}{12}$) are also tabulated in Table 1, which shows that the transmission under DTDR is less vulnerable to interference when the beamwidth is quite narrow.

The analytical results under DTDR show that the successful transmission probability is quite high when the beamwidth is lessened enough. For example, when $\theta \le \frac{\pi}{12}$ (i.e., 15°), the success probability is always above 98%. A beamwidth of 15° is a feasible angle for most directional antennas. Thus, intuitively, there is an

Table 1. The probability of a successful transmission under the very narrow beam.

	$\theta = \frac{\pi}{48}$	$\theta = \frac{\pi}{36}$	$\theta = \frac{\pi}{24}$	$\theta = \frac{\pi}{12}$
$N = 4$	0.9999	0.9998	0.9995	0.9978
$N = 8$	0.9997	0.9995	0.9989	0.9956
$N = 12$	0.9996	0.9993	0.9984	0.9934
$N = 16$	0.9995	0.999	0.9978	0.9913
$N = 20$	0.9993	0.9988	0.9973	0.9891

interesting question: can the transmission continue even if there exist few collisions? In other words, when the beamwidth of antennas is narrow enough and the collision probability is quite low, can the transmission be collision-tolerated?

4. Lightweight MAC Protocol

In this section, we propose a lightweight MAC scheme denoted as Basic Directional Transmission and Directional Reception (B-DTDR), which turns off the control frames of request to send (RTS) and clear to send (CTS). It has a rival termed RTC/CTS Directional Transmission and Directional Reception (RTS-DTDR). Then, we compare the performance of B-DTDR with that of RTS-DTDR and discuss the implications from this lightweight scheme.

4.1. Quick Review of RTS/CTS Mechanisms with Directional Antennas

B-DTDR scheme keeps the basic collision avoidance features, such as the exponential backoff and the control frame of acknowledge (ACK). Thus, if there is a collision with data packets, those data packets need to be retransmitted. In IEEE 802.11 distributed coordination function (DCF) [13], an exponential backoff scheme can be used to avoid further collisions of data packets. At each packet transmission, the backoff time is uniformly chosen in a range of (o, w) where w is called the contention window. At the first transmission, w is set to be the minimum value of $W_{\min}$. After each unsuccessful transmission (no ACK received), the size of w is doubled until it reaches the maximum value of $W_{\max}$. This mechanism can effectively reduce the collision probability.

Most of current directional MAC schemes are using another four-way handshaking technique which turns on RTS/CTS frames. In stead of sending a data packet, a transmitter sends a short control frame called request to send (RTS). After the reception of RTS, the receiver responds the transmitter with the frame called clear to send (CTS). After shaking hands of RTS and CTS, the data transmission begins. This mechanism is useful to reduce the hidden nodes in wireless networks using omni-directional antennas. However, it cannot mitigate the hidden terminal problem and the deafness problem with directional antennas [5]. In this paper, we consider a general MAC scheme (RTSDTDR) which can be used to represent the current directional MAC schemes since it keeps has the main features of them. In both B-DTDR and RTS-DTDR, RTS, CTS, data packets and ACK are transmitted directionally.

One of the difficult problems with B-DTDR and RTSDTDR is to find the location of a node's neighbors, or *neighbor discovery*. This problem can be solved by using DOA caching [4] or similar mechanisms. A specific problem with B-DTDR is how to help a receiver to know that a transmitter is trying to send a frame to it. Zhang [10] proposes a scheduling mechanism to address this problem. In this paper, we assume that both B-DTDR and RTS-DTDR can solve the neighbor discovery problem.

On the other hand, when a node receives any frames (RTS, CTS and data frames), it will record the corresponding information into its DNAV (Directional Network Allocation Vector), which is a directional version of NAV of IEEE 802.11, proposed in [4,5]. DNAV excludes the directions and sets the corresponding durations, toward which the node is not allowed to initiate a transmission to avoid collisions with data or control frames. When a node receives a frame and the frame is for this node, it beamforms toward the transmitter (switch to directional mode) and replies the frame with a CTS (or ACK) frame. If the frames are not for itself, it will update the sender's information and set the corresponding DNAVs. DNAVs are used in both B-DTDR and RTS-DTDR.

4.2. Performance Model

In this paper, we adopt a discrete Markov chain model used in [16, 20] to evaluate the saturation throughput and the overhead of wireless networks (as shown in Figure 4). We extend the model to support directional antennas. Range extension and overhead calculation are also considered in our model. We also adopt the assumption that each node operate in time-slotted mode, with a time slot τ. If the time slot τ is very small, the performance of the time-slotted protocol is very close to that one of the asynchronous version of the protocol [16, 20]. The period of time during which RTS, CTS, data and ACK frames are transmitted can be depicted as multiples of τ, i.e., t_{rts}, t_{cts}, t_{data} and t_{ack}, respectively.

The throughput is calculated by the proportion of time that a node spends transmitting data packets successfully on the average. Let $P(S)$, $P(I)$ and $P(C)$ denote the steady-state probability of *SUCCESS*, *IDLENESS* and *COLLISION*, respectively. From the Markov chain model

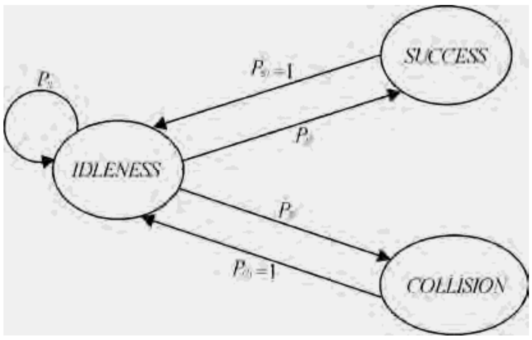


Figure 4. The Markov chain model for a node.

shown in Figure 4, we have the following equation to calculate the throughput.

$$Throughput = \frac{P(S) \cdot t_{data}}{P(C)T_C + P(S)T_S + P(I)T_I} \quad \textbf{(5)}$$

where T_C, T_S and T_I are the duration of *COLLISION*, *SUCCESS* and *IDLENESS*, respectively.

The duration of time that a node stays in the *SUCCESS* state, T_S or the collision state, T_C, depend on the mechanisms of different MAC protocols. Thus, the detailed calculation will be given in the following subsections. The duration of a node in *IDLENESS* state T_I is 1τ.

Then, we need to calculate the probabilities that the node stays in different states. From Figure 4, the steady-state probability of *IDLENESS* equals:

$$P(I) = P(I) \cdot P_{II} + P(S) + P(C) \quad \textbf{(6)}$$

Note that $P(S) + P(C) = 1 - P(I)$, thus,

$$P(I) = 1/(2 - P_{II}) \quad \textbf{(7)}$$

From Figure 4, the steady-state probability of *SUCCESS* can be calculated by $P(S) = P(I) \cdot P_{IS}$. Before deriving the transition probability P_{IS} from *IDLENESS* to *SUCCESS*, we need to calculate $P_{IS}(r)$ that node X_i successfully shakes hands with node X_j which is a distance r way. The detailed calculation of $P_{IS}(r)$ will be stated as follows.

We also derive the MAC overhead by calculating the portion of time that a node spends transmitting control frames on the average when data packets are successfully transmitted.

$$Overhead = \frac{P(S) \cdot t_{ctrl}}{P(C)T_C + P(S)T_S + P(I)T_I} \quad \textbf{(8)}$$

where t_{ctrl} is depicted as time slots which are used to transmit control frames such as RTS, CTS and ACK.

In the following subsections, we derive the steady-state probabilities, transition probabilities and times spent at different states of the two MAC schemes, respectively.

4.3. RTS/CTS Based Directional Transmission and Directional Reception (RTS-DTDR)

In this subsection, we calculate the throughput and the overhead of RTS-DTDR. From the throughput model presented above, we need to calculate the transition probability P_{IS} first. Figure 5 indicates that the nodes within the four regions (named 1, 2, 3, 4) may interfere

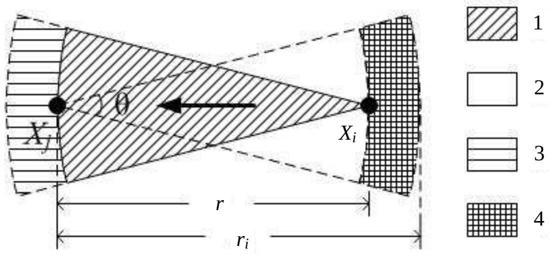


Figure 5. The interference region for DTDR.

with node X_i and node X_j. The transmission range between X_i and X_j is denoted as r. So, the interference range $r_i = \sqrt[4]{\sigma} r$, which can be easily derived from the results presented in Section 3.2. Since the number of nodes depends on the area size, we need to calculate the four areas of regions 1, 2, 3 and 4, which are denoted as S_1, S_2, S_3 and S_4, respectively:

$$\begin{aligned} S_1 &= \pi r^2 \cdot \theta/(2\pi) \\ S_2 &= \pi r^2 \cdot \theta/(2\pi) - r^2 \tan(\theta/2)/2 \\ S_3 &= \pi r_i^2 \cdot \theta/(2\pi) - \pi r^2 \cdot \theta/(2\pi) \\ S_4 &= \pi r_i^2 \cdot \theta/(2\pi) - \pi r^2 \cdot \theta/(2\pi) \end{aligned} \quad \textbf{(9)}$$

$P_{IS}(r)$ equals the probability that X_i transmits in a given time slot, and X_j does not transmit in the same time slot, and none of the nodes within the four regions interferes with the handshake between nodes X_i and X_j. Therefore, we have:

$$P_{IS}(r) = p(1-p) \cdot P_1 \cdot P_2 \cdot P_3 \cdot P_4 \quad \textbf{(10)}$$

Since RTS-DTDR does not prevent interference from neighboring nodes in regions 3 and 4, the handshake might be interrupted at any time. Hence, the *COLLISION* period T_C lasts from $T_1 = t_{rts} + 1$ till $T_2 = t_{rts} + t_{cts} + t_{data} + t_{ack} + 4$, where one propagation delay 1τ is also considered. T_C is the mean value of the truncated geometric distribution. Then, we obtain the following equation to calculate T_C.

$$T_C = (1-p)/(1-p^{T_2 - T_1 + 1}) \sum_{i=0}^{T_2 - T_1} p^i (T_1 + i) \quad \textbf{(11)}$$

The probability that no nodes in region 1 interferes with the handshake between nodes X_i and X_j is equal to the probability that no node in this area transmits as node X_i does, which can be depicted as:

$$\begin{aligned} P_1 &= \sum_{i=0}^{\infty} (1 - p\frac{\theta}{2\pi})^i \cdot \frac{(\rho S_1)^i}{i!} e^{-\rho S_1} \\ &= e^{-p\rho S_1 \frac{\theta}{2\pi}} \end{aligned} \quad \textbf{(12)}$$

The probability P_2 is equal to the probability that no node transmits toward node X_j within the duration of time t_{rts} and no node transmits within the slot when node X_i begins to transmit toward node X_j Thus, we have the following equation to calculate P_2 .

$$P_2 = e^{-p\frac{\theta}{2\pi}\rho S_2(t_{rts}+1)} \cdot e^{-p\frac{\theta}{2\pi}\rho S_2} \tag{13}$$

P_3 is equal to the probability that no node can interfere with the reception of CTS and ACK frames of node X_i . Hence, we have:

$$P_3 = e^{-p\frac{\theta}{2\pi}\rho S_3(t_{cts}+1)} \cdot e^{-p\frac{\theta}{2\pi}\rho S_3(t_{ack}+1)} \tag{14}$$

In region 4, there is no interference if no node transmits toward node X_j when node X_i is sending a data frame. Then, we get:

$$P_4 = e^{-p\frac{\theta}{2\pi}\rho S_4} \cdot e^{-p\frac{\theta}{2\pi}\rho S_4(t_{data}+1)} \tag{15}$$

Because each transmitter can choose its receiver with the equal probability and the average number of nodes within a region of radius r is proportional to r^2 , the probability density is the function of distance r between nodes X_i and X_j , i.e., $f(r) = 2r$, where $0 < r < R_t$. Therefore, P_{IS} is equal to:

$$\begin{aligned} P_{IS} &= \int_0^{R_t} P_{IS}(r) f(r) dr \\ &= \int_0^{R_t} p(1-p) \cdot P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot 2r dr \end{aligned} \tag{16}$$

The duration in time slots of a node in the *SUCCESS* state is

$$\begin{aligned} T_S &= (t_{rts}+1) + (t_{cts}+1) + (t_{data}+1) + (t_{ack}+1) \\ &= t_{rts} + t_{cts} + t_{data} + t_{ack} + 4 \end{aligned} \tag{17}$$

where t_{rts} , t_{cts} , t_{data} and t_{ack} are the duration times of transmitting RTS, CTS, data and ACK frames, respectively.

After the corresponding parts are replaced in Equation (5), the throughput of RTS-DTDR is obtained. Following the similar process, we can calculate the overhead of RTS-DTDR from Equation (8).

4.4. Basic Directional Transmission and Directional Reception (B-DTDR)

Since there is no RTS and CTS frames, B-DTDR has a narrower bound on T_C (from $T_1 = 1\tau$ to $T_2 = t_{data} + t_{ack} + 2$). Then we can calculate T_C by using Equation (11).

And the success period time is

$$T_S = t_{data} + t_{ack} + 2 \tag{18}$$

P_1 keeps the same as RTS-DTDR. P_2 is equal to the probability that no node transmits toward node X_j within t_{data} period and does not transmit in the slot when node X_i begins to transmit with node X_j , therefore, we have

$$P_2 = e^{-p\frac{\theta}{2\pi}\rho S_2(t_{data}+1)} \cdot e^{-p\frac{\theta}{2\pi}\rho S_2} \tag{19}$$

Similarly, we have

$$\begin{aligned} P_3 &= e^{-p\frac{\theta}{2\pi}\rho S_3(t_{ack}+1)} \\ P_4 &= e^{-p\frac{\theta}{2\pi}\rho S_4} \cdot e^{-p\frac{\theta}{2\pi}\rho S_4(t_{data}+1)} \end{aligned} \tag{20}$$

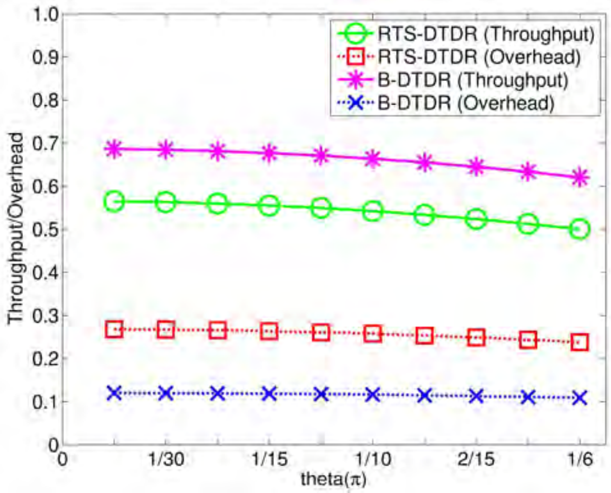
Then after replacing the corresponding parts in Equation (5), we get the throughput of B-DTDR. Since $t_{ctrl} = t_{ack} + 2$ in B-DTDR, we can calculate the overhead of B-DTDR from Equation (8).

4.5. Numerical Results

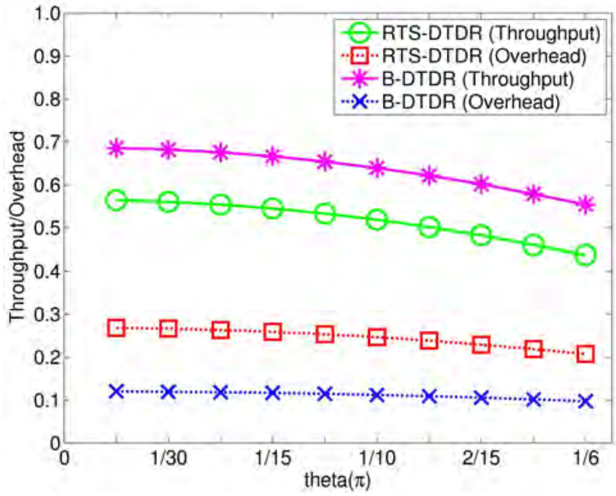
We compare the performance of the RTS-DTDR and BDTDR under the different configurations and present the results in Figure 6 and Figure 7.

Figure 6 shows the saturation throughput and overhead of RTS-DTDR and B-DTDR under different node density ($N = 10, 20, 30, 40$, respectively) when the beamwidth is less than $\frac{\pi}{6}$. The results are obtained under a short data length, i.e., $t_{data} = 40\tau$. With the increased node density, both the throughputs of RTS-DTDR and B-DTDR begin to degrade although B-DTDR has a much higher throughput than RTSDTDR protocol. The peak value of B-DTDR is almost 20% higher than that of RTS-DTDR. One possible reason is that when the beamwidth is quite narrow, the number of the interfering nodes is so small that those nodes cause nearly no collisions. In this situation, RTS/CTS frames are not necessary to be used. On the contrary, they only contribute additional overhead on the throughput.

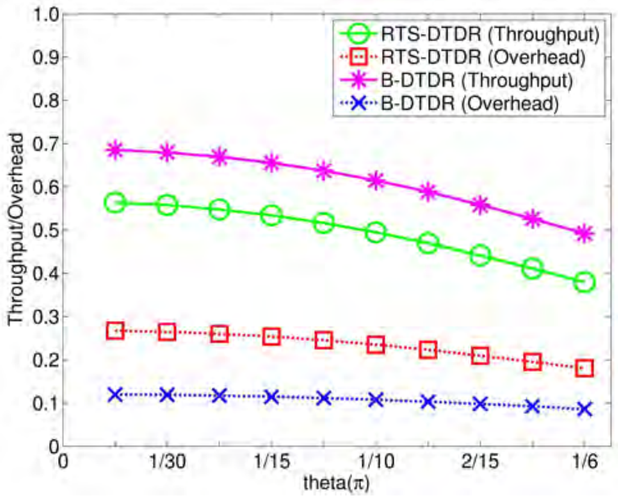
Then we calculate the throughput and overhead under the long data length setting (i.e., $t_{data} = 120\tau$ and the results are shown in Figure 7. Similarly, both RTS-DTDR and B-DTDR perform well under a narrow beam (e.g., $\frac{\pi}{15}$). Under this setting, B-DTDR still has a higher throughput than RTS-DTDR because it gets rid of the bulky RTS/CTS mechanism. However, when the beamwidth is increased further, the collisions caused by



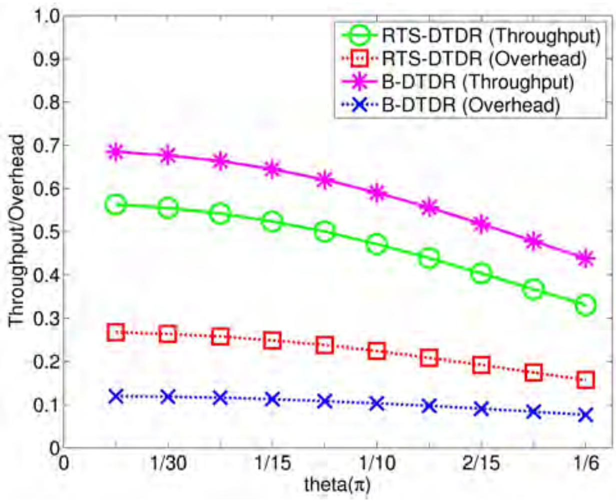
(a) *N*=10



(b) *N*=20



(a) *N*=30



(b) *N*=40

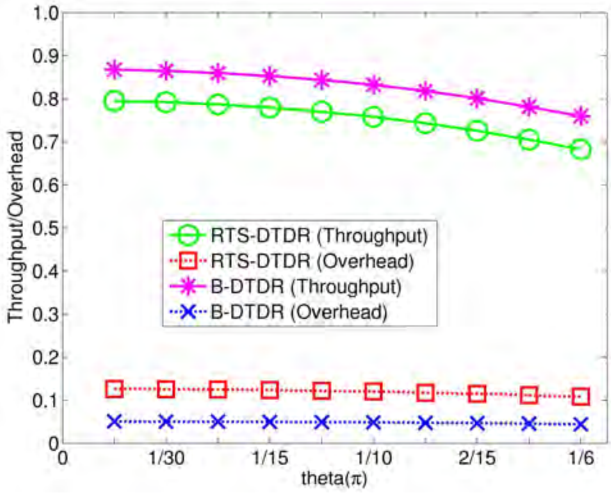
Figure 6. Throughput comparison when $p = 0.1, t_{rts} = t_{cts} = t_{ack} = 5\tau, t_{data} = 40\tau$ (short data frame).

interfering nodes become remarkable, both the throughput of RTS-DTDR and B-DTDR degrades.

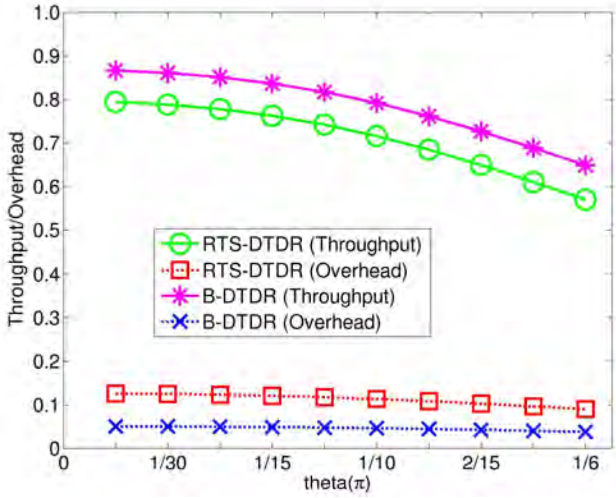
5. Discussions

The results in Figure 6 and Figure 7 show that, when the beamwidth is decreased, a higher network throughput can be obtained. The capacity analysis in [15] also proves that the capacity grows with the lessened beamwidth. However, the capacity will not grow arbitrarily high when the beamwidth decreases further and even approaches to zero. Yi et al. [15] have also observed that when the beamwidth is too small, the interference has been fully reduced and there is no further improvement by decreasing the beamwidth of the antennas.

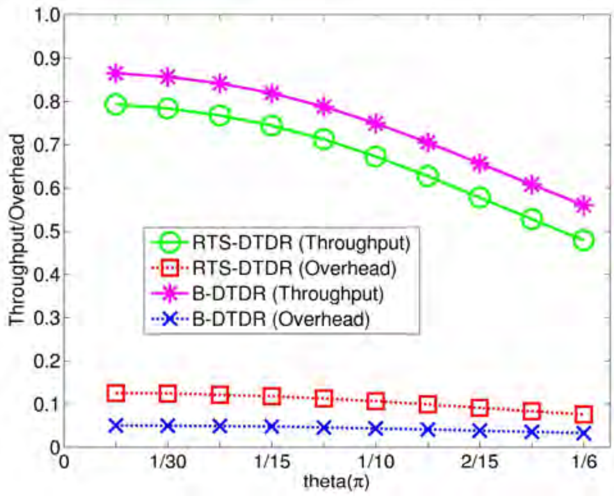
Actually, when the beamwidth is narrow enough (more specifically, less than a certain angle) a trans-mission can yield a high success probability. As shown in Section 3.3, if the beamwidth is less than $\frac{\pi}{12}$ (i.e., 15°) and both directional antennas are used at the transmitter and the receiver, then the probability of a successful transmission is greater than 99%. The transmission under this situation can be regarded as a *collision-tolerant* transmission (the collision probability is quite small). Hence, DTDR should be the best way to use directional antennas. Meanwhile, the angle 15° is feasible in most *intelligent* directional antennas. Under this condition, the complicated collision avoidance mechanisms, such as RTS/CTS, are not necessary to be used because they only contribute excessive overhead on the performance. At that time, using some simple collision avoidance mechanisms, such as the exponential back-off, might be enough to reduce the interference.



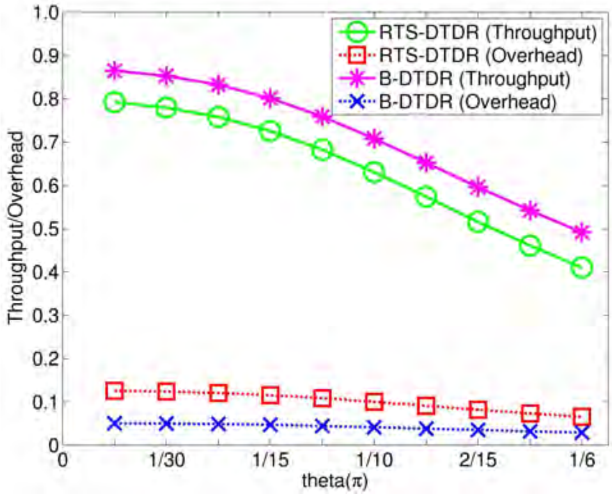
(a) *N*=10



(b) *N*=20



(a) *N*=30



(b) *N*=40

Figure 7. Throughput comparison when $p = 0.1, t_{rts} = t_{cts} = t_{ack} = 5\tau, t_{data} = 120\tau$ (long data frame).

This collision-tolerant transmission gives us some important implications on MAC design. Directional antennas have different properties, e.g., higher spatial reuse and the smaller interfering region. Although RTS/CTS schemes work well in wireless networks using omni-directional antennas, they cannot mitigate interference caused by directional antennas completely [5]. Thus, the MAC layer design with directional antennas should start from another different perspective. For example, when the beamwidth is narrow enough and the collision probability is small, we can turn off RTS/CTS. On the contrary, we should consider other techniques, such as power control and multi-channel schemes to reduce interference.

6. Conclusions

This paper studies the performance of wireless networks using directional antennas with a narrow beam. In particular, we examine the probability of a successful transmission under Directional Transmission and Directional Reception. The numerical results show that the interference probability is quite low when the antenna beamwidth is narrow enough. These results encourage us to design a lightweight MAC protocol which turns off RTS/CTS. The evaluation results prove that the protocol has a higher throughput than the typical MAC protocol based on RTS/CTS. The results also demonstrate that a collision-tolerant transmission is feasible when the beamwidth is narrow enough. One of our future works is to implement the lightweight MAC protocol in simulators and conduct experiments in real environments.

7. Acknowledgements

This research was partially supported by Natural Science

Foundation of China grant No.60573138 and 60773091, and the National Grand Fundamental Research 973 Program of China under Grant No.2006CB303000.

8. References

[1] Y. B. Ko, V. Shankarkumar, and N. H. Vaidya, "Medium access control protocols using directional antennas in ad hoc networks," in Proceedings IEEE INFOCOM, 2000.

[2] A. Nasipuri, S. Ye, and R. E. Hiromoto, "A MAC protocol for mobile ad hoc networks using directional antennas," in Proceedings IEEE WCNC, 2000.

[3] T. Korakis, G. Jakllari, and L. Tassiulas, "A MAC protocol for full exploitation of directional antennas in adhoc wireless networks," in Proceedings MobiHoc, 2003.

[4] M. Takai, J. Martin, R. Bagrodia, and A. Ren, "Directional virtual carrier sensing for directional antennas in mobile ad hoc networks," in Proceedings MobiHoc, 2002.

[5] R. R. Choudhury, X. Yang, N. H. Vaidya, and R. Ramanathan, "Using directional antennas for medium access control in ad hoc networks," in Proceedings MobiCom, 2002.

[6] C. S. Z. Huang, C.-C. Shen, and C. Jaikaeo, "A busytone based directional MAC protocol for ad hoc networks," in Proceedings MILCOM, 2002.

[7] R. R. Choudhury and N. H. Vaidya, "Deafness: a MAC problem in ad hoc networks when using directional antennas," in Proceedings ICNP, 2004.

[8] H. Singh and S. Singh, "Smart-802.11b MAC protocol for use with smart antennas," in Proceedings IEEE ICC, 2004.

[9] L. Bao and J. Garcia-Luna-Aceves, "Transmission scheduling in ad hoc networks with directional antennas," in Proceedings MobiCom, 2002.

[10] Z. Zhang, "Pure directional transmission and reception algorithms in wireless ad hoc networks with directional antennas," in Proceedings IEEE ICC, 2005.

[11] K. Xu, M. Gerla, and S. Bae, "How effective is the IEEE 802.11 RTS/CTS handshake in ad hoc networks," in Proceedings IEEE GLOBECOM, 2002.

[12] H. N. Dai, K. W. Ng, and M. Y. Wu, "A busy-tone based MAC scheme for wireless ad hoc networks using directional antennas," in Proceedings IEEE Globecom, 2007.

[13] IEEE 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications - Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band, IEEE Std., 2003.

[14] A. Spyropoulos and C. S. Raghavendra, "Capacity bounds for ad-hoc networks using directional antennas," in Proceedings IEEE ICC, 2003.

[15] S. Yi, Y. Pei, and S. Kalyanaraman, "On the capacity improvement of ad hoc wireless networks using directional antennas," in Proceedings ACM MobiHoc, 2003.

[16] Y. Wang and J. J. Garcia-Luna-Aceves, "Directional collision avoidance in ad hoc networks," Performance Evaluation Journal, Vol. 58, pp. 215–241, 2004.

[17] M. M. Carvalho and J. J. Garcia-Luna-Aceves, "Modeling wireless ad hoc networks with directional antennas," in Proceedings IEEE INFOCOM, 2006.

[18] R. Ramanathan, "On the performance of ad hoc networks with beamforming antennas," in Proceedings Mobi-Hoc, 2001.

[19] T. S. Rappaport, "Wireless communications: principles and practice," 2nd edition Upper Saddle River, N.J.: Prentice Hall PTR, 2002.

[20] L. Wu and P. K. Varshney, "Performance analysis of CSMA and BTMA protocols in multihop networks (i). single channel case," Information Sciences, Vol. 120, pp. 159–177, 1999.

***I. J. Communications, Network and System Sciences*, 2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

An Experimental Study of the Printed-Circuit Elliptic Dipole Antenna with 1.5-16 GHz Bandwidth

Chunchi LEE[1], Hsinsheng HUANG, Chengda YANG, Chiawei WANG
Department of Computer and Communication, Shu-Te University, Kaohsiung, Taiwan, R.O.C
E-mail: [1]homer1@mail.stu.edu.tw
Received August 1, 2008; revised September 22; accepted September 29, 2008

Abstract

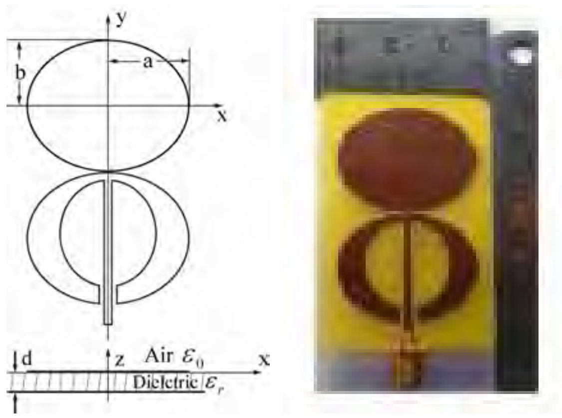
Printed-circuit board (PCB) elliptic antennas with useful bandwidth exceeding 10:1 ratio are suitable for wideband radar, wireless ultra wideband (UWB) and other wireless communication applications. We present wideband PCB elliptic dipole antennas which are capable of achieving the bandwidth requirements for all the applications. A set of elliptic dipole antennas with varying eccentricities have been fabricated for demonstration. We find one specific size (specific eccentricity) dipole that can yield an impressive 1.5-16 bandwidth exceeding the currently available. A couple of elliptic dipole antennas suitable for UWB application have been presented. We have measured swept frequency response, impedance and radiation patterns of all dipoles. An empirical formula is given for calculating the starting resonant frequency within the operating band. The calculated values are found in good agreement with measured results.

Keywords: Elliptic Dipole Antenna, Wideband Radar Antenna, Printed-Circuit Board (PCB), Ultra Wideband (UWB)

1. Introduction

The printed-circuit elliptical antenna has proven to be an efficient and effective radiator with broadband performance. In its simplest configuration, circular or elliptical antenna can be designed to produce a broad beamwidth as well as broadband with linear polarization and a radiation pattern having a broadside maximum. The most direct approach to provide a broad beamwidth and broadband performance from such an antenna is to use a printed-circuit dipole with the upper and lower radiation elements having a circular or elliptical shape. The antenna dipole could be fed with a 50 ohm micro-stripe line, extending into the dipole center (the point where the two adjacent circular or elliptical radiators join). It was found that the current on the radiator at all frequencies is largely concentrated on the peripheral edge with very low current density approaching inward towards the center. For elliptic (circular) dipoles, one can picture that numerous semi-elliptic thin-line dipoles of varying lengths are effectively formed to excite multi-linear modes hence resulting in a wide bandwidth with linear polarization.

Several methods have been proposed to study the impedance of elliptical printed-circuit antennas [1–3]. A printed crescent patch antenna [4] and a bottom fed elliptical antenna [5] were investigated experimentally to provide broadband performance with linear polarization without added complexities inherent in the feed circuit.



(a) (b)

Figure 1. (a) Geometry of an elliptical printed-circuit dipole antenna. (b) An elliptical dipole antenna etched on PCB with dielectric constant ε_r =4.2.

Broadband printed-circuit elliptical dipole antennas covering 750 MHz to 6.0 GHz for WLAN and WiMax applications have been fabricated and tested [6]. Recently, Powell [7] has shown that broadband linear polarization could also be achieved by two differential crescent patches fed at the dipole center of the two adjacent elliptic elements. They achieve a broadband performance covering the 3.1-10.6 GHz ultra wideband (UWB) spectrum with a swept frequency return loss of about 11dB.

In this paper, experiments are carried out to investigate the impedance bandwidth and swept frequency measurement for several elliptic dipole radiators using various eccentricities (equivalently, various b/a ratios, see Figure 1). The radiation patterns are also measured.

The rest of the paper is organized as follows. Section 2 presents experimental results of the bandwidth performance for a set of elliptic dipoles of various sizes (various eccentricities) by measuring the frequency return loss. An optimum dipole size is found to yield widest bandwidth among all. Section 3 discusses the starting resonant frequency. Section 4 discusses the antenna products for UWB application. Then, Section 5 measures the radiation patterns of the optimum dipole. Finally, Section 6 concludes the paper.

2. Experimental Results of Bandwidth Performance

The geometry of a printed-circuit elliptic antenna dipole is shown in Figure 1(a). A photograph of an elliptical dipole antenna etched on PCB with dielectric constant of 4.2 is shown in Figure 1(b). In the lower elliptic radiator, portion of the area is cut off in the shape of an ellipse to accommodate the 50 ohm micro stripe feed line. The first part of our experiment study is to investigate the broadband properties. A set of dipoles with varying *b/a* ratios were etched on the printed-circuit board (PCB). We varied the *b/a* ratio progressively from 1.00 (a circle), 0.945, 0.897, 0.852 to 0.813. The minor diameter 2*b* was held constant at 26 mm, while the major diameter 2*a* was progressively increased from 26 mm to 32 mm. Thus, five different size dipoles are etched on microwave printed-circuit boards (PCBs) using 41.2 mm×88.1 mm to 38.1 mm×53.4 mm FR4 with thickness *d* = 0.762 mm and dielectric constant 4.2.

The swept frequency return loss for an elliptic dipole with 2*a*=32 mm and 2*b*=26 mm *(b/a* = 0.813) is first measured using a network analyzer. We can see from Figure 2 that, in the 1.5 GHz to 16 GHz range (bandwidth ratio 10.66:1), the return losses are all better than -10 dB. This is an impressive broadband result as the bandwidth has exceeded the 10:1 ratio. Next, Figure 3 presents the swept frequency measurement of this particular dipole in Smith chart format. Then, a plot of the real and imaginary parts of the input impedance against frequency is given in Figure 4. From this figure, we can see multiple resonance peaks indicating that the radiator effectively consists of multi-elliptically-shaped thin-line dipoles of various lengths exciting many linear modes thus resulting in a broad bandwidth with linear polarization.

Next, we vary 2*a* from 26.0, 27.5, 29.0, and 30.5 mm while holding 2*b* constant at 26 mm (*b/a* ratios of 1.000, 0.945, 0.897, and 0.852) and repeat the experiment. For these *b/a* ratios, the measured swept frequency return losses are compared and results are presented in Figure 5. Typical swept frequency input impedance derived from corresponding Smith chart measurements are given in Figure 6 and Figure 7. The first dipole with 2*a*=32mm outperforms all. Then, we have also fabricated elliptic dipoles with reducing *b/a* ratios of 0.800 and 0.750 (increasing 2*a* beyond 32mm) and found that the performance also starts to fall.

The bandwidth performance (bandwidth is defined here as the frequency range with return loss better than

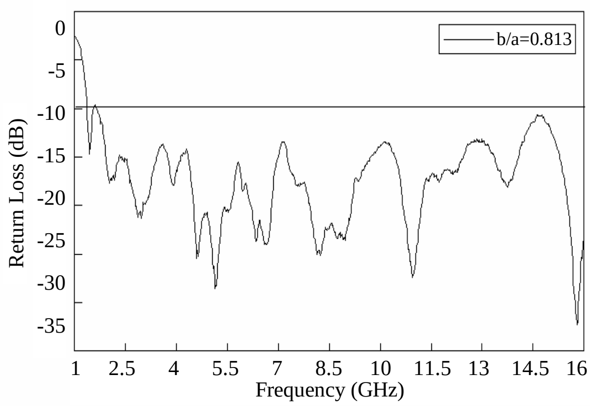


Figure 2. Measured return loss for elliptic dipole with 32 mm × 26 mm, *b/a* = 0.813.

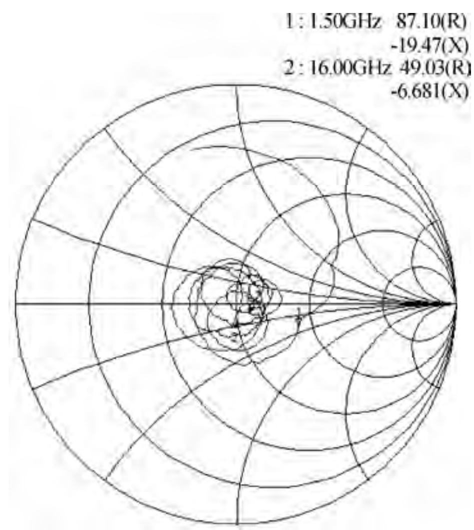


Figure 3. Smith chart display for elliptic dipole with 32 mm × 26 mm, *b/a* = 0.813.

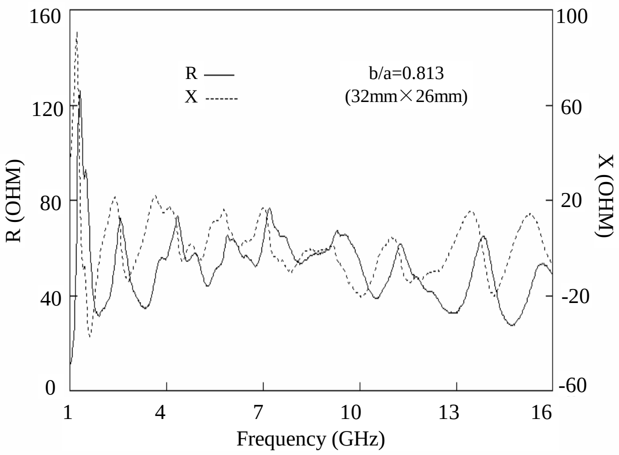


Figure 4. Impedance vs. frequency for elliptic dipole with 32 mm × 26 mm, *b/a* = 0.813.

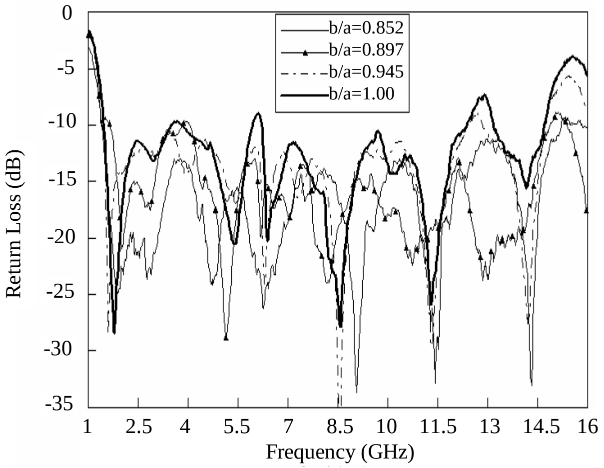


Figure 5. Comparison of measured return loss between elliptic dipoles (*b/a* = 0.852, 0.897, 0.945, and 1.000).

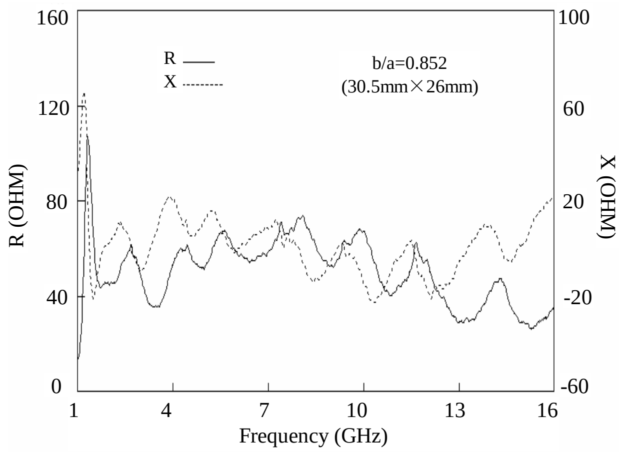


Figure 6. Impedance vs. frequency for elliptic dipole with 30.5 mm × 26 mm, *b/a* = 0.852.

-10 dB) versus *b/a* ratio for the above five dipoles is given in Figure 8. It is observed that as the radiator shape becomes less elliptical, the number of effective semi-elliptical-shaped thin-line dipoles of various lengths appears to decrease resulting in a narrower bandwidth.

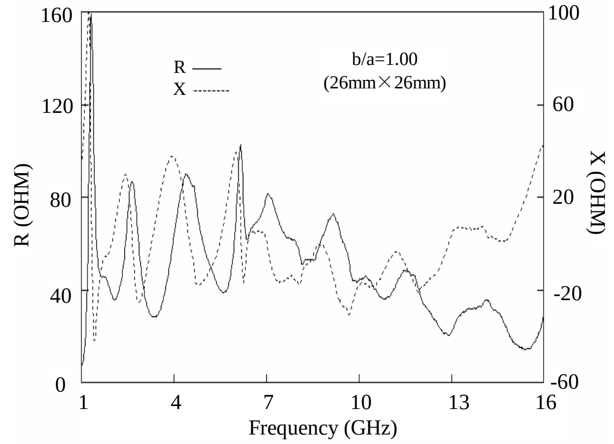


Figure 7. Impedance vs. frequency for elliptic dipole with 26 mm × 26 mm, *b/a* = 1.000.

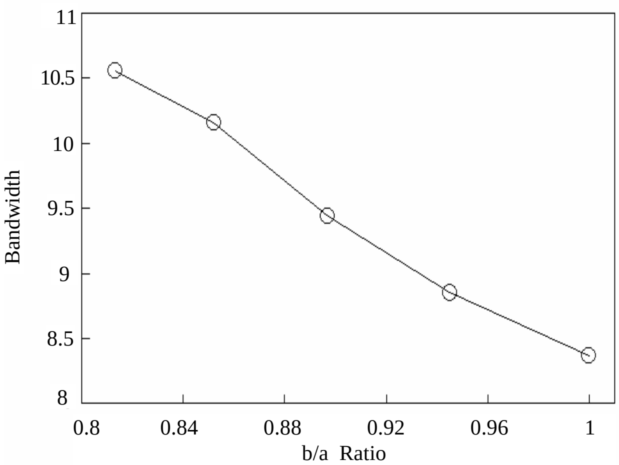


Figure 8. Bandwidth vs. *b/a* ratio for elliptic dipoles with minor diameter 26mm.

3. The Starting Frequency

The second part of our study is to investigate the starting frequency in the operating band of the elliptic dipole. It is well-known that the minor diameter 2*b* of a dipole radiator determines the resonant frequency. To investigate the resonant frequency of the elliptic dipole experimentally, we vary 2*b* from 27.0, 28.0, 29.0, and 31.0 mm while holding 2*a* constant at 32 mm (*b/a* ratios of 0.844, 0.875, 0.906, and 0.969) and repeat the experiment. For these *b/a* ratios, the swept frequency Smith charts are measured and the typical input impedances derived from the corresponding Smith charts are given in Figures 9 through 12.

For a PCB material of dielectric constant ε_r and a given 2*b* (mm), we found an empirical formula for calculating the starting resonant frequency (defined for return loss < -10 dB) in GHz as

$$f_s = \frac{40.8}{b\sqrt{\varepsilon_r}} \quad \textbf{(1)}$$

For elliptic antenna dipoles with 2*b* varied from 26.0, 27.0, 28.0, 29.0, and 31.0 mm and with 2*a* held constant at 32 mm, using (1), the calculated values of corresponding starting resonant frequencies are respectively about 1.53, 1.47, 1.42, 1.37, and 1.28 GHz. These calculated frequencies are found in good agreement with the measured frequencies (at frequencies, the impedance values correspond to better than 10 dB return loss) as shown in Figure 4 and Figures 9 through 12. Thus, (1) proves to be quite useful and handy.

4. Products for UWB Application

Utilizing Equation (1) in Section 3, printed-circuit elliptical dipole antennas for 3.1-10.6 GHz UWB applications have been fabricated and tested [8]. Figure 13 shows a photograph of elliptical dipole antennas etched on PCB with dielectric constant of 4.2 and 10.2, respectively.

To achieve the required 3.42:1.00 UWB impedance bandwidth properties, low eccentricity elliptic dipole radiators were first etched on FR4 PCB with an overall size of 24 mm × 46 mm. The measured return loss, as shown in Figure 14, in the 3.1 GHz to 10.6 GHz range is generally better than -12.6 dB. Figure 15 presents the swept frequency measurement in Smith chart format. Next, to reduce he size of the product, an elliptical dipole antenna of the same design is etched on a flexible laminate PCB with a thickness d=0.635 mm and ε_r =10.2. With an overall PCB size of 15 mm × 28 mm, this elliptic dipole provides suitable impedance properties across major portions of the frequency spectrum. The swept frequency return loss of this elliptic dipole fabricated on the flexible laminate PCB is presented in Figure 16. Due to the non-uniform properties of the flexible laminate PCB, this antenna can only achieve close to -10 dB return loss performance in the specified 3.1 GHz to 10.6 GHz frequency band.

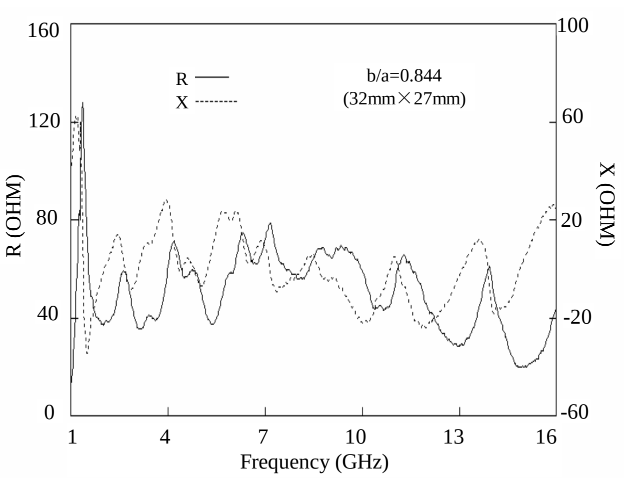


Figure 9. Impedance vs. frequency for elliptic dipole with 32 mm × 27 mm, *b/a* = 0.844.

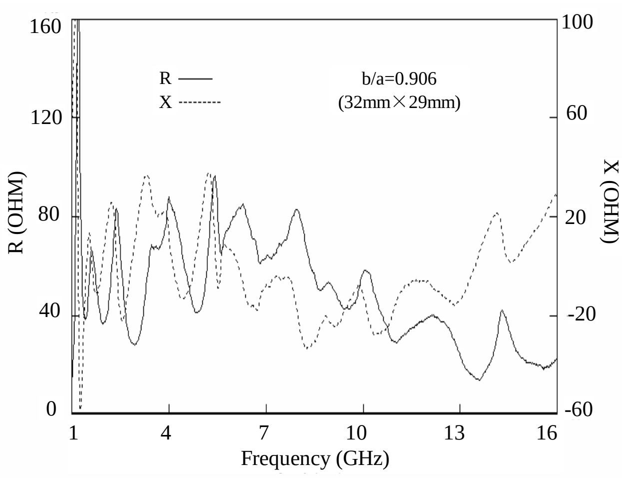


Figure 11. Impedance vs. frequency for elliptic dipole with 32 mm × 29 mm, *b/a* = 0.906.

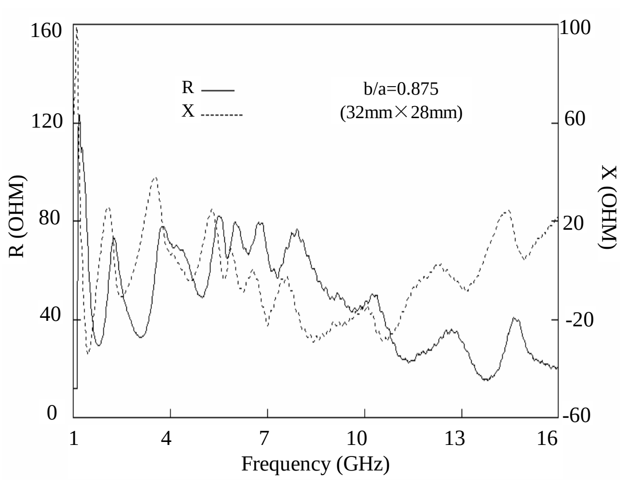


Figure 10. Impedance vs. frequency for elliptic dipole with 32 mm × 28 mm, *b/a* = 0.875.

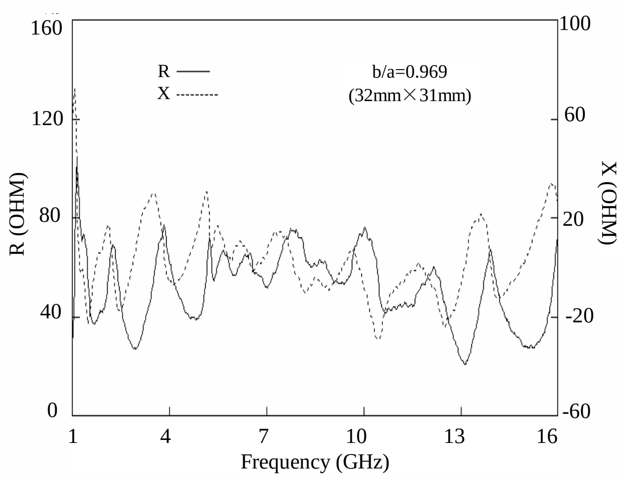


Figure 12. Impedance vs. frequency for elliptic dipole with 32 mm × 31 mm, *b/a* = 0.969.

Different from in wideband radar applications, for UWB impulse radio application, antenna requires sufficient impedance matching, linear ungroup phase response or near constant group delay throughout the entire 3.1 to 10.6 GHz band. As shown in Figure 13, the presented antennas are small, compact, and should exhibit fixed phase center property. Owing to that, this antenna tends to radiate a mostly non-dispersive waveform which cause less pulse shape distortion to the transmitted waveform and provides suitable frequency domain characteristics and performance. For UWB applications, we are preparing equipment to perform the time domain transmission tests required to assess the impulse response and fidelity characteristics of these antennas.

5. Radiation Patterns

The third part of our experiment is to investigate the radiation patterns of the 1.5-16 GHz elliptic dipole. For demonstration purpose, we only present the measurements for our optimum dipole (32 mm × 26 mm, b/a = 0.813). The radiation patterns on the x-z, x-y, and y-z planes of the optimum elliptic dipole at 2.0, 4.0, 10.0, and 14.0 GHz are measured in an anechoic chamber and shown in Figure 17, Figure 18, and Figure 19 respectively. Results indicate reasonable omnidirectional radiation patterns on all the three planes. Consistency of the patterns, similar to a typical dipole radiation pattern, can be observed across major portions of the frequency band. The measured gain is about 2 dBi.

6. Conclusions

For elliptic dipole antennas with 2b varied from 26.0, 27.0, 28.0, 29.0, and 31.0 mm and with 2a held constant at 32 mm, we found an optimum size 32 mm × 26 mm, b/a = 0.813 can yield a maximum bandwidth performance in the 1.5-16 GHz range. Thus, we have

Figure 13. Elliptical dipole antennas etched on PCB with dielectric constant ε_r = 4.2 (Left) and 10.2 (Right).

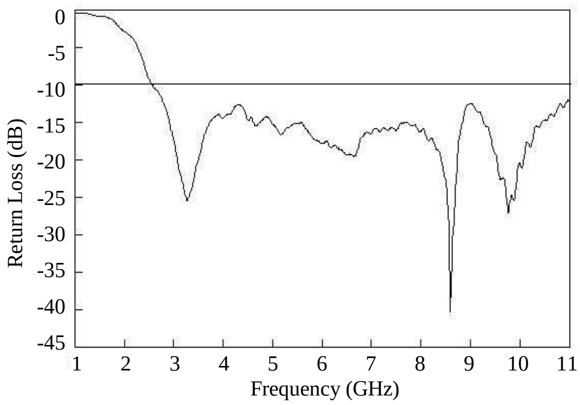


Figure 14. Measured return loss of the (24 mm × 46 mm) elliptic dipole etched on FR4 with ε_r = 4.2.

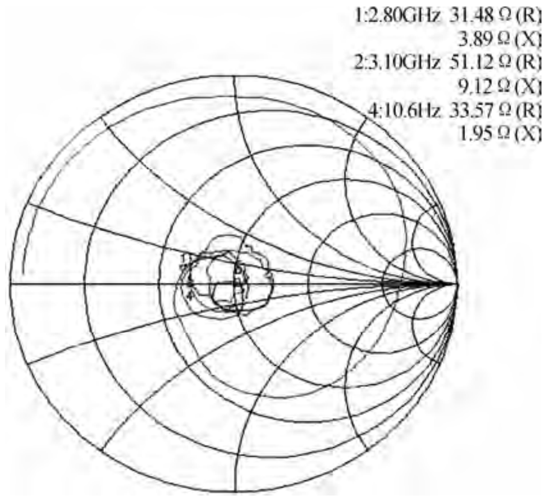


Figure 15. Smith chart display of the (24 mm × 46 mm) elliptic dipole etched on FR4 with ε_r = 4.2.

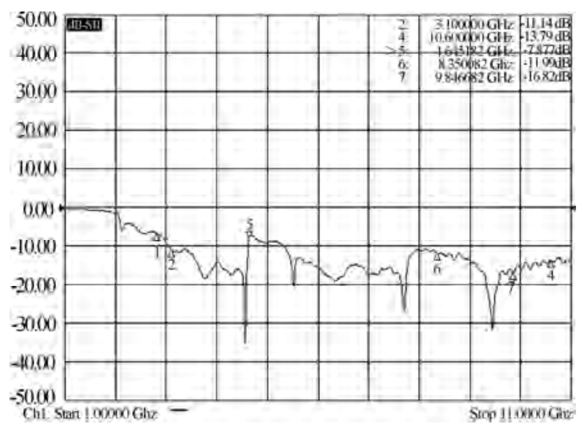


Figure 16. Measured return loss of the (15 mm × 28 mm) elliptic dipole etched on a flexible laminate PCB with ε_r = 10.2.

shown that, with a proper choice of minor to major axis ratio, a printed-circuit elliptic dipole antenna using a simple single-feed network can provide a useful operating bandwidth exceeding the 10:1 ratio. Aside from swept frequency and impedance measurements for elliptic dipoles of various *b/a* ratios, the radiation patterns for the optimum dipole are measured and found to be in consistency with typical dipole radiation patterns. By properly choosing the minor diameter for the dipole made of PCB of known dielectric constant, the starting operating frequency can be easily calculated using an empirical formula for system design. A couple of elliptic dipole antennas for 3.1-10.6 GHz (3.42:1 bandwidth ratio) UWB application have been presented to demonstrate that these dipole antennas can be designed for other fixed operating frequency bandwidth ratio.

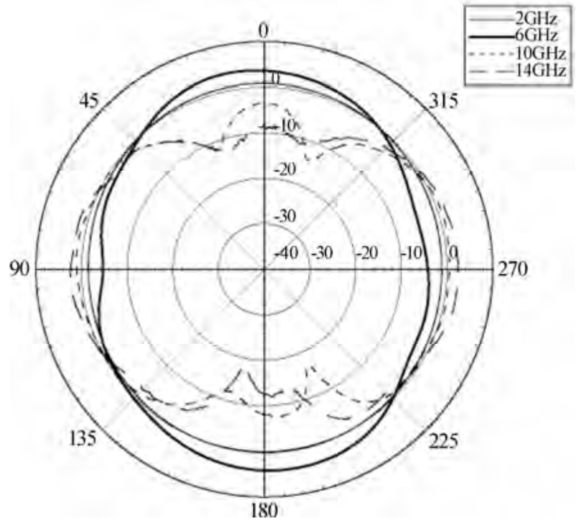


Figure 17. Measured radiation pattern on x-z plane of antenna (32 mm × 26 mm, *b/a* = 0.813).

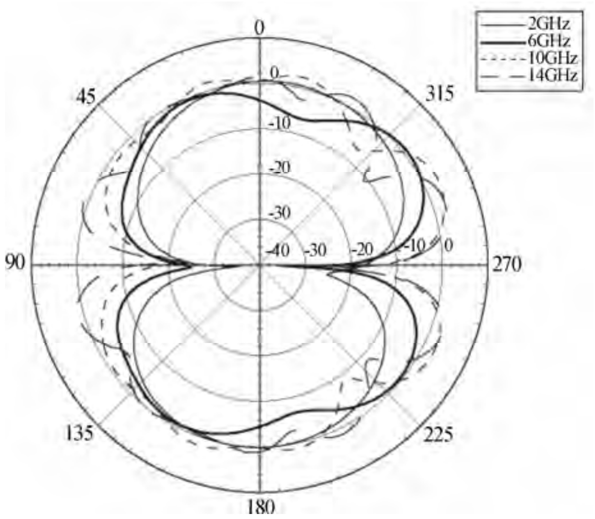


Figure 18. Measured radiation pattern on x-y plane of antenna (32 mm × 26 mm, *b/a* = 0.813).

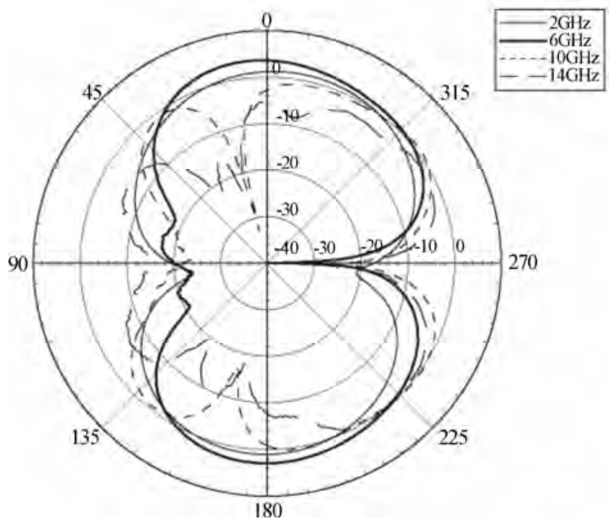


Figure 19. Measured radiation pattern on y-z plane of antenna (32 mm × 26 mm, *b/a* = 0.813).

7. References

[1] L. C. Shen, "The elliptical microstrip antenna with circular polarization," IEEE Transactions on Antenna and Propagation, Vol. AP–29, No. 1, January 1981.

[2] S. A. Long and M. W. McAllister, "The impedance of an elliptical printed-circuit antenna," IEEE Transactions on Antenna and Propagation, Vol. AP–30, No. 6, November 1982.

[3] S. A. Long, L. C. Shen, D. H. Schaubert, and F. G. Farrar "An experimental study of the circular-polarized elliptical printed-circuit antenna," IEEE Transactions on Antenna and Propagation, Vol. AP–29, No. 1, January 1981.

[4] N. C. Azenui and H. Y. D.Yang, "A printed crescent patch antenna for ultrawideband applications," IEEE Antennas and Wireless Propagation Letters, Vol. 6, 2007.

[5] H. Schantz, " Apparatus for establishing signal coupling between a signal line and an antenna structure," U. S. Patent 6, 512, 488, January 28, 2003.

[6] C. C. Lee, C. W. Wang, R. Y. Yen, and H. S. Huang, "Broadband printed-circuit elliptical dipole antenna covering 750 MHz-6.0 GHz," International Conference on Microwave and Millimeter Wave Technology Proceedings, Nanjing, China, Vol. 3, pp. 1207–1209, April 2008.

[7] J. Powell and A. Chandrakasan, "Differential and single ended elliptical antennas for 3.1-10.6 GHz ultra wideband communication," 2004 IEEE Antennas and Propagation Society Student Paper Competition, 1st Place.

[8] C. C. Lee, H. S. Huang, R. Y. Yen, C. D. Yang, and S. C. Nan, "Printed-circuit elliptical dipole antenna for 3.1-10.6 GHz UWB application," The 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computering, October 2008, Dalian, China.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Analysis of a Cyclic Multicast Proxy Server Architecture

Olatunde ABIONA[1], Tricha ANJALI[2], Clement ONIME[3], Lawrence KEHINDE[4]
[1]*Department of Computer Information Systems, Indiana University Northwest, Garry, IN 46408, USA*
[2]*Electrical and Computer Engineering Department, Illinois Institute of Technology, Chicago, IL 60616, USA*
[3]*Information and Communication Technology Section, Abdus Salam International Centre for Theoretical Physics, I-34014 Trieste, Italy*
[4]*Department of Engineering Technologies, Texas Southern University, Houston, Texas 77004, USA*
E-mail: [1]oabiona@iun.edu, [2]tricha@ece.iit.edu, [3]onime@ictp.it, [4]kehindelo@tsu.edu
Received September 6, 2008; revised October 7, 2008; accepted October 10, 2008

Abstract

The exponential growths of the World Wide Web (WWW) users have made the deployment of proxy servers popular on a network with limited resources. WWW clients perceive better response time, improved performance and speed when response to requested pages are served from the cache of a proxy server, resulting in faster response times after the first document fetch. This work proposes cyclic multicast as a scalable technique for improving proxy server performance for next generation networks. The proposed system uses a cyclic multicast engine for the delivery of popular web pages from the proxy server cache to increasingly large users under limited server capacity and network resources. The cyclic multicast technique would be more efficient for the delivery of highly requested web pages from the cache to large number of receivers. We describe the operation of the cyclic multicast proxy server and characterized the gains in performance.

Keywords: Caching, Cyclic Multicast, Reliable Unicast, Scalable Data Delivery, Next Generation Network

1. Introduction

A proxy server is a server that sits between a client application, such as a web browser, and a real server. It intercepts all requests to the real server to see if it can fulfill the requests by itself. Otherwise, it forwards the request to the real server. Proxy servers have two main purposes on a network, firstly, to improve network performance through the delivery of previously request objects from the cache and secondly to filter request, i.e. preventing users from accessing some specific sets of website. Proxy caching has been widely used to cache static (text/image) objects on the Internet so that subsequent requests to the same objects can be served directly from the proxy server without contacting the real server.

In order to reduce client perceived access latencies as well as server/network loads, caching of frequently used data at proxies close to the client is an effective technique. This will enhance the availability of objects and reduce packet losses, as local transmission is more reliable than remote transmission.

Proxy caching has become one of the vital components in all web systems. Streaming media, in particular those prestored can have significant performance improvements from proxy caching, due to their static nature in content. Hence proxy servers have found useful applications in media streaming, video on demand and large scale multimedia applications [1–5].

Over the last several years, the WWW has gained tremendous popularity; similarly, the number of WWW users on the Internet has grown exponentially. Making the system administrator to continually battle with ways of improving response times due to large volumes of users' request. Different approaches have been used to solve the problem of scalability; one of such approaches is to simply buy more powerful hardware to upgrade the servers. This is not a cost effective or scalable solution as this approach may fail with increasing WWW users. Another solution is the improvement of the Hyper Text Transport Protocol (HTTP) to reduce the latency associated with HTTP communication by removing overhead of creating a new connection for each HTTP request [6]. Another solution is replicating transparent servers' at the most popular websites [7], caching of hot pages [8] and multicast delivery [9].

The focus of this work is to investigate cyclic multicast architecture for the delivery of WWW pages to increasingly large numbers of user given limited server capacity and network resources for next generation networks. Access pattern to files follows a Zipf-like distribution [10]. Access to website typically follows a skewed pattern, namely; small number of popular pages (hot pages) accessed very frequently, a large number of warm pages accessed with moderate frequency and a large number of cold pages accessed a few times or rarely. We explore the cyclic multicast for the transmission of popular (hot and warm) requested web pages and reliable unicast for other (cold) pages. With this option, web pages are delivered to multiple requesting clients using a single server response based on the network support for point to multipoint communication.

The cyclic multicast option is expected to be more efficient for the delivery of highly requested web pages (hot and warm pages) to large number of users. With this option, the web page is broken into chunks, cyclically transmitted and clients can listen at any point in time in the transmission, and continue to listen until all of the data is received.

The rest of the paper is organized as follows. In section 2 we review related work and in section 3 we discuss the architecture of a cyclic multicast proxy server. In section 4 we present the operation of the cyclic multicast proxy server and in section 5 the analysis of cyclic multicast. In section 6 we present the simulation of the server and in section 7 we discuss the result of our performance analysis. The paper concludes in section 8.

2. Related Work

Large popular files can be delivered efficiently from a server to several clients concurrently using multicast or broadcast. Some previous work has shown the use of multicast to provide scalable services [9,11–15]. Some other applications of multicast for the delivery of information, news to large audience and general data base access were described in [14,16,17]. The use of multicast support within the Internet has been largely tied to the delivery of videoconferencing, audio, video and streaming of multimedia applications to large recipients. In this work, we propose the User Datagram Protocol (UDP) best effort multicast for the delivery of popular pages to large numbers of receivers, with repetitive, cyclic transmission of the requested page to ensure reliability. This solution is expected to be scalable and more efficient when used for the delivery of the same content to large numbers of requesters or receivers.

3. The Basic Design of a Cyclic Multicast Proxy Server

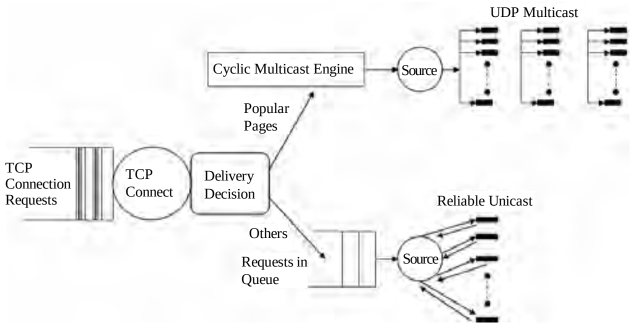


Figure 1. Cyclic multicast proxy server architecture.

Figure 1 shows the basic design of a cyclic multicast proxy server. This server is capable of delivering web pages using cyclic multicast and reliable unicast. When a request for Transmission Control Protocol (TCP) connection arrives from a client for a page, the requests are queued until the server can process them. When the request is about to be processed, the server establishes a TCP connection and the client request is transmitted.

A delivery decision is made based on the popularity of the requested page. There are two possible options for the delivery of a page, cyclic multicast or reliable unicast. The most popular (hot) pages and moderately popular (warm) pages are served using cyclic multicast engine, while other unpopular (cold) pages are served using the traditional TCP unicast connections. The decision for the pages to serve using cyclic multicast is based on the document hit rate of the server. The cyclic multicast operation involves a number of processes ranging from chunking of the page, joining a multicast group to receive all the chunks correctly in one or more cycles and finally leaving the multicast group after receiving all chunks correctly. This architecture is further described in detail in the following section.

4. The Cyclic Multicast Proxy Server Operation

The proposed cyclic multicast engine built in the proxy server is an effective way to deliver most popular and heavily requested pages. The cyclic multicast engine is capable of delivering multiple pages simultaneously using multiple multicast groups; however, in this work we only describe the use of this engine to deliver a single page to several receivers. For the delivery of multiple pages simultaneously, the operation is replicated for every page intended for delivery using cyclic multicast. The cyclic multicast delivery scheme may be summarized in the following steps similar to [18].

- The page including embedded files is divided into a number of chunks.
- A multicast address is used to transmit all chunks in a page sequentially from the server to the group of receivers. A cycle is the transmission of all chunks in a page.

- A receiver joins an appropriate multicast group and remains a member until all chunks are received correctly. If a receiver misses a chunk, the receiver must remain in the group until the missed chunk is re-transmitted and received correctly in subsequent cycle.
- The server (cyclic multicast engine) continues the cyclic transmission if there is at least one receiver in the group waiting to receive the page. The server stops transmission when all members of the group have finished receiving the page.

5. Analysis of a Cyclic Multicast Engine

We use the following analysis to compare the performance of cyclic multicast with reliable unicast. We assume our web page is broken into *C* equal-size chunks and those transmissions out of the server are in packets for both unicast and cyclic multicast with each packet containing one chunk. We assume that *N* receivers make requests for same page at the same time. If the probability that a packet (chunk) is received correctly is *P* and losses are independent of packet and receivers, for reliable unicast, U_c the number of packets (chunks) that will be transmitted such that all *N* receivers get all chunks making up a page is given by:

$$U_C = \frac{N * C}{P} \qquad \textbf{(1)}$$

Similarly, for cyclic multicast, the server will continue to cycle through the chunks until all *N* receivers get all chunks. Assume is the number of cycles required, M_c the packets (chunks) transmitted is given by:

$$M_C = C * \alpha \qquad \textbf{(2)}$$

Since all *N* receivers make their request at the same time, they will all be waiting to start receiving just before the transmission of the first cycle. We use a discrete time Markov chain to represent the behavior of the system. The chunks represent a state and a page has *K* chunks. Figure 2 shows the Markov chain for one user receiving all *K* chunks correctly.

Similarly, a discrete time Markov chain may be used to represent how a chunk is received by all N receivers. Let the state of the system represent the number of receivers that have received a particular chunk. If *m* receivers have

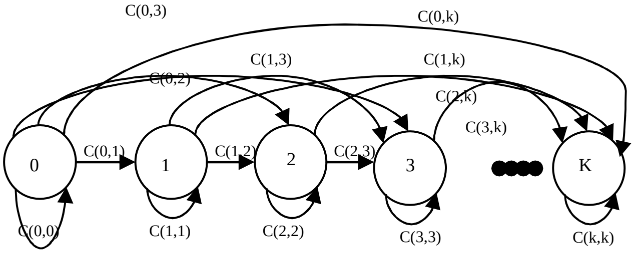


Figure 2. Markov chain for all chunks.

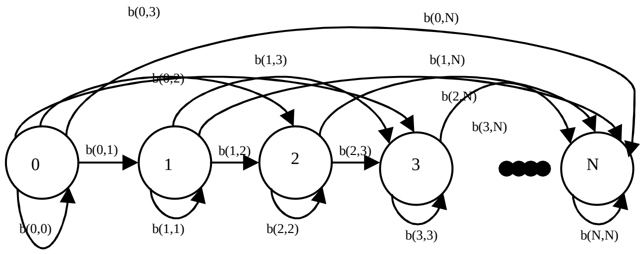


Figure 3. Markov chain for receiving one chunk.

received a particular chunk at the end of cycle *t*, then the probability of *m+n* receivers receiving the chunk at the end of cycle *t+1* will be:

$$b_{(m,m+n)} = \binom{N-m}{n} P^n * (1-P)^{(N-m-n)} \qquad \textbf{(3)}$$

The Markov chain for all N receivers receiving one chunk is shown in figure 3. Equation 3 gives the transition between states; the end state is reached when all N receivers have received the chunk correctly.

The transition probability matrix is given by:

$$B = \begin{bmatrix} 1 & b_{01} & b_{02} & b_{03} & \bullet & \bullet & \bullet & b_{0N} \\ 0 & 1 & b_{12} & b_{13} & \bullet & \bullet & \bullet & b_{1N} \\ 0 & 0 & 1 & b_{23} & \bullet & \bullet & \bullet & b_{2N} \\ 0 & 0 & 0 & 1 & \bullet & \bullet & \bullet & b_{3N} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & 0 & 0 & \bullet & \bullet & \bullet & 1 \end{bmatrix}$$

Let $P(k,t)$ be the probability that *k* receivers have already received a particular chunk at the end of cycle *t*. Then

$$P(k,t) = \sum_{i=0}^{k} b_{(i,k)} * P(i,t-1) \qquad \textbf{(4)}$$

The following initial conditions apply to $P(k,t)$:

$P(i,j)=1$ *for* $i=j$, $(i=0,1,2,3...k)$

$P(i,0)=0$ *for* $(i=1,2,3\ldots k)$

$P(i,j)=0$ *for* $i \geq 2, j=i-1$

If $P_{RCVD}(N,C,t)$ represent the probability that all *N* receivers have received all *C* chunks at the end of cycle *t*, and we assume independence of loss then,

$$P_{RCVD}(N,C,t) = [P(N,t)]^C \qquad \textbf{(5)}$$

By computing $P_{RCVD}(N,C,t)$ in Equation 5 we obtain the minimum number of cycles, α for the delivery of all *C* chunks to all *N* receivers with increasing values of *t*.

6. Simulation of the Cyclic Multicast Proxy Server Architecture

In this section, we present the results of simulations for the cyclic multicast proxy server which supports cyclic multicast and reliable unicast. Our objective is to provide a comparison in performance based on throughput, response time, end-to- end delay and jitters experienced by clients using a cyclic multicast proxy server and how it compares with clients using a caching or unicast proxy server.

For our simulation we used the ns-2 [19] network simulator. We assume a large number of clients making request which follow a Poisson process and each server will have N pages with all pages of same size. We assume that access pattern follows a Zipf distribution. We use the time to transmit all the chunks that make up a page (time for one cycle) as our time unit. We also assume that there is no propagation delay in making a request and in receiving chunks from the server and that the list of popular pages are know. The reliable unicast server is capable of transmitting streams out of the server using selective reject Automatic Repeat Request (ARQ) protocol, while the first request to the cyclic multicast server for a page is used to start the cyclic multicast engine.

The experiment scenario is shown in figure 4. From Figure 4 the centre node 0 is the server surrounded by client's node 1 to node 8 receiving packets from the server. Each client node (node 1 to node 8) has a link speed of 1Mbps with a delay of 10ms and a drop tail buffer to the central server node 0. There is a TCP/FTP flow from the server node 0 to all the eight clients' nodes. A complete page (all chunks making up the page) can be transmitted in one cycle to all clients receiving transmission from the server using cyclic multicast, but for unicast a cycle is the time to transmit all chunks to a single client. We conducted the experiment for unicast i.e. each node receiving transmission from the server one node at a time and for cyclic multicast which allows several nodes to receive transmission from the server at the same time. For the cyclic multicast, we also vary the time a client joins the multicast group using joining time of (*t=0.2s and t=0.5s*) in our simulation.

7. Performance of a Cyclic Multicast Proxy Server

Throughput is one of the performance parameters studied in our simulation. Throughput is defined as the amount of data (bits) that can be sent in a unit time. For the graph below, our time interval length (TIL) is 0.1s.

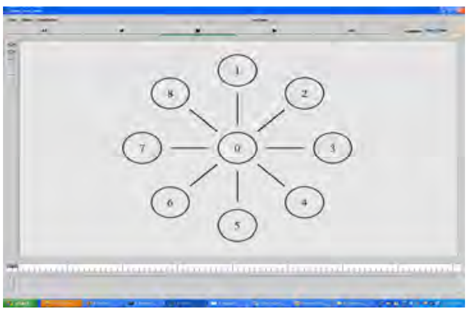


Figure 4. Cyclic multicast proxy server simulation scenario.

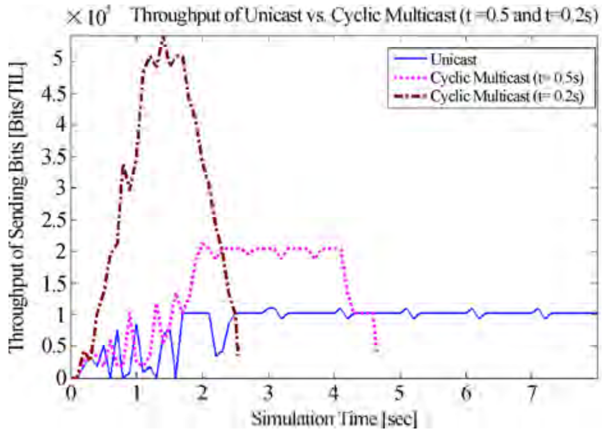


Figure 5. Throughput comparison.

Figure 5 shows the comparison of throughput for unicast and cyclic multicast servers. The throughput achieved by the unicast server was about 10,000Mbps while the throughput of the cyclic multicast server with a joining time t=0.5s was about 20000Mpbs. Reducing the joining time to t=0.2s allowed more clients to join the multicast group, further increasing the throughput to about 50,000Mbps.

This results shows a better performance by the proxy server when the number of receivers in a multicast group increases.

Similarly, we studied the response time. Response time is the time it takes to completely receive a page. In Figure 5 we can see a better response time for the cyclic multicast server.

The response time reduces as more clients join the multicast group to receive a page. For the unicast, the response time for all clients to completely receive a page is 8s. For the cyclic multicast with joining time t=0.5s the response time is 4.5s, while for cyclic multicast with joining time t=0.2s the response time is about 2.5s. Hence the load on the server is zero for cyclic multicast *(t*=0.2*)* after 2.5s and cyclic multicast *(t*=0.5*)* after 4.5s since there are no more receivers waiting to receive a page.

End-to-End delay is another performance parameter considered. End-to-End delay is defined as the time taken for a packet to be transmitted across a network from source to destination. From Figure 6, the End-to-End delay experienced by the unicast proxy server increases with the simulation time, while the End-to-End delay for the cyclic multicast reduces as the simulation time increases, showing again that the cyclic multicast proxy server out performs the unicast server. Similarly, for cyclic multicast, the server load drops to zero after 4.5s since there are no more requests to serve by the proxy after the last receiver exits the multicast group.

Jitter is another performance parameter considered. Jitter is an unwanted variation of signal characteristics. Jitter may be defined as the variation in the delay. Figure 7 shows the comparison of Jitter for unicast and cyclic multicast. The cyclic multicast experienced less jitters, indicating lower variation in the delay of packets.

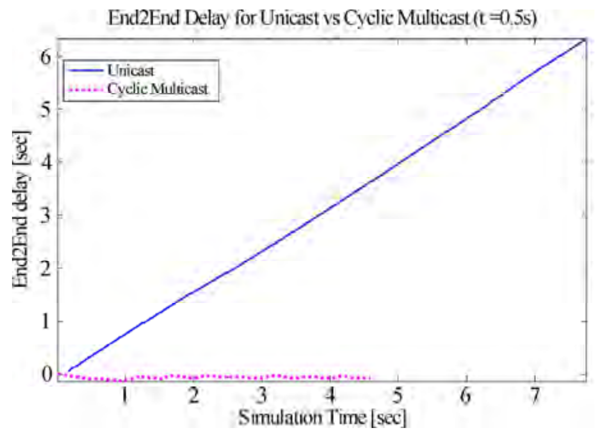


Figure 6. End-to-end delay comparison.

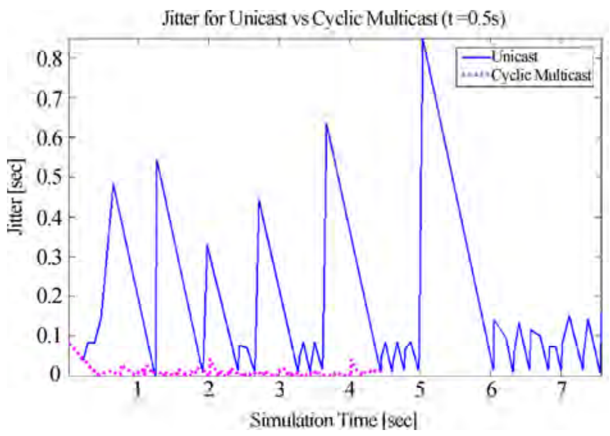


Figure 7. Jitter comparison.

We also plot the Cumulative Distribution Function (CDF) for End-to-End delay and jitter for unicast and cyclic multicast. Figure 8 shows the CDF for delay in unicast and cyclic multicast proxy server.

$$F_X(x) = P(X \le x)$$

For unicast proxy server,

$\Pr(delay \le 3) \approx 0.7$

For cyclic multicast server,

$\Pr(delay \le 0.1) \approx 0.7$

This shows that the cyclic multicast proxy server performs better than the unicast proxy server with respect to end-to-end delay.

Figure 9 shows the CDF for jitter in unicast and cyclic multicast server.

$F_X(x) = P(X \le x)$

For unicast proxy server,

$\Pr(jitter \le 0.1) \approx 0.7$

For cyclic multicast server,

$\Pr(jitter \le 0.01) \approx 0.7$

Again Figure 9 shows that the cyclic multicast proxy server performs better than the unicast proxy server with respect to jitter.

8. Conclusions

In this paper, we propose a proxy server based on the cyclic multicast for next generation networks, as a scalable delivery option for the delivery of web pages to increasingly large number of users under limited server capacity and network resources. Our proposed solution uses a cyclic multicast engine attached to the proxy server to deliver a popular page using UDP multicast with reliability achieved through repetitive, cyclic multicast transmission of a requested page. This solution is expected to be scalable and more efficient when used for the delivery of the same content to large numbers of receivers. Our simulation results show the performance gains achievable with this technique. Our result also shows that the performance of a proxy server can be further enhanced by integrating both delivery options in the proxy server for the next generation networks. A practical implementation of the cyclic multicast proxy server with squid [20], and detailed analysis of the behavior of the cyclic multicast engine using a discrete time Markov chain will be considered for future work.

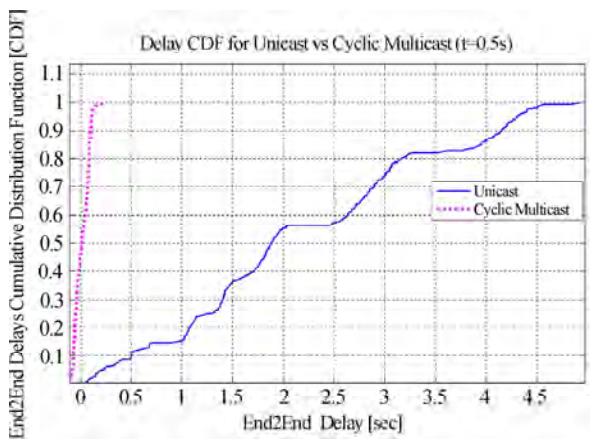


Figure 8. Delay CDF comparison.

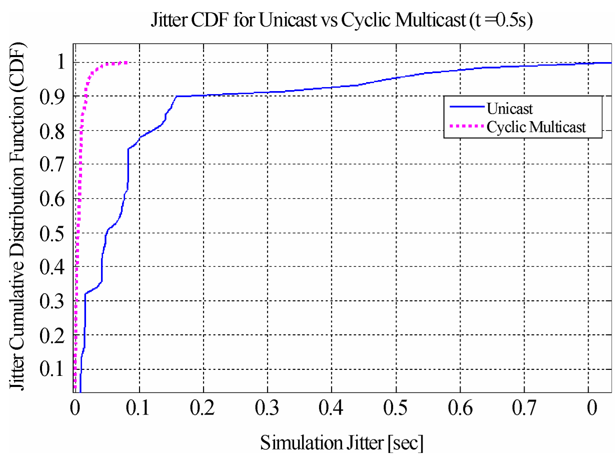


Figure 9. Jitter CDF comparison.

9. References

[1] Z. Miao and A. Ortega, "Scalable proxy caching of video under storage constraints," IEEE Journal on Selected Areas in Communications, Vol. 20, pp. 1315–1327, September 2002.

[2] Z. L. Zhang, Y. Wang, D. Du, and D. Su, "Video staging: a proxy-server-based approach to end-to-end video delivery over Wide-Area Networks," IEEE/ACM Transactions on Networks, Vol. 8, No. 4, pp. 429–442, August 2000.

[3] H. Fahmi, M. Latif, S. Sedigh-Ali, A. Ghafoor, P. Liu, and L. Hsu, "Proxy servers for scalable interactive video support," IEEE Computer, Vol. 43, No. 9, pp. 54–60, September 2001.

[4] L. Guo, S. Chen, S. Ren, X. Chen, and S. Jiang, "PROP: A scalable and reliable P2P assisted proxy streaming system," Proceedings IEEE International Conference on Distributed Computing Systems, Tokyo, Japan, March 2004.

[5] J. Liu and J. Xu, "Proxy caching for media streaming over the Internet," IEEE Communications, pp 88–94 August 2004.

[6] V. N. Padmanabhan and J. C. Mogul, "Improving http latency", 2nd World Wide Web Conference '94, available: http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings, October 1994.

[7] E. Katz, M. Butler, and R. McGrath, "A scalable http server: The ncsa prototype," Computer Networks and ISDN Systems, Vol. 27, pp. 155–164, 1994.

[8] H. Braun and K. Claffy, "Web traffic characterization: An assessment of the impact of caching documents from NCSA's web server," Computer Networks and ISDN Systems, Vol. 28, pp. 37–51, December 1995.

[9] R. J. Clark and M. H. Ammar, "Providing scalable web services using multicast communication," Computer Networks and ISDN Systems, Vol. 29, No. 7, pp. 841–858, 1997.

[10] L. Breslau, P. Cao, L. Fan, G. Philips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," Infocom'99, March 1999.

[11] J. W. Wong and M. H. Ammar, "Analysis of broadcast delivery in a videotext system," IEEE Transactions on Computers , Vol. 34, pp. 863–866, September 1985.

[12] J. W. Wong and M. H. Ammar, "Response time performance of videotext systems," IEEE Journal on Selected Areas in Communication, Vol. 4, pp. 1174–1180, October 1986.

[13] M. Ammar and J. Wong, "On the optimality of cyclic transmission in teletext systems," IEEE Transactions on Communications, Vol. 35, pp. 68–73, January 1987.

[14] G. Herman, G. Gopal, K. Lee, and A. Weinrib, "The datacycle architecture for very high throughput database systems," Proceedings ACM SIGMOD, 1987.

[15] S. Acharya, M. Franklin, and S. Zdonik, "Dissemination based data delivery using broadcast disks," IEEE Personal Communications, Vol. 2, pp. 50–60, December 1995.

[16] K. Lidl, "Drinking from the firehose: Multicast USENET news," in USENIX Winter '94, USENIX Association Press, pp. 33–45, January 1994.

[17] D. K. Gifford, "Polychannel systems for mass digital communications," Communications of the ACM, 33(2), pp. 141–151, February 1990.

[18] K. C. Almeroth, M. H. Ammar, and Z. Fei, "Scalable delivery of web pages using cyclic best effort multicast," Proceedings INFOCOM, pp. 1214–1221, 1998.

[19] UCB/LBNL/VINT Network Simulator - ns (version 2), 1997, available: http://www.isi.edu/nsnam/ns.

[20] D. Wessels, "Squid internet object cache," 1996, available: http://www.nlanr.net/squid.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Improving Throughput in Wireless Local Area Networks

Imam AL-WAZEDI, Anjali AGARWAL, Ahmed K. ELHAKEEM
Department of Electrical and Computer Engineering,
Concordia University, Montreal, Quebec, Canada
Email: i_alwaze@encs.concordia.ca, {aagarwal,ahmed}@ece.concordia.ca
Received September 6, 2008; revised October 16, 2008; accepted October 21, 2008

Abstract

The medium access control (MAC) technique of standard WLANs, called the distributed coordination function (DCF), is carrier sense multiple access based on collision avoidance (CSMA/CA) scheme with binary slotted exponential backoff. It has a two way handshaking technique for packet transmission and also defines an additional four way handshaking technique called RTS/CTS mechanism, which is used to reduce the hidden terminal problem. The RTS/CTS frames carry the information of the packet length to be transmitted which can be read by any listening stations, to update a network allocation vector (NAV) about the information of the period of time in which the channel is busy. In this paper a method is proposed called the table driven technique (which has two parts called table driven DCF and table driven RTS/CTS) which is similar to the standard DCF (IEEE802.11) and RTS/CTS (IEEE802.11) system without having the exponential backoff. In this technique users use the optimum transmission probability by estimating the number of stations from the traffic conditions in a sliding window fashion one period at a time, thereby increasing the throughput compared to the standard DCF (IEEE802.11) and RTS/CTS (IEEE802.11) mechanism while maintaining the same fairness and the delay performance.

Keywords: Standard DCF (IEEE802.11), Standard RTS/CTS (IEEE802.11), Table Driven DCF, Table Driven RTS/CTS

1. Introduction

Wireless local area networks (WLANs) have been widely deployed for the past decade. Their performance has been the subject of intensive research. In [1] an improvement of throughput and fairness is shown by optimizing the backoff without estimating the number of active nodes in the network. In [2], the authors proposed a MAC layer based WLAN technique in which they gave higher priority to access points so as to improve the throughput and the channel utilization. A technique is proposed where the backoff is tuned based on collision avoidance and fairness to improve the channel utilization [3]. In [5] a DCF model is proposed where the arrival and the service of the packets in the queue are controlled to improve the throughput and delay performance.

Cali in [6] pointed out that depending on the network configuration, DCF may deliver a much lower throughput compared to the theoretical limit. Cali derived a distributed algorithm that enables the stations to tune its backoff at run time where a considerable improvement in the throughput is shown. In [7] a contention based MAC protocol named fast collision resolution is presented where the backoff is also utilized. A model named DCF+ in [8] is proposed which uses the backoff to improve the fairness.

It is evident that the throughput, delay, fairness performances are improved by tuning the backoff in different scenarios considered by the authors in [1–8].

RTS/CTS mechanism with (NAV) is used solve the hidden terminal problem. In [9] Khurana proposed a concept of *Hearing graph* to model the hidden terminals in static environment and analyzed the performance. Also in [11] Fullmer, proposed a three way handshaking technique to solve the hidden terminal problems of single channel WLANs. However our paper does not concentrate on the hidden terminals but contributes on a modification of the standard DCF and standard RTS/CTS mechanisms.

In this paper table driven DCF and table driven RTS/CTS systems are proposed, which are similar to IEEE 802.11 (Both DCF and RTS/CTS) standards without, the use of the exponential backoff. In table driven DCF and table driven RTS/CTS the users estimate the number of active stations and transmit with an optimum probability measured from the traffic conditions (by sensing the channel) in a sliding window fashion, which is described elaborately later on. Simulation results show that our systems out perform the standard in terms of throughput while maintaining same delay and fairness.

2. The IEEE 802.11 MAC Protocol

Figure 1 shows one of many transmission scenarios possible with the IEEE 802.11 DCF mode. In this mode a node with a packet to transmit initializes a backoff timer with a random value selected uniformly from the range [0, CW-1], where CW is the contention window in terms of time slots. After a node senses that the channel is idle for an interval called DIFS (DCF interframe space), it begins to decrease the backoff timer by one for each idle time slot observed on the channel. When the channel becomes busy due to other nodes transmission ativity the node freezes its backoff timer until the channel is sensed idle for another DIFS. When the backoff timer reaches zero, the node begins to transmit. If the transmission is successful, the receiver sends back an acknowledgement (ACK) an interval called the SIFS. Then, the transmitter resets its CW to CW_{min}. In case of collisions the transmitter fails to receive the ACK from its intended receiver within the specified period, it doubles its CW subject to maximum value CW_{max}, chooses a new backoff timer, and starts the above processes again.

In 802.11, DCF also provides a more efficient way of transmitting data frames that involves transmission of special

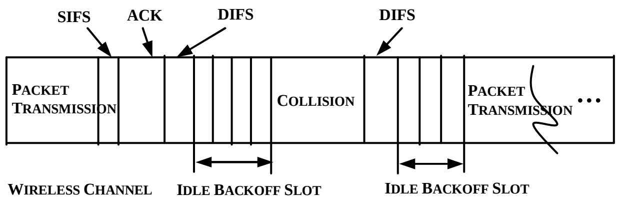


Figure 1. IEEE 802.11 MAC mechanism.

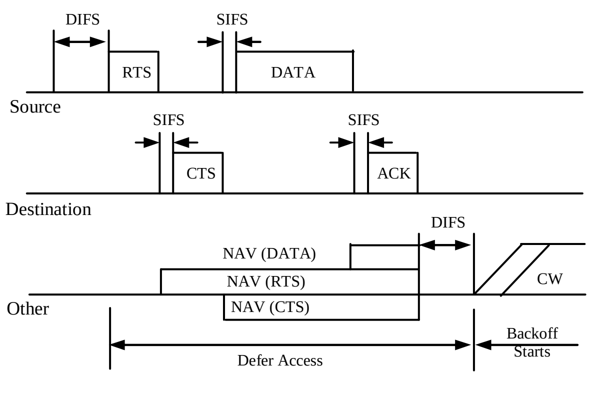


Figure 2. Standard RTS/CTS access mechanism in IEEE 802.11.

short RTS and CTS frames prior to the transmission of actual data frame. As shown in Figure 2, an RTS frame is transmitted by a node, which needs to transmit a packet. When the destination receives the RTS frame, it will transmit a CTS frame after SIFS interval immediately following the reception of the RTS frame. The source station is allowed to transmit its packet only if it receives the CTS correctly. Note that all the other stations are capable of updating their knowledge about other nodes transmission duration by receiving a certain field in RTS, CTS, ACK, and packets transmission called network allocation vector (NAV). This helps to combat the *hidden terminal* problem. In fact, a node that is able to receive the CTS frames correctly, can avoid collisions even when it is unable to sense the data transmissions directly from the source station. If a collision occurs with two or more RTS frames, much less bandwidth is wasted when compared with the situations where larger data frames in collision, thus justifying the case for RTS, CTS mode of operation [4].

3. Analysis of Table Driven DCF and Table Driven RTS/CTS

Let p be the transmission probability of each node and M be the number of active stations. Assuming each user tries to transmit randomly in each slot following the DIFS period. According to table driven DCF (Figure 3) the probability of successful transmission, is thus given by Equation (1).

$$P_s = Mp(1-p)^{M-1} \qquad \textbf{(1)}$$

The probability of an idle slot in table driven DCF is

$$P_o = (1-p)^M \qquad \textbf{(2)}$$

and the probability of unsuccessful transmission for table driven DCF is

$$P_c = 1 - P_s - P_o \qquad \textbf{(3)}$$

Let i be the number of idle periods (cycles) before a successful transmission as shown in Figure 3 and j be the number of idle slots in each idle period lengths ($W_1, W_2, \ldots$). The throughput (η_1) is given by Equation (7) for table driven DCF.

It is easily seen that the average length of each idle period except the last one before packet success in table driven DCF is given by

$$W_1 = W_2 = \ldots\ldots W_{L-1} = E(j) = \sum_{j=1}^{\infty} j(P_o)^j P_c$$

$$= \frac{P_o P_c}{(1-P_o)^2} \text{ slots} \qquad \textbf{(4)}$$

which means Equation (4) determines the number of idle slots before a collision.

The last idle period before a success, has an average of

$$W_L = \frac{P_o P_s}{(1-P_o)^2} \text{ slots} \qquad \textbf{(5)}$$

The average number of cycles (I) is given by,

$$I = \sum_{i=1}^{\infty} iP_i$$

where $P_1(No.\,of\;\;cycles = 1) = P_s + P_0P_s + P_0^{\,2}P_s + ...$

$$= \frac{P_s}{1-P_0}$$

$$P_2(No.\,of\;\;cycles = 2) = P_c\left[\frac{P_s}{1-P_0}\right] + P_0P_c\left[\frac{P_s}{1-P_0}\right] + P_0^{\,2}P_c\left[\frac{P_s}{1-P_0}\right].......$$

$$= \frac{P_cP_s}{(1-P_0)^2}$$

$$P_2(No.\,of\;\;cycles = 3) = \frac{P_c^{\,2}P_s}{(1-P_0)^3}$$

$$P_i = \frac{P_c^{\,i-1}P_s}{(1-P_0)^i}$$

Therefore

$$I = \frac{(1-P_o)}{P_s} \quad \textbf{(6)}$$

Let the number of collisions be $C= I–1$. This C and W_L are calculated from different values of M and p which are stored in two different tables (not shown for space consideration). So for particular values of M and P there is a particular value of C and W_L. The throughput for table driven DCF (η_1) and table driven RTS/CTS (η_2) are given in Equations (7) and (8) respectively based on the transmission activity on the wireless channel as shown in Figure 3.

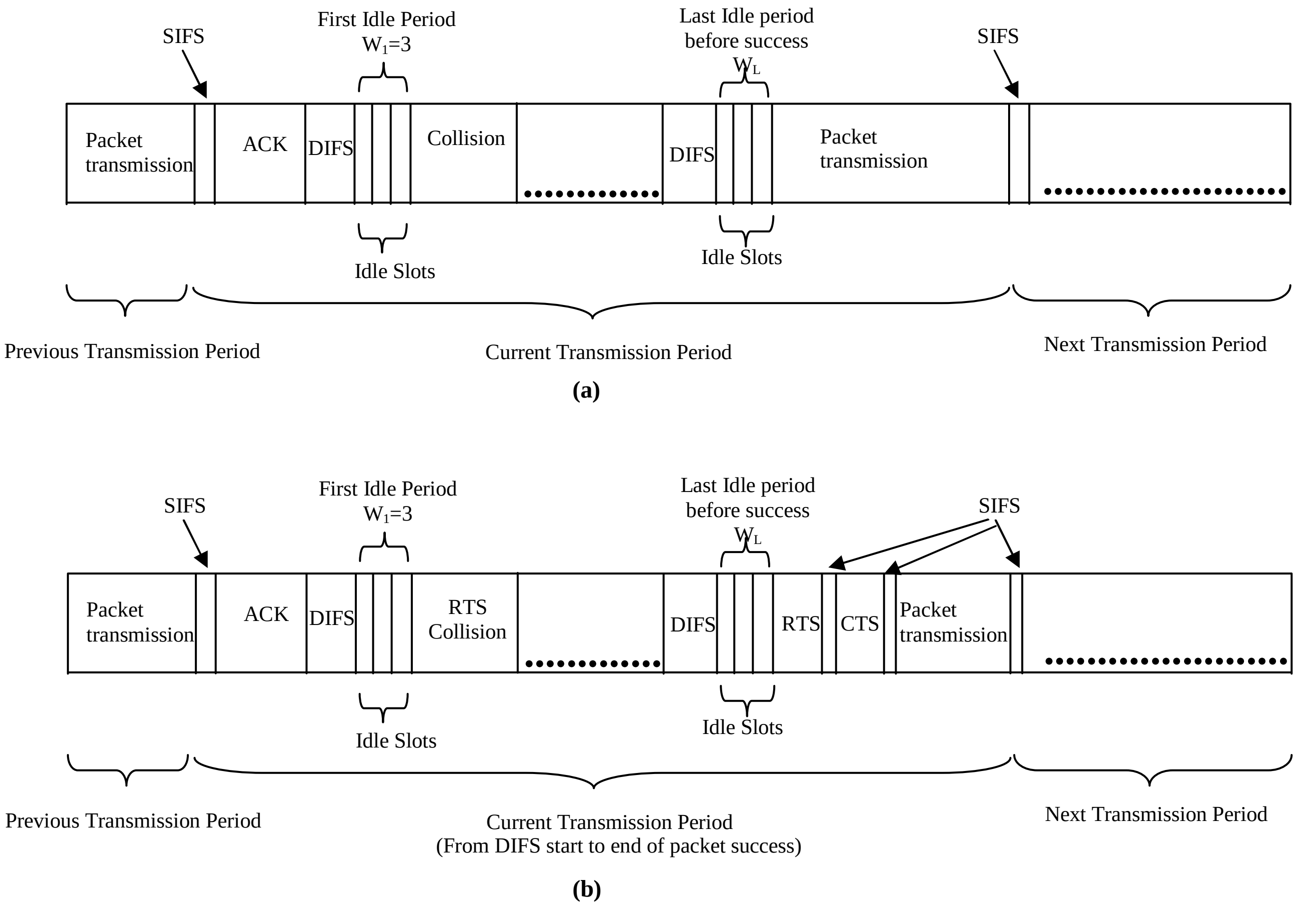


Figure 3. Transmission Activity on the Wireless Channel for (a) table driven DCF and (b) table driven RTS/CTS.

$$\eta_1 = \frac{T_{Payload}}{(W_1 + W_2 + ..W_{L-1})T_{Slot} + (I-1)\{T_{DIFS}\} + T_{DIFS} + T_{ACK} + T_{SIFS} + T_{Payload} + W_LT_{Slot}} \quad \textbf{(7)}$$

$$\eta_2 = \frac{T_{Payload}}{(W_1 + W_2 + ..W_{L-1})T_{Slot} + (I-1)\{T_{RTS} + T_{DIFS} + T_{SIFS}\} + T_{DIFS} + T_{RTS} + T_{CTS} + T_{ACK} + 3T_{SIFS} + T_{Payload} + W_LT_{Slot}} \quad \textbf{(8)}$$

The throughput η_1 (for table driven DCF) is calculated for different values of M and p as in Figure 4. Table 1 depicts the probabilities at which the maximum throughput occurs for different values of M.

Similarly for the table driven RTS/CTS, to calculate the C and W_L, Equations (1)–(6) are used. However the throughput is calculated from Equation (8) which includes the RTS/CTS frames (Figure 3).

For the case of table driven RTS/CTS all cycles leading to no success (RTS heard but no CTS) will each have a cost of $T_{RTS}+T_{DIFS}+T_{Slot}$ seconds.

4. Operation of Table Driven Technique

In the proposed protocol, if the nodes sense that the channel is idle for an interval called DIFS (DCF interframe space), they try to send a packet with a probability p, which is dependent on the traffic condition i.e. the number and activities of the nodes, as follows.

A user continuously monitors the channel in each idle slot following the DIFS idle period. If the previous slot is idle, it calls a uniform random generator (0,1). If the value

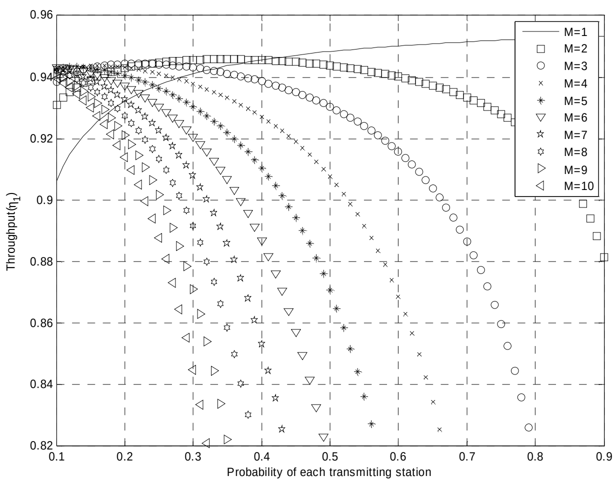


Figure 4. Throughput for different probabilities and different number of stations for table driven DCF.

Table 1. Optimum throughput for different probabilities and different number of stations for table driven DCF.

No of Stations	Probability	Optimum throughput
1	0.9000	0.9532
2	0.3400	0.9458
3	0.2200	0.9444
4	0.1600	0.9437
5	0.1300	0.9434
6	0.1000	0.9431
7	0.1000	0.9428
8	0.1000	0.9420
9	0.1000	0.9410
10	0.1000	0.9397

of this generator is less than or equal to p, it tries to start its packet transmission in the given next slot. If the value is larger than p, the user persist on listening and repeats trials as stated. However if the channel is sensed busy the user defers his transmission till the next DIFS idle period heard.

The users measure the number of collisions $C= I–1$ and the length W_L by monitoring the channel over a large enough window (which is explained latter on). With the help of these values the users can use the tables formulated in Section 3 to obtain the corresponding p and M.

Users having a non-empty queue start by monitoring the channel for the first n transmission periods. This active user will average the length of the idle period preceding the correct packet transmission over n transmission periods, i.e. $\overline{W_L}$ and $\overline{C}$, the average number of collisions over the same period. Aided with these values the users obtain the operating values of p and M and uses p to control their activities for the head of line packet in their queue. Active users continuously monitor the channel and use a sliding window technique to estimate W_L and I and hence obtain (M, p). For example the first sliding window averages W_L and C of the first n transmission periods. The second window averages W_L and C of the l=2,3,...,n+1 transmission periods. The sliding window averaging process reflects the changing traffic, so transmission activity of active users are dependent only on the current traffic and not on past history.

It is possible that the tables relating (M, p) to (W_L, I) yield more than one possibility for (M, p) for certain (W_L, I) measurement values from the sliding window. In this case, the user averages the obtained values of M and use Table 1 to find the optimum p at this averaged value of M. This Table 1 is obtained from Figure 4 in an evident manner. The operation of this table driven technique is similar to the DCF standard (IEEE 802.11) [4] except for using this optimized transmission probability p. The active users just estimate (M, p) from the traffic conditions (by sensing the channel) in a sliding window fashion one period after another.

We note that old and new users both measure the traffic and adapt to the same traffic conditions fairly and obtain the same p. However having same p does not mean all users will repeatedly collide in the same slot because of feeding a random number generator with p.

The above procedure is followed for both table driven DCF and table driven RTS/CTS shown in Figure 3.

5. Simulation Results

For numerical calculations the following parameters are taken from “Bianchi” in [4].

In the table driven DCF, as per the standards, following the observance of each DIFS, users try to transmit with probability p obtained from $\overline{W_L}$ and $\overline{C}$. If two or more stations try to transmit at the same time, collisions occur. If no stations transmit (Figure 3), the number of idle slots will increase. If one station is successful after certain

$T_{Payload}$	*10msec*
PHYheader	*128bits*
ACK	*112bits+PHY header*
RTS	*160bits+PHY header*
CTS	*112bits+PHY header*
Channel bit rate	*1 Mbits/s*
Slot time (T_{Slot})	$50\ \mu s$
T_{SIFS}	$28\ \mu s$
T_{DIFS}	$128\ \mu s$
service rate	$\mu = \dfrac{1}{T_{payload}}$
offered traffic	λ *packets/sec*
	$M\lambda \le \mu$
No of stations	*M*

number of idle and collision periods, the transmission period ends. As a result the total time for one successful packet transmission include T_{DIFS}, T_{SIFS} T_{Idle}, $T_{Payload}$. The throughput is calculated at the end of the simulation at certain values of M, λ and p i.e. ,

$$\eta = \frac{T_{Payload} \times No\ of\ Transmission\ Periods\ in\ the\ whole\ simulation}{Time^{(n)}}$$

where $Time^{(n)}$ is the total simulation time that depends on T_{DIFS}, T_{SIFS}, T_{Slot}, $T_{Payload}$. Initially $Time^{(n)} = T_{DIFS}$ and is subsequently increased based on the user's activity, e.g.,

$$Time^{(n)} = Time^{(n)} + T_{Slot}\ ;\ for\ each\ idle\ slot\ period$$
$$Time^{(n)} = Time^{(n)} + T_{DIFS};\ for\ each\ collision$$
$$Time^{(n)} = Time^{(n)} + T_{DIFS} + T_{SIFS} + T_{Payload};$$
$$for\ each\ successful\ packet$$

For the Table driven RTS/CTS the total simulation time is calculated by the following equations,

$$Time^{(n)} = Time^{(n)} + T_{Slot}\ ;\ for\ each\ idle\ slot\ period$$
$$Time^{(n)} = Time^{(n)} + T_{RTS} + T_{DIFS};\ for\ each\ collision$$
$$Time^{(n)} = Time^{(n)} + T_{RTS} + T_{CTS} + T_{DIFS} + 3T_{SIFS} + T_{Payload};$$
$$for\ each\ successful\ packet$$

Figure 5 shows a comparison of the throughput between the table driven DCF and the standard DCF (IEEE 802.11) for 10 stations. The values of standard DCF are taken from [5] which uses the same parameters as in [4]. It is evident the table driven DCF performs better than the standard DCF (IEEE 802.11).

Figure 6 shows a comparison of average delay between the table driven DCF and the standard DCF (IEEE 802.11). The values of standard DCF are again taken from [5]. It is noticeable that the delay performances are the same.

Figure 7 shows the throughput curve for different offered loads for the table driven RTS/CTS technique. It shows that the throughput rises and becomes saturated at higher values of the load. The maximum throughput calculated by "Bianchi" in [4] for the standard RTS/CTS (IEEE 802.11) mechanism is 0.837281 when M=10.

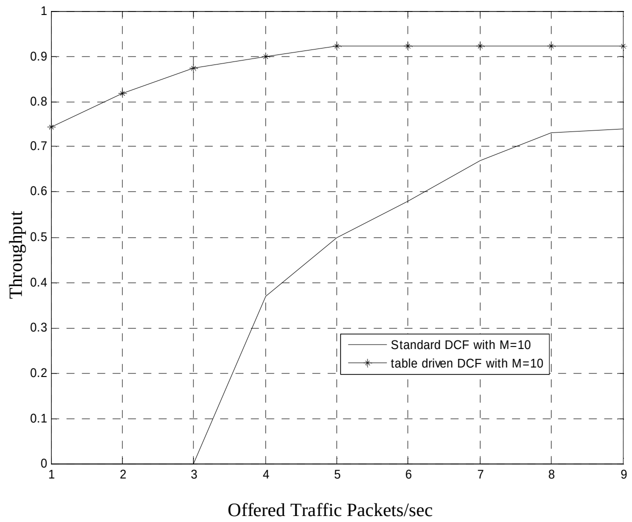


Figure 5. Throughput comparison between the table driven DCF and the standard DCF (IEEE 802.11).

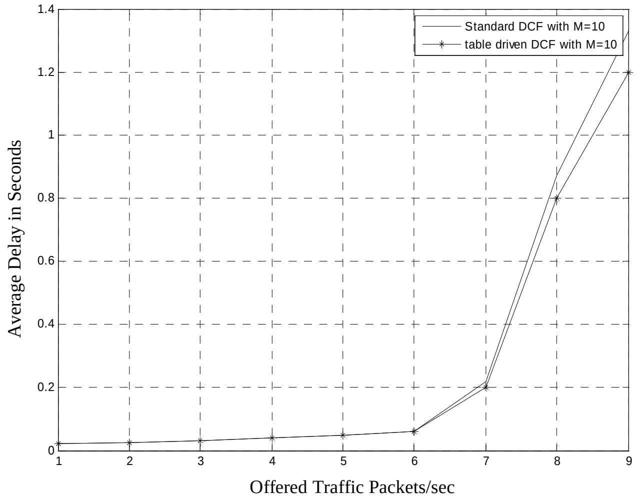


Figure 6. Average Delay Comparison between the table driven DCF and standard DCF (IEEE 802.11).

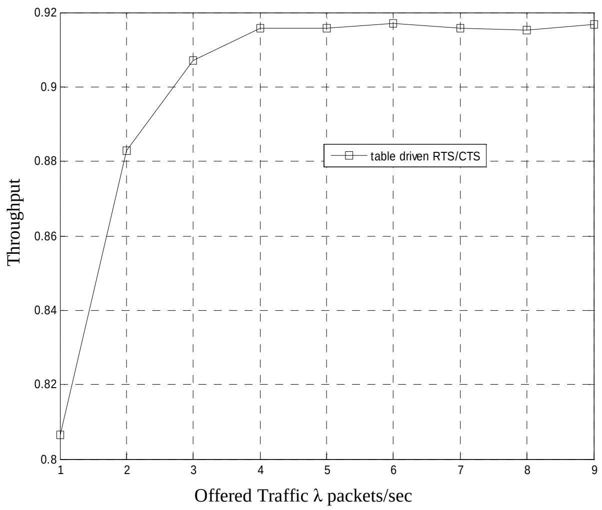


Figure 7. Throughput corresponding to different offered traffic for table driven RTS/CTS.

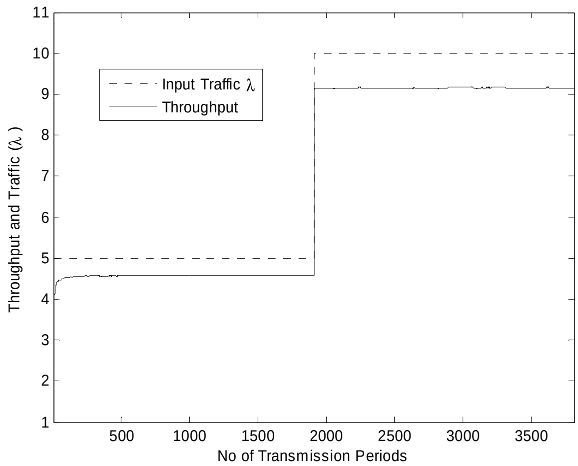


Figure 8. Throughput and input traffic corresponding to the number of transmission periods (table driven RTS/CTS).

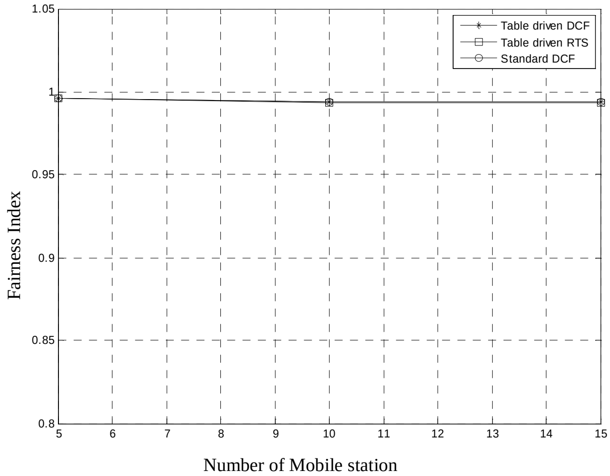


Figure 9. Fairness index for different number of stations for table driven DCF, table driven RTS/CTS and standard DCF (IEEE 802.11).

From the Figure 7 it is evident that table driven RTS/CTS performs better than the standard RTS/CTS in terms of throughput.

The table driven RTS/CTS technique has an extra advantage as it is a load adaptive system. It means that it has the capability to adapt to the input traffic as quickly as possible. Figure 8 shows a case where the input traffic suddenly increases from 5 packets/sec to 10 packets/sec. In this case the throughput $(\eta \times Input\ traffic\ rate(\lambda))$ is shown to follow the offered traffic λ.

Fairness (FI) is another important issue considered in this paper. To express this, we take the fairness index defined in [10] and [2] to measure the fair packet capacity allocation. In [10] fairness index is represented as $\frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^r}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)}$. For example if m dollars are to be distributed among n people and we favor k people by giving them m/k dollars each and discriminate against n-k people, then the above function becomes $FI = \left(\frac{k}{n}\right)^{r-1}$. Favoring 10% would result in a fairness index of $FI = 0.1^{r-1}$ and discriminate index of $1-0.1^{r-1}$. Therefore r should be equal to 2. That is, $FI = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n\left(\sum_{i=1}^{n} x_i^2\right)}$, where FI is the fairness index, n is the number of stations, x_i is the packets transmitted by the i^{th} active station during the simulation time (current traffic in which the offered traffic λ is same for all stations).

Figure 9 shows the comparison of the fairness index performance of table driven DCF, table driven RTS/CTS and standard DCF (IEEE 802.11). It can be observed that, for the three cases up to 15 active stations the performance is fair.

6. Conclusions and Future Work

In this paper a new approach that is based on the table driven technique is proposed for DCF and RTS/CTS mechanism in WLANs. While maintaining the same delay the table driven DCF outperforms the standard DCF (IEEE 802.11) in terms of throughput. The table driven RTS/CTS also demonstrates that its throughput is more than the standard RTS/CTS mechanism. Moreover the table driven DCF and table driven RTS/CTS gives very good fairness performance. In the table driven technique (for both DCF and the RTS) a simple search mechanism is used to find the values of M and p from $\overline{W_L}$ and $\overline{C}$. However an efficient lookup mechanism is required for this purpose.

The subnet technique presented in this paper is amenable to implementation with two hops or more from the SS (subscriber stations) of the IEEE802.16. This SS is typically aware of the number of the nodes of the subnet of the IEEE802.11 standard and will broadcast such number to the nodes of the IEEE802.11 subnet. The nodes may use this value as rule of thumb against the actual estimated value of M obtained by the new table driven technique. This current paper estimates p and M from the nodes activities on the channel. Further, the value of M from SS and its favorable consequences are for future research. Our table based technique shall hence improve the performance of such subnet.

7. References

[1] X. J. Tian, X. Chen, T. Ideguchi, and Y. G. Fang, "Improving throughput and fairness in WLANs through dynamically optimizing backoff," IEICE Transactions on Communications 2005 E88-B(11), pp. 4328–4338.

[2] L. Zhang, Y. T. Shu, and O. Yang, "Performance improvement for 802.11 based wireless local area networks,"

IEICE Transactions on Communications, Vol. E90-B, No. 4, April 2007.

[3] V. Bharghvan, "Performance evaluation of algorithms for wireless medium access," IEEE International Computer Performance and Dependability Symposium IPDS'98, pp. 86–95, 1998.

[4] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," IEEE Journal of Selected Areas on Communications, Vol. 18, No. 3, pp. 535–547, March 2000.

[5] S. Pasupuleti and D. Das, "Throughput and delay evaluation of a proposed-DCF MAC protocol for WLAN," IEEE INDIA annual conference 2004, lndicon 2004.

[6] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," IEEE/ACM Transactions on Networking, Vol. 8, No. 6, pp. 785–799, December 2000.

[7] Y. Kwon, Y. Fang, and H. Latchman, "A novel MAC protocol with fast collision resolution for wireless LANs," IEEE INFOCOM'03, April 2003.

[8] J. Weinmiller, H. Woesner. J. P. Ebert, and A. Wolisz, "Analyzing and tuning the distributed coordination function in the IEEE 802.11 DFWMAC draft standard," Proceedings MASCOT, San Jose, CA, February 1996.

[9] S. Khurana, A. Kahol, S. K. S. Gupta, and P. K. Srimani, "Performance evaluation of distributed co-ordination function for IEEE 802.11 wireless LAN protocol in presence of mobile and hidden terminals," Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1999, Proceedings, 7th International Symposium, Digital Object Identifier 10.1109/MASCOT.1999.805038, pp. 40–47, October 24–28, 1999.

[10] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Technical Report TR-301, DEC Research Report, 1984.

[11] C. L. Fullmer and J. J. Garcia-Luna-Aceves, "Complete single-channel solutions to hidden terminal problems in wireless LANs," Communications, 1997, ICC 97 Montreal, 'Towards the Knowledge Millenniu', 1997 IEEE International Conference, Vol. 2, Digital Object Identifier 10.1109/ICC.1997.609856, pp. 575–579, 8–12 June 1997.

[12] H. S. Chhaya and S. Gupta, "Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol," ACM/Baltzer Wireless Networks, Vol. 3, pp. 217–234, 1997.

I. J. Communications, Network and System Sciences, 2008, 4, 285-385
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

An Active Rule Approach for Network Intrusion Detection with Enhanced C4.5 Algorithm

L Prema RAJESWARI, Kannan ARPUTHARAJ
College of Engineering, Guindy, Anna University, Chennai, India
E-mail: jlprema@cs.annauniv.edu, kannan@annauniv.edu
Received February 28, 2008; revised August 20, 2008; accepted August 29, 2008

Abstract

Intrusion detection systems provide additional defense capacity to a networked information system in addition to the security measures provided by the firewalls. This paper proposes an active rule based enhancement to the C4.5 algorithm for network intrusion detection in order to detect misuse behaviors of internal attackers through effective classification and decision making in computer networks. This enhanced C4.5 algorithm derives a set of classification rules from network audit data and then the generated rules are used to detect network intrusions in a real-time environment. Unlike most existing decision trees based approaches, the spawned rules generated and fired in this work are more effective because the information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found. The main advantage of this proposed algorithm is that the generalization ability of enhanced C4.5 decision trees is better than that of C4.5 decision trees. We have employed data from the third international knowledge discovery and data mining tools competition (KDDcup'99) to train and test the feasibility of this proposed model. By applying the enhanced C4.5 algorithm an average detection rate of 93.28 percent and a false positive rate of 0.7 percent have respectively been obtained in this work.

Keywords: Decision Tree, Intrusion Detection, KDD Cup Dataset, Enhanced C4.5

1. Introduction

To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be preprocessed and converted to the format suitable for mining processing. Next, the reformatted data will be used to develop a clustering or classification model. The classification model can be rule-based, decision-tree based, and association-rule based. Classification approach can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection.

1.1. Intrusion Detection System

Intrusion Detection Systems (IDS) have been used to monitor network traffic thereby detect if a system is being targeted by a network attacks such as a DoS, Probe, U2R and R2L. The two main intrusion detection techniques are misuse detection and anomaly detection. Misuse detection systems, for example, IDIOT [1] and STAT [2], use patterns of well known attacks or weak spots of the system to match and identify known intrusions. For example, a signature rule for the "guessing password attack" can be "there are more than four failed login attempts within two minutes".

The existing intrusion detection methods [3,4] like misuse detection and anomaly detection are generally incapable of adapting detection systems to the change of circumstances resulting in high false positive rate. The most popular ways to detect intrusions are by using audit trail data generated by operating systems. An audit trail is a record of activities on a system that are logged to file in chronologically order. Manual inspection of these logs is not feasible due to incredibly large sizes of audit data generated. Therefore, data mining is used to automate the wading through audit data [5].

1.2. Classification System

Classification is similar to clustering in that it also partitions customer records into distinct segments called classes. But unlike clustering, classification analysis requires that the end-user/analyst know ahead of time how classes are defined. It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. As each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but to decide how new records should be classified. Classification is used to assign examples to pre-defined categories. Machine learning software performs this task by extracting or learning discrimination rules from examples of correctly classified data. Classification models can be built using a wide variety of algorithms. Classification categorizes the data records in a predetermined set of classes used as attribute to label each record; distinguishing elements belonging to the normal or abnormal class. This technique has been popular to detect individual attacks but has to be applied with complementary fine-tuning techniques to reduce its demonstrated high false positives rate. Classifications algorithms can be classified into three types [6] extensions to linear discrimination (e.g., multilayer perceptron, logistic discrimination), decision tree and rule-based methods (e.g., C4.5, AQ, CART), and density estimators (Naïve ayes, k-nearest neighbor, LVQ).

Decision trees are among the well known machine learning techniques. A decision tree is composed of three basic elements: – A decision node is specifying a test attribute. – An edge or a branch corresponding to the one of the possible attribute values which means one of the test attribute outcomes. A leaf which is also named an answer node contains the class to which the object belongs. In decision trees, two major phases should be ensured: 1) Building the tree. 2) Classification. This process will be repeated until a leaf is encountered. The instance is then being classified in the same class as the one characterizing the reached leaf. Several algorithms have been developed in order to ensure the construction of decision trees and its use for the classification task. The ID3 and C4.5 algorithms developed by [6,7] are probably the most popular ones.

This paper proposes an enhanced C4.5 algorithm towards developing a more-robust Intrusion Detection System through the use of data-mining techniques. Signature-based intrusion-detection systems are normally known as misuse-detection systems. Misuse-detection systems apply a rule-based approach that uses stored signatures of known intrusion instances to detect attacks. The attribute selection measure allowing to choose an attribute that generates partitions where objects are distributed less randomly. In other words, this measure should consider the ability of each attribute to determine training objects' classes. The measure is the gain ratio of Quinlan, based on the Shannon entropy, where for an attribute Ak and a set of objects T. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

These data mining techniques will dynamically model what a normal network should look like and reduce the false negative alarm rates in the process. We will use classification-tree techniques to accurately predict probable attack sessions.

The subsequent sections are organized as follows. Section 2 presents a general survey in the field of misuse detection in network intrusion detection. Section 3 describes the systems architecture of the new misuse intrusion detection and Enhanced C4.5 algorithm for generating active rules. Section 4 depicts the results and its possible implications. Section 5 gives the conclusions on this work and suggests some possible future works.

2. Literature Survey

This section discusses the related works on IDS and classification for IDS and compares them with the proposed enhanced C4.5 algorithm. There are many works in the literature that deal with classification. [1,3,5,8]. Decision tree [9] is a widely used tool for classification in various domains that need to handle large data sets. One major advantage of the decision tree is its interpretability, i.e., the decision can be represented in terms of a rule set. Each leaf node of the tree represents a class and is interpreted by the path from the root node to the leaf node in terms of a rule such as: "If A1 and A2 and A3, then class C1," where A1, A2, and A3 are the clauses involving the attributes and C1 is the class label. Thus, each class can be described by a set of rules.

Xiang et al. [8] say that intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions.

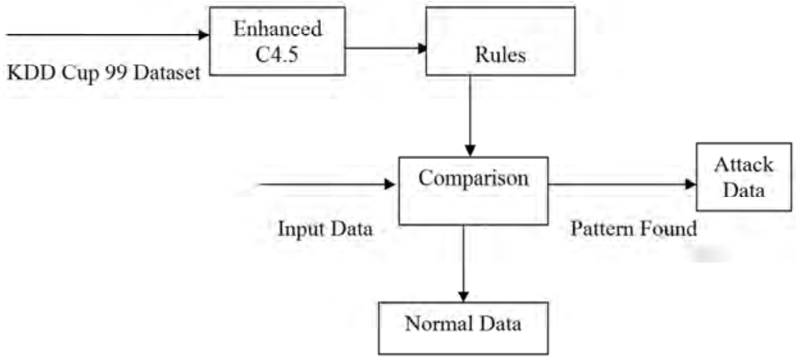


Figure 1. Misuse detection.

For accurate intrusion detection, we must have reliable and complete data about the target system activities. Similarly, routers and firewalls provide event logs for network activity. These logs might contain simple information, such as network connection openings and closings, or a complete record of every packet that appeared on the wire [10].

Decision trees are special classifiers built from instances for which the classification is known, and they are used to classify new cases taken from the same application domain [11]. Each internal node of the tree is associated to an attribute describing a feature of the dataset (domain data), and each outgoing arc is labeled with a possible value (set of values) for that feature. Each leaf node is associated to the attribute we want to predict a value for (the classification attribute), and to a value for that attribute. As stated in [12], learning algorithms based on decision trees generally adopt the divide-and-conquer strategy, i.e. they build the decision tree by recursively dividing the data set of examples into subsets according to some splitting criterion (splitting test). The splitting criterion is very important in the process of building the tree, because it determines if we must attach a node or a leaf as next element in the tree. Some of the well known splitting techniques are information Gain and Information Gain Ratio [13], Gini Criterion and Twoing rule proposed in [14].

Shavlik et al. [15] provide a framework for kdd cup 1999 dataset. The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies. To do so, a simulation is made of a factitious military network consisting of three “target” machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks.

3. System Architecture

The main task of the intrusion detection system is to discover the intrusion from the network packet data or system audit data as shown in Figure 1. One of the major problems that the IDS might face is that the packet data or system audit data could be overwhelming. Some of the features of audit data may be redundant or contribute little to the detection process. The intrusions are simulated and the detection of the intrusions is done and proper alert message is displayed.

In decision trees, two major phases should be ensured:

1) Building the tree. Based on a given training set, a decision tree is built. It consists of selecting for each decision node the ‘appropriate’ test attribute and also to define the class labeling each leaf.
2) Classification. In order to classify a new instance, we start by the root of the decision tree, then we test the attribute specified by this node. The result of this test allows moving down the tree branch relative to the attribute value of the given instance. This process will be repeated until a leaf is encountered. The instance is then being classified in the same class as the one characterizing the reached leaf.

Several algorithms have been developed in order to ensure the construction of decision trees and its use for the classification task. Using enhanced C4.5 algorithm we have generated rules. Comparison is take place for the input data. If the pattern is found we can detect the data is an attack, otherwise the data is normal.

4. Mathematical Preliminary

In this section, a brief introduction of the classification algorithms used in the i.e., the Enhanced C4.5 algorithm for building decision trees algorithm, will be given.

4.1. Enhanced C4.5 Algorithm

In a number of applications like pattern recognition, we need to classify data items into discrete set of applicable categories. A classifier, which is a function (or model) that assigns a class label to each data item described by a set of attributes, is often needed in these classification tasks.

There are quite a few machine learning approaches for generating classification models, among which decision tree learning is a typical one. As an example, C4.5 in [6] builds a decision tree where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top-most node in a tree is the root node. The tree is a model generated by the classification algorithm. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. The C4.5 algorithm builds a decision tree, from the root node, by choosing one remaining attribute with the highest information gain as the test for the current node. In this paper, Enhanced C4.5, a later version of the C4.5 algorithm, the gain ratio, expresses the proportion of useful information generated by split, i.e., that appears helpful for classification will be used to construct the decision trees. The specific algorithm is given below. The reader is referred to [7] for further details.

Algorithm: Generate a decision tree from the given training data.

Input: Training samples, represented by discrete/ continuous attributes; the set of candidate attributes, attribute-list.

Output: A decision tree

Method**:**

1) Create a node N;
2) If samples are all of the same class, C, then
3) Return N as a leaf node labeled with the class C;
4) If attribute-list is empty then
5) Return N as a leaf node labeled with the most common class in samples; (majority voting)
6) Select test-attribute, the attribute among attribute-list with the highest information gain ratio;
7) Label node N with test-attribute;
8) For each known value ai of test-attribute
9) Grow a branch from node N for the condition test-attribute = ai;
10) Let si be the set of samples in samples for which test-attribute = ai;
11) If si is empty then
12) Attach a leaf labeled with the most common class in samples;
13) Else attach the node returned by Generate_decision_tree (si, attribute-list).

Attribute Selection:

The information gain measure used in step (6) of above Enhanced C4.5 algorithm is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Existing Algorithm: Information Gain:

Let S be a set of training set samples with their corresponding labels. Suppose there are m classes and the training set contains si samples of class I and s is the total number of samples in the training set. Expected information needed to classify a given sample is calculated by:

$$I\left(S_1, S_2, ..., S_m\right) = \sum_{i=1}^{m}\left(S_i/S\right)\log_2\left(S_i/S\right) \quad \textbf{(1)}$$

A feature F with values $\{f_1, f_2, \ldots, f_v\}$ can divide the training set into v subsets $\{S_1, S_2, \ldots, S_v\}$ where S_j is the subset which has the value f_j for feature F [Zhi-xin, 2005]. Furthermore let S_j contain S_{ij} samples of class i. Entropy of the feature F is

$$E\left(F\right) = \sum_{j=1}^{V}\left(S_{1j} + ... + S_{mj}\right)\Big/ S\ X\ I\left(S_{1j}...S_{mj}\right) \quad \textbf{(2)}$$

Information gain for F can be calculated as:

$$Gain\left(F\right) = I\left(S_1, S_2, ..., S_m\right) - E\left(F\right) \quad \textbf{(3)}$$

In this experiment, information gain is calculated for class labels by employing a binary discrimination for each class. That is, for each class, a dataset instance is considered in-class, if it has the same label; out-class, if it has a different label. Consequently as opposed to calculating one information gain as a general measure on the relevance of the feature for all classes, we calculate an information gain for each class. Thus, this signifies how well the feature can discriminate the given class (i.e. normal or an attack type) from other classes.

4.2. Proposed Enhancement: Gain Ratio Criterion

The notion of information gain introduced earlier tends to favor attributes that have a large number of values. For example, if we have an attribute D that has a distinct value for each record, then *Info*(D,T) is 0, thus *Gain*(D,T) is maximal. To compensate for this, it was suggested in [6] to use the following ratio instead of gain.

Split info is the information due to the split of T on the basis of the value of the categorical attribute D, which is defined by

$$SplitInfo\left(X\right) = -\sum_{i=1}^{n}\frac{|T_i|}{|T|}\,x\log_2\left(\frac{|T_i|}{|T|}\right) \quad \textbf{(4)}$$

And the *gain ratio* is then calculated by

$$GainRatio\left(D,T\right) = \frac{Gain\left(D,T\right)}{SplitInfo\left(D,T\right)} \quad \textbf{(5)}$$

The *gain ratio*, expresses the proportion of useful information generated by split, i.e., that appears helpful for classification. If the split is near trivial, split information will be small and this ratio will be unstable. To avoid this, the gain ratio criterion selects a test to maximize the ratio above, subject to the constraint that the information gain must be large, at least as great as the average gain over all tests examined.

5. Implementation and Results

Since information gain is calculated for discrete features, continuous features should be discretized. To this end, continuous features are partitioned into equalized partitions by utilizing equal frequency intervals [16]. In equal frequency interval method, the feature space is partitioned into arbitrary number of partitions where each partition contains the same number of data points. That is to say, the range of each partition is adjusted to contain *N* dataset instances. If a value occurs more than N times in a feature space, it is assigned a partition of its own. In “10% KDD” dataset, certain classes such as denial of service attacks and normal connections occur in the magnitude of hundreds of thousands whereas other classes such as R2L and U2R attacks occur in the magnitude of tens or hundreds. Therefore, to provide sufficient resolution for the minor classes *N* is set to 10, (i.e. maximum 50,000 partitions).

Enhanced C4.5 Rules

Enhanced C4.5rules read the decision tree produced by Enhanced C4.5 and generates a set of production rules from each and from all trees together. Single Enhanced C4.5 acquires pruned decision tree with 117 nodes on train data. Total classification error rate is 47%. However, we found that Enhanced C4.5 has high classification capability for buffer_overflow and guess_passwd. Following results is a part of Enhanced C4.5 rules for buffer_overflow and guess_passwd on given 29313 training patterns.

Detecting Misuse Data

The anomaly data is detected using the rules generated. The rules are given as SQL queries and then the test data is given as input. The data are separated and stored in the different database either as normal or anomaly data. By this classification we can detect the misuse data in the testing dataset.

Table 1. Rules for training data.

Rule No	Conditions	Actions
Rule 1	num_failed_logins > 0 dst_host_same_srv_rate > 0	class guess_passwd
Rule 2	hot > 2 root_shell > 0	class buffer_overflow
Rule 3	src_bytes <= 70 dst_bytes > 5006 dst_host_same_src_port_rate > 0	class buffer_overflow

Different databases are created for different data. The network accepts the normal packets. The abnormal data are then detected and then deleted. As the anomaly data is detected in the network an alert message is given to the user of the system or administrator.

RESULTS

Tree Generation with Training Data

Enhanced C4.5 summarizes its results in a table of the following form:

Evaluation on training data (4000 items):

Before Pruning		After Pruning		
Size	Errors	Size	Errors	Estimate
1085	496(12.4%)	873	5 46(13.7%)	(26.9%)

Evaluation on test data (4000 items):

Before Pruning		After Pruning		
Size	Errors	Size	Errors	Estimate
1085	1232(30.8%)	873	1206(30.1%)	(26.9%)

Decision Tree Generation:

Subtree [S1]

diff_srv_rate <= 0.9 : portsweep < 5.0/2.3 >
diff_srv_rate > 0.9 : macspoofing < 7.0/1.3 >

Subtree [S2]

dst_host_same_src_port_rate<=0.75:others <103.0/5.0>
dst_host_same_src_port_rate>0.75: land < 4.0/2.2 >

Subtree [S3]

dst_host_src_rerror_rate<=0.5:macspoofing <14.0/1.3>
dst_host_src_rerror_rate>0.5:portsweep <4.0/1.2 >

Subtree [S4]

dst_host_src_diff_host_rate<=0.25:macspoofing <8.0/1.3 >
dst_host_src_diff_host_rate>0.25: others <4.0/1.2 >

Evaluation on training data (105131 items):

Before Pruning		After Pruning		
Size	Errors	Size	Errors	Estimate
955	74 (0.1%)	446	109(0.1%)	(0.2%)

The above rule depicts the decision tree generation. Most of this should be self-explanatory. The “Size” column gives the number of nodes in the decision tree. The “Errors” column gives the number (and percentage) of examples that are misclassified. The “Estimate” column gives the predicted error rate for new examples (this is the so-called “pessimistic” estimate, and it is computed internally by the tree algorithm). In this case, we see that the unpruned decision tree had 1,085 nodes and made 496 errors on the training data and 1,232 errors (or 30.8%) on the test data. Pruning made the tree significantly smaller (only 873 nodes) and, while it hurt performance on the training data, it slightly improved performance on the test data. The pessimistic estimate (26.9%) was actually a bit optimistic, but not too far off the mark (30.1%). You should use the error rate on the test data to plot your learning curves.

Enhanced C4.5 also prints out a confusion matrix that has one row and column for every class. The number shown in row *i*, column *j* is the number of examples that we classified into class *i* but whose true class was *j*. The perfect confusion matrix has entries along the diagonal only.

Generation of Rules from Tree

After Enhanced C4.5 has been run, then convert the decision tree into a set of rules. To execute the program, use the following command line:

C4.5rules -f stem -u >> stem.log

Enhanced C4.5rules will read the stem.names, stem.data and stem.unpruned files and append its output to the file stem.log. It will evaluate its rules on the examples in stem.test. This program can be quite slow.

This rules act as a platform for misuse detection. These rules are saved in a separate document and then they are used as SQL query. Enhanced C4.5 ules displays all of the rules and then summarizes the rule performance in the following manner

Read 105131 cases (41 attributes)

Processing tree 0

Final rules from tree 0:

Table 2. Rules for testing data.

Rule No	**Conditions**	**Actions**
Rule 10	srv_count <= 1 same_srv_rate <= 0.26 diff_srv_rate > 0.9 dst_host_srv_count < = 31	class macspoofing [99.9%]
Rule 6	srv_bytes > 0 dst_bytes <= 1 count > 9 srv_count < = 1	class macspoofing [99.5%]
Rule 32	service = eco_i src_bytes <= 26	class macspoofing [99.5%]
Rule 8	protocol_type = udp service = other src_bytes <= 17 srv_count <= 1	class macspoofing [99.2%]
Rule 16	Service = other src_bytes <= 26 srv_count <= 1 dst_host_same_src_port_rate >0.08	class macspoofing [99.2%]

The columns in the confusion matrix given below have the following meaning.

- "Rule": The number of the rule. There is one row for each rule.
- "Size": The number of tests in the rule.
- "Wrong": The number of times the rule made an error (also expressed as a percentage).
- "Advantage": A heuristic quantity used by the rule to select and prune rules.
- The class given in the conclusion part of the rule.
- "Error": The estimated error rate for this rule.

The classified results, the false positive and detect rate was obtained in confusion matrix shown in the following Table 3.

Default class: smurf

Evaluation on training data <105131 items>:

Tested 105131, errors 18343 (17.4%)

In Table 3 we see that the rules achieved an error rate 17.4% on the test data. The rules are grouped according to their output classes. Furthermore, the classes are ordered. The rules are applied in th order given. If none of the rules applies to an example, then the example is assigned to the "default" class. The top left entry in the confusion matrix shows that 6038 of the actual "normal" test examples were predicted to be normal by this entry. The last column indicates that in total 99.6% of the actual

Table 3. Confusion matrix obtained by C4.5.

Predicted / Actual	smurf	Ipsweep	nmap	neptune	portal	land	teard	% correct
smurf	72535			1				99.6
ipsweep	2							99.1
nmap	375		737	2	3			97.4
neptune	504		2	2922				93.2
portal	598			91	376		3	89.4
land	47		9			242	2	0
teard	9							
% correct	83.3	93.2	99.2	94.6	94.2	84.3	0	0

Table 4. Performance comparison of C4.5 vs enhanced C4.5.

	C4.5 Algorithm			**Enhanced C4.5 Algorithm**		
Class	Detection Rate	False Positive	False Negative	Detection Rate	False Positive	False Negative
Normal	0.9603	0.0397	-	0.9826	0.0174	-
DoS	0.906	-	0.094	0.9455	-	0.0545
Probe	0.84	-	0.16	0.8801	-	0.1199
U2R	0.8366	-	0.1634	0.8830	-	0.117
R2L	0.5376	-	0.4624	0.5580	-	0.442

normal examples were recognized correctly. The bottom row shows that 89.3% of test examples said to be normal were indeed normal in reality. From the last column, we can obtain the average detect rate of 84.9%, the false positive rate is 15.1%.

Classifying and Detecting Anomalies

Misuse detection is done through applying rules to the test data. Test data is collected from the DARPA. The test data is stored in the database. The rules are applied as SQL query to the database. This classifies data under different attack categories as follows:

1) DOS (Denial of Service)
2) Probe
3) U2R (User to Root)
4) R2L (Root to Local)

The C4.5 algorithm builds a decision tree, from the root node, by choosing one remaining attribute with the highest information gain as the test for the current node. In this paper, Enhanced C4.5, by choosing one remaining attribute with the highest information gain ratio as the test for current node is considered a later version of the C4.5 algorithm, will be used to construct the decision trees for classification. From the table 5.4 it is clear that Enhanced C4.5 outperforms the classical C4.5 algorithm. R2L attack was a bit challenging to Enhanced C4.5 but even in that case, it has reported a subtle improvement over its counter part.

Split info is the information due to the split of T on the basis of the value of the categorical attribute D, which is defined by

$$SplitInfo(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} x \log_2 \left(\frac{|T_i|}{|T|} \right) \quad \textbf{(4)}$$

And the gain ratio is then calculated by

$$GainRatio(D,T) = \frac{Gain(D,T)}{SplitInfo(D,T)} \quad \textbf{(5)}$$

In Enhanced C4.5 the gain ratio, expresses the proportion of useful information generated by split, i.e., that appears helpful for classification. If the split is near-trivial, split information will be small and this ratio will be unstable. To avoid this, the gain ratio criterion selects a test to maximize the ratio above, subject to the constraint that the information gain must be large, at least as great as the average gain over all tests examined.

It has been observed that the class distribution in training data will affect classifier learning significantly since the decision trees are built with the statistical information of the samples, and naturally occurring class distribution may not be the best in terms of classification performance. Thus, not all the connection records from original training set were used. Some heuristics have been employed in selecting the training set. Basically, all the samples will be used if the number of samples is relatively small, and only a small subset will be randomly selected from the original training data if the size of the training data is huge.

6. Conclusions and Future Enhancement

In this paper, a novel architecture for NIDS has been proposed which uses an Enhanced C4.5 algorithm for intrusion detection in the system. The NIDS monitors network packets and connection status from the network layer to the application layer of the TCP/IP protocol stack. The NIDS detects anomaly behaviors through state transition and classification that has been carried out using Bayes-like decision. The major advantage of this architecture is that it creates profile using enhanced C4.5 algorithm and utilizes the original C4.5 algorithm to implement behavior classification. The system has been tested with a set of attacks using KDD-99 data set. The test show that this NIDS detects low-level network attacks effectively with low false positive rate, and it has a good performance in detection of unknown attacks, especially for PROBE, DOS and U2R attacks.

In the future work, number of inside attacks can be simulated and the attack can be detected. More attention shall be paid for U2R and R2L attacks. The detection of U2R and R2L attack is more difficult because of their close resemblance with the normal connections. More accurate detection methods can be used for U2R and R2L attacks. In the future work the anomaly is going to be detected by the clustering technique.

7. References

[1] S. Kumar and E. H. Spafford, 1995, “A software architecture to support misuse intrusion detection,” in Proceedings of the 18th National Conference on Information Security, pp.194–204, April 2001.

[2] K. Ilgun, R. A. Kemmerer, and P. A. Porras, 1995, “State transition analysis: A rule-based intrusion detection approach,” IEEE Transaction on Software Engineering Vol. 28, No. 2, pp. 181–199, March 2005.

[3] C. T. Lu, A. P. Boedihardjo, and P. Manalwar, “Exploiting efficient data mining techniques to enhance intrusion detection systems,” IEEE International Conference, Blacksburg, USA, pp. 512–517, April 2005.

[4] J. Zhang and M. Zulkernine, “A hybrid network intrusion detection technique using random forests,” Proceedings of the First International Conference on Availability, Reliability and Security, Vienna, University of Technology, Austria, pp. 121–132, April 2006.

[5] J. F. Tian, Y. Fu, Y. Xu, and J. L. Wang, “Intrusion detection combining multiple decision trees by fuzzy logic,” Proceedings of the Sixth International IEEE Conference on Parallel and Distributed Computing, Applications and Technologies, Dalian, China, pp. 256–258, May 2005.

[6] R. J. Henery, “Classification,” Machine Learning Neural and Statistical Classification, 1994.

[7] J. R. Quinlan, “C4.5: Programs for machine learning,” Morgan Kaufmann, San Mateo, California, 1993.

[8] C. Xiang and S. M. Lim, “Design of multiple-level hybrid classifier for intrusion detection system,” in IEEE Transaction on System, Man, Cybernetics, Part A, Cybernetics, Vol. 2, No. 28, Mystic, CT, pp. 117–122, May 2005.

[9] P. Kabiri and A. A. Ghorbani, “Research on intrusion detection and response: A survey,” International Journal of Network Security, Vol. 1, No. 2, pp. 84–102, February 2005.

[10] J. Shavlik and M. Shavlik, “Selection, combination, and evaluation of effective software sensors for detecting abnormal computer usage,” Proceedings of the First International Conference on Network security, Seattle, Washington, USA, pp. 56–67, May 2003.

[11] J. R. Quinlan, “Decision trees as probabilistic classifiers,” in Proceedings of the Fourth International Workshop on Machine Learning, Morgan Kaufmann, pp. 31–37, June 1987.

[12] W. Lee, S. Stolfo, and K. Mok, “A data mining framework for building intrusion detection models,” in Proceedings of IEEE Symposium on Security and Privacy, Oakland, CA, USA, pp. 120–132,1 June 1999.

[13] S. Sarasamma, Q. Zhu, and J. Huff, “Hierarchical Kohonen net for anomaly detection in network security,” IEEE Transactions on System, Man, Cybernetics, Part B, Cybernetics, Vol. 35, No. 2, pp. 302–312, April 2005.

[14] L. Breiman, “Pasting small votes for classification in large databases and on-line,” Machine Learning, Vol. 36, No. 1–2, pp. 85–103, 1999.

[15] J. W. Shavlik, S. Calcari, T. Eliassi-Rad, and J. Solock, “An instructable, adaptive interface for discovering and monitoring information on the World-Wide Web,” Intelligent User Interfaces, pp. 157–160, 1999.

[16] Z. X. Yu, J. R. Chen, and T. Q. Zhu, “A novel adaptive intrusion detection system based on data mining,” in Proceedings of the Fourth International Conference on Machine learning and Cybernetics, Guangzhou, pp. 2390–2395, June 2005.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

CA-AODV: Congestion Adaptive AODV Routing Protocol for Streaming Video in Mobile Ad Hoc Networks

Boraiah RAMESH[1], Dhanabalachandran MANJULA[2]
[1]*Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, India*
[2]*Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, India*
Email: sanchara@gmail.com[1], manju@annauniv.edu[2]
Received July 20, 2008; revised October 10, 2008; accepted October 14, 2008

Abstract

Whenever streaming of multimedia based data such as video, audio and text is performed traffic will be more and network becomes congested in mobile ad hoc networks. The present routing protocols are not able to cope up with this situation. It is observed that network congestion is the dominant reason for packet loss, longer delay and delay jitter in streaming video. Most of the present routing protocols are not designed to adapt to congestion control.

We propose a new routing protocol, Congestion Adaptive AODV Routing Protocol (CA-AODV), to address the congestion issues considering delay, packet loss and routing overhead. To evaluate their performance, we have considered mpeg4 for streaming video data using network simulator (NS2). CA-AODV outperforms present protocols in delivery ratio and delay, while introducing less routing protocol overhead. The result demonstrates that integrating congestion adaptive mechanisms with AODV is a promising way to improve performance for heavy traffic load in multimedia based mobile ad hoc networks.

Keywords: AODV, Streaming, Congestion, Multimedia, Video

1. Introduction

A mobile ad hoc network (MANET) is a collection of mobile nodes that form a wireless network without the use of a fixed infrastructure i.e., base stations or access points or any centralized administration. Ad hoc wireless networks are self-creating, self-organizing, and self-administering. They come into existence solely by interacting among their constituent mobile nodes and only such interactions are used to provide the necessary control and administration functions supporting such networks. The ad hoc wireless networks offer unique benefits and versatility for certain environments and certain applications. The preexisting fixed infrastructure and base stations are not being prerequisite to such networks. They can be created and used any time, anywhere. Such networks could be intrinsically fault-resilient, for they do not operate under the limitations of a fixed topology. Indeed, since all nodes are allowed to be mobile, the composition of such networks is necessarily time varying. Addition and deletion of nodes occur only by interactions with other nodes, no other agency is involved. Such perceived advantages elicited immediate interest in the early days among military, police, and rescue agencies in the use of such networks, especially under disorganized or hostile environments like isolated scenes of natural disaster and armed conflict. In recent days, home or small office networking and collaborative computing with laptop computers in a small area (e.g., a conference or classroom, single building, convention center) have emerged as other major areas of potential application. In addition, people also recognize that ad hoc networking has obvious potential application in all the traditional areas of interest for mobile computing.

Streaming multimedia type of data is very challenging issue in mobile ad hoc networks. Many researchers have considered these factors very seriously and are working in this direction [1,2]. Thus, our aim is to develop a routing protocol that provides alternate non congested path if node become congested. The congested node will immediately provide congestion status to concerned node in order to take necessary action. Providing congestion status is desirable for many applications, as this allows them to alter the data they transmit. For example, several visual compression techniques, such as MPEG-4 [3] and H. 263 [4], are designed to meet various channel conditions. Without this congestion status information, the node may not be alert to change its path, causing congestion in the network and a large number of dropped packets.

In mobile ad hoc networks, a message sent by a mobile node may be received simultaneously by all of its neighboring nodes. Messages directed to mobile nodes not within the sender's transmission range must be forwarded by neighbors, which thus act as routers. Due to mobility it is not possible to establish fixed paths for message delivery through the network. Mobile Ad hoc networks are composed of mobile stations communicating solely through wireless links [5]. Routing protocols are classified as proactive or reactive, depending on whether they keep routes continuously updated, or whether they react on demand.

The routing protocols [6] can also be categorized based on congestion-adaptive versus congestion-un adaptive routing. The congestion unawareness in routing in MANETs may lead to the following issues.

1) *Maximum delay to find a new route:* Traditional routing protocol takes maximum time for detecting congestion using a suitable control mechanism. In severe congestion situations, it may be better to use a new route. The problem with an on-demand routing protocol is the delay it takes to search for the new route.
2) *Huge routing overhead:* In case a new route is needed, it takes processing and communication effort to discover it. If multi-path routing is used, though an alternate route is readily found, it takes effort to maintain multiple paths.
3) *Heavy packet loss:* Many packets may have already been lost by the time congestion is detected. A typical congestion control solution will try to reduce the traffic load, either by decreasing the sending rate at the sender or dropping packets at the intermediate nodes or doing both. The consequence is a high packet loss rate or a small throughput at the receiver.

The above problems become more visible in large-scale transmission of traffic intensive data such as multimedia data. In such situation congestion is more probable and the negative impact of packet loss on the service quality is more of significance. We have proposed Congestion Adaptive AODV Protocol which tries to prevent congestion from occurring in the first place and be adaptive should a congestion occur. The ns-2 simulation results show that the proposed protocol significantly improves the packet loss rate and end-to-end delay while enjoying small protocol overhead and high-energy efficiency as compared to AODV [7], DSDV [8], DSR [9], and TORA [10]. Our proposed Adaptive Congestion AODV protocol tries to prevent congestion from occurring in the first place and it is adaptive to network congestion.

The remainder of the paper is organized as follows: Related Works is presented in Section 2. The proposed congestion adaptive routing protocol is presented in Section 3. In Section 4 we investigate simulation results and analysis of obtained results. Finally, Section 5 concludes the paper.

2. Related Works

Since Mobile ad network nodes are highly dynamic in nature, congestion is a main factor for more packet loss and longer delay. Traditional AODV is not an effective method under this situation in ad hoc networks [11]. It doesn't take any necessary precautions to handle the nodes which become congested under heavy network traffic. The modified version of AODV called CADV favors nodes with short queuing delays by adding it into the route to the destination. While this modification may improve the route quality, the issues of long delay and high overhead when a new route needs to be discovered remain unsolved. Furthermore, CADV is not congestion adaptive. It offers no remedy when an existing route becomes heavily congested. A similar routing protocol has been proposed in [12], Dynamic Load Aware Routing protocol, which favors low routing load in the routing path during route discovery. CADV and DLAR are both are single path on-demand routing protocols. Some of the multi path protocols were suggested in [13–15] which are extensions of AODV and DSR.

AODV (Ad hoc On-demand Distance Vector) [7] is a dynamic, self-starting, multi-hop on-demand routing protocol for mobile wireless ad hoc networks. AODV discovers paths without source routing and maintains table instance of route cache. This is loop free and uses destination sequence numbers. AODV also maintains active routes only while they are in use and delete the stale (unused) route. AODV performs Route Discovery using control messages Route Request (RREQ) and Route Reply (RREP) whenever a node wishes to send packet to destination. The source node in network broadcasts RREQs to neighbors and uses an expanding ring search technique. The forward path sets up in intermediate nodes in its routing table with a lifetime association using RREP. When route is broken, destination or intermediate node moves RERR to the source node. When RERR is received, source node reinitiate discovery is still needed.

DSR (Dynamic Source Routing) [9] is reactive, simple and efficient routing protocol for multi-hop wireless ad hoc networks of mobile nodes. DSR uses source routing and this protocol is composed of two main mechanisms: Route Discovery and Route Maintenance. Both the mechanisms work together entirely, on-loop-free routing. They can rapidly discover the changes in the network routes and are designed for mobile ad hoc networks of up to about two hundred nodes and they work well even with high rates of mobility. The source route is needed when some nodes originate a new packet destined for some node by searching its route cache or initiating route discovery using RREQ and RREP messages. On detecting the break, DSR sends RERR message to source for new route.

The Destination-Sequenced Distance-Vector (DSDV) [8] Routing Algorithm is based on the idea of the classical Bellman-Ford Routing Algorithm with certain improvements. Every mobile station maintains a routing table that lists all available destinations, the number of hops to reach the destination and the sequence number assigned by the destination node. The sequence number is used to distinguish stale routes from new ones and thus avoid the formation of loops. The stations periodically transmit their routing tables to their immediate neighbors. A station also transmits its routing table if a significant change has occurred in its table from the last update sent. So, the update is both time-driven and event-driven.

The routing table updates can be sent in two ways: – a “full dump” or an incremental update. A full dump sends the full routing table to the neighbors and could span many packets whereas in an incremental update only those entries from the routing table are sent that has a metric change since the last update and it must fit in a packet. If there is space in the incremental update packet then those entries whose sequence number has changed will be included into it. When the network is relatively stable, incremental updates are sent to avoid extra traffic and full dump are relatively infrequent. In a fast-changing network, incremental packets can grow big so full dumps will be more frequent.

The Temporally-Ordered Routing Algorithm (TORA) [10] is “an adaptive routing protocol for multi-hop networks”. TORA is a distributed algorithm so that routers only need to maintain knowledge about their neighbors. TORA also maintains states on a per destination basis like other distance-vector algorithms. It uses a mix of reactive and proactive routing. Sources initiate route requests in a reactive mode. At the same time, selected destinations may start proactive operations to build traditional routing tables. Usually, routes to these destinations may be consistently or frequently required, such as routes to gateways or servers. TORA supports multiple path routing. It is said that TORA minimizes the communication overhead associated with adapting to network topology changes. The reason is that TORA keeps multiple paths and it does not need to discover a new route when the network topology changes unless all routes in the local route cache fail. Hence, the trade off is that since multiple paths are used, routes may not always be the shortest ones.

TORA uses the concept of height associated with a certain destination to describe the routing metric used by routers. Like water flows in pipes, routers with higher heights may forward packet flows to neighbors with lower heights. Note that since heights for routers are associated with particular destinations, the paths to forward packets are also associated with the corresponding destinations. In networks using TORA, an independent copy of TORA runs for each possible destination. So for different destinations, routers may have different heights and links can have different directions.

3. CA-AODV: Congestion Adaptive Routing Protocol

The proposed routing protocol is designed to ensure the availability of primary route as well as alternative routes and reduce the route overhead. If congestion happens at any point of time between source and destination nodes on primary route, concerned node warns its previous node about congestion. The previous node uses a non congested alternate route to the destination node. Since video data is very sensitive in delay and packet loss, the measurement of congestion has been considered here depending on average packet delivery time and packet delivery ratio. The Congestion adaptive AODV is reactive routing protocol and has the following three divisions.

- Congestion status setup
- Route Discovery Process
- Route Maintenance Process

3.1. Congestion Status Setup

Calculate time taken at every intermediate node from source node periodically. The calculated time at intermediate node is called as calculated delay C_d. The average delivery time should be calculated by source node i.e. expected delay(E_d) to reach destination. Check C_d with expected delay E_d in the following manner and set up value of C_s congestion status.

The status of congestion can be indicated by three levels: Forward, Alert and Drop. The Forward level means that packet can be forwarded to the next node, Alert means continue with remaining packets, but not for a longer time, and Drop level means there is no alternate way to forward packet, just drop it.

- If calculated time $C_d < E_d$, the value of C_s will be in Forward level.
- If calculated time $C_d <= E_d$, the value of C_s will be in Alert level.
- If calculated time $C_d > E_d$, the value of C_s will be in Drop level.

3.2. Route Discovery Process

During the route-request phase, each node which receives a RREQ packet will determine level of congestion status. If Cs is in forward level, RREQ packet will be forwarded to next neighbor node. If C_s is in alert level, RREQ packet might be forwarded but this is not continued for a longer time. If Cs is in drop level, RREQ packet will be dropped.

We have modified conventional AODV as per our requirements. For all the nodes on the main route, RREQ packet will be forwarded to next node. The destination node will send the RREP packet back to the source as the conventional AODV does and then complete the route discovery process.

The data structures used are as follows:

The main routing table is represented by MRT[S_n,D_n]. It specifies the entry for destination D_n in the routing table of node S_n. The MRT[S_n,D_n].attr specify the value for the attribute attr. The traffic can be reduced by dropping RREQ packets when congestion status is "Drop" and also stop broadcasting RREQ packets.

1) The main routing table metric attribute is set to 1: MRT[S_n,D_n].nc_metric=1, for n nodes. Set destination node and its congestion status level: MRT[S_n, D_n].nc_hop = D and MRT[S_n,D_n].hop_status= "Forward". Set main table as not congested node: MRT[S_n,D_n].nc_hop=−1 for every other node.
2) Whenever node S_n receives an updated packet from its neighbor node S_{next}, it will check if MRT[S_{next}, D_n].nc_status="Drop" and MRT[S_{next},D_n].nc_status ="Alert" then node S_n will initiate non congested path discovery process towards the node S_{next} obtained from the update packet.
3) Searching for non congested optimal route: Set TTL to 2 x k in non congested request packets, where k is distance between node S_n and non-congested node S_i on the main route.
4) If non congested node is already present in main routing table, drop non congested request packet.
5) If timeout occurs after certain period, delete the entries in the non-congested alternate table.
6) The traffic splitting can be done effectively as follows: If next main node MRT [S_n, D_n].hop = "drop" the incoming packets will follow main Link.
   S_n → MRT[S_n, D_n].hop and with probability p = MRT [S_n, D_n].prob = 0.5.
   Non congested link S_n → MRT[S_n, D_n].nc_hop will have equal chance (1-p = 0.5).

3.3. Route Maintenance Process

The route maintenance of our proposed modified version of AODV algorithm will take necessary actions compared to traditional AODV. If there is a broken route detected by monitoring error message RERR and/or node does not receive any reply message from a specific neighbor within a predefined interval of time, it remarks that routes as invalid and sends an error message to the upstream nodes. Once the error message has been received by previous node, it would select best alternate route.

4. System Simulation Design

We have implemented proposed protocol using Network Simulator NS-2 [16] version 2.28. We have compared CA-AODV to DSR, AODV, DSDV and TORA, the most popular MANET routing protocols. In following sections observations are discussed.

4.1. System Simulation model

We built a simulation model to support video transmission which consists of 100 mobile nodes to form ad hoc network within the 1500m x 800m rectangular field. The nodes were equipped with omni-directional antennas. To test the performance of our proposed routing protocol with other protocols in MANET environment we have used MPEG4 video traffic generator [17]. The packet size used in our simulations is 512 bytes and the raw channel bandwidth is 2 Mbps. The ten pairs of source destination flows are randomly chosen to observe congestion. The routing buffer at the network layer could store up to 128 data packets. The MAC layer was based on IEEE 802.11 CSMA and interface queue at MAC layer could hold 50 packets. The random waypoint model [18] was used with maximum node speed of 4m/s as suggested in [19].

The simulations were run for 900 seconds for ten different simulations with different pause times, where a higher pause time reflects lower mobility, 0 indicates a high mobility scenario, while a pause time of 900 is considered a stable network. For each connection, the source generated 512-byte data packets at a constant bit rate (CBR) in the traffic model.

4.2. Parameters Monitored

We have evaluated the performance of CA-AODV by considering three important parameters: Data packet delivery ratio, Normalized routing overhead and End-to-end data packet delay [20].

Data packet delivery ratio: The ratio of the number of packets sent by the source nodes to the number of packets received by the destination nodes.

Normalized routing overhead ratio: The ratio of the total number of routing packets transmitted to the number of data packets delivered. For packets sent over multiple hops, each transmission of a packet over a hop counts as one transmission. Protocols that generate large amounts of routing overhead increase the probability of packet collision and data packet delays in network interface queues.

End-to-end data packet delay: The delay in transmitting data packets through wireless links plus the delay in the network interface queues due to network congestion. This metrics includes all the possible delays caused by buffering during the route discovery latency, queuing at the interface queue, retransmission delays at the MAC layer.

4.3. Simulation Result Analysis

The performance of proposed CA-AODV was evaluated by comparing it with AODV, DSR, DSDV and TORA. We considered various numbers of sessions with different packet rates and mobility models.

Figures 1 to 3 show the performance of CA-AODV and other routing protocols with different packet rates (20 packets/sec, 30 packets/sec and 40 packets/sec), and a pause time of 500s. The proposed routing protocol provide a high packet delivery ratio, small normalized routing overhead and low end-to-end delay than AODV and other routing protocols. As a consequence, AODV has slightly same performance compare to CA-AODV under light traffic - this is also observed at 10 packets/sec. This is because of less traffic load. When the packet rate increases to 20, CA-AODV performs better than AODV and other routing protocols, and this becomes more pronounced as the traffic load increases. At 20 packets/s, the packet delivery ratio of CA-AODV outperforms AODV by 12%, DSR by 17% and DSDV by 20% and TORA by 22%. Whenever the packet rate is increased, CA-AODV uses less routing overhead (up to 20% less) than AODV. The end-to-end delay is slightly more than AODV because of some alternate routes that are not the shortest.

Thus with CA-AODV the traffic load is more balanced, and the probability of packet loss is reduced. Furthermore, in congested nodes, alternate routes should be selected, resulting in a significant increase in packet delivery, decrease in routing overhead and increase in delays.

Different mobility models were simulated by using different pause times. It was observed that mobility has a great impact on the performance of CA-AODV and other routing protocols. Performance is always worse with high mobility, but mobility has a slightly greater impact on CA-AODV than AODV and other routing protocols. Figures 4 to 6 show the performance of CA-AODV and other protocols with 20 CBR sessions. The CA-AODV has a lower packet delivery ratio than AODV by 2% at 0s pause time (high mobility), and the routing overhead is 10% less than with AODV. At 900s pause time (low mobility), CA-AODV has a 5% to10% higher packet delivery ratio than AODV and other protocols, and requires 30% to 40% less routing overhead than AODV and other protocols.

The delay variation is less than that of AODV and DSR which makes our protocol more suitable for multimedia kind of applications as shown in Figures 2 and 5.

The routing overhead incurred by CA-AODV is very less when compared to other routing protocols. This is shown in Figure 3 and 6. When packet rate was 50 packets per second the proposed protocol incurred less routing overhead and delivered 21.34% more data than AODV. This is because, upon link breakage, AODV tried to establish a new route to the destination by broadcasting RREQ and RREP packets, CA-AODV protocol tried to make use of non congested available route and uses route request packets very often. The overhead to maintain non-congested paths in proposed algorithm is kept small by minimizing the use of multiple paths.

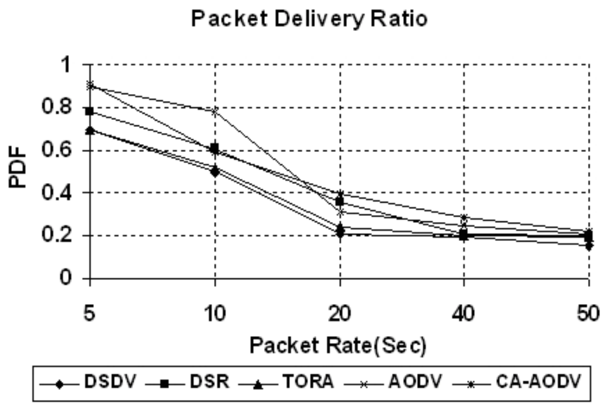


Figure 1. Comparison of packet delivery ratio vs. packet rate.

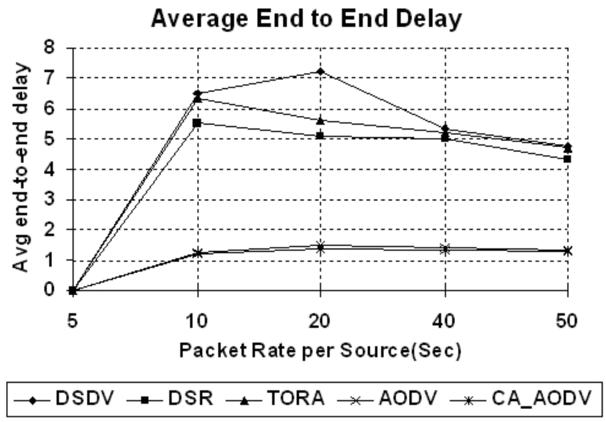


Figure 2. Comparison of average end to end delay vs. packet rate.

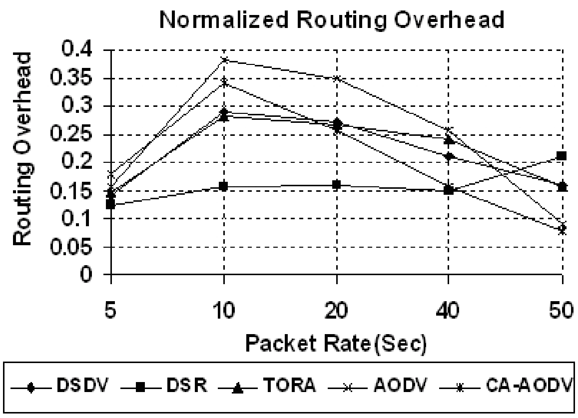


Figure 3. Comparison of normalized routing overhead vs. packet rate.

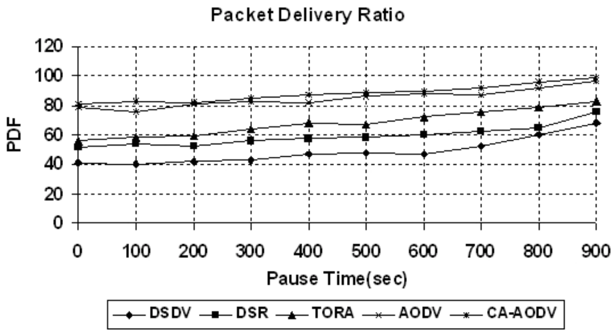


Figure 4. Comparison of packet delivery ratio vs. different pause time.

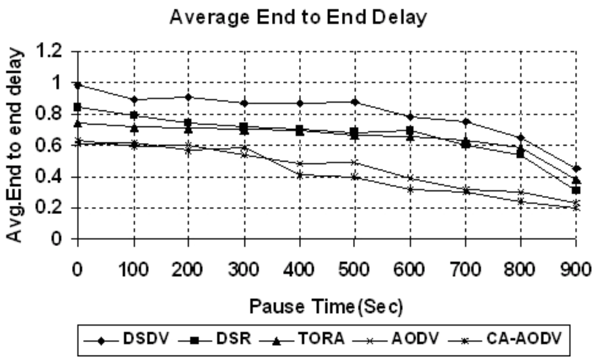


Figure 5. Comparison of average end to end delay vs. different pause time.

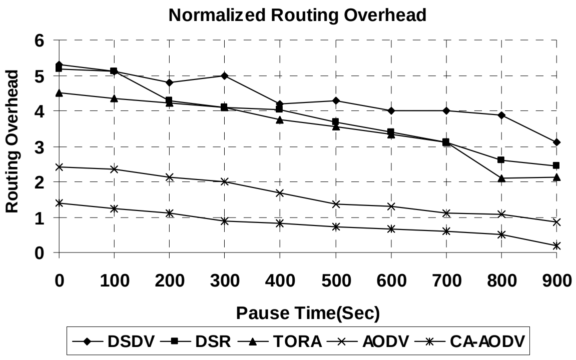


Figure 6. Comparison of Normalized Routing Overhead vs. different Pause time.

5. Conclusions

In this paper, we have proposed a new method to adapt to network congestion for video streaming in mobile ad hoc networks. The existing MANET protocols are not adaptive to network congestion and cannot handle the heavy traffic load while offering services to multimedia applications. The proposed Congestion Adaptive AODV Routing Protocol (CA-AODV) reduces packet losses than other routing protocols in real time transmission. The non-congested alternate route concept in the proposed method help next node that may go congested.

Whenever a node becomes aware of congestion ahead, it finds a non-congested alternate route that will be used to avoid congestion that is about to happen. The part of incoming traffic is split and then sent on the non-congested route, making the traffic coming to the congested node less. Thus congestion can be avoided. Proposed Algorithm does not incur heavy overhead due to maintenance of non-congested alternate paths. It also offers high packet delivery when the traffic in heavy. The delay incurred while establishing new connection is low because of using existing non-congested paths. Thus the proposed algorithm in mobile ad hoc networks is especially designed for multimedia applications. The difference between CA-AODV and other protocols were examined. Results were presented considering various situations to show the effectiveness of our proposed method. It is clear that CA-AODV can provide good performance comparable to AODV and other protocols in light traffic, and better performance in heavy traffic at a cost of slight longer delays.

6. References

[1] C. H. Ke, C. H. Lin, C. K. Shieh, and W. S. Hwang. "A novel realistic simulation tool for video transmission over wireless network," IEEE (SUTC2006), June 5–7, 2006.

[2] A. Lo, G. Heijenk, and I. Niemegeers, "Performance evaluation of MPEG-4 video streaming over UMTS networks using an integrated tool environment," International Conference, Netherlands, 2005.

[3] Y. Zhang, "Very low bit rate video coding standards," in Proceedings of Visual Communication and Image Processing, Vol. 2501, May 1995, pp. 1016–1023. 572 IEEE Journal on Selected Areas in Communications, Vol. 23, No. 3, March 2005.

[4] G. Coete, B. Erol, M. Gallant, and F. Kossentini, "H.263+: Video coding at low bit rates," IEEE Transactions on Circuits Systems for Video Technology, Vol. 8, No. 7, pp. 849–866, November 1998.

[5] G. S. Lauer, "Packet-radio routing, in: Routing in communications networks", Editor: M. E. Steenstrup (Rentice-Hall, 1995), Chapter 1, 35, pp. 1–396.

[6] S. Ramanathan, M. E. Steenstrup, "A survey of routing techniques for mobile communications networks, mobile networks and applications," Vol. 1, pp. 98–104, 1996.

[7] C. E. Perkins, "Ad hoc on demand distance vector (AODV) routing," IEFT Internet Draft, November 1998, http://www.ieft.org/internetdrafts/draft-ietf-manet-aodv-02.txt.

[8] C. E. Perkins and P. Bhagwat, "Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers," Proceedings of ACM SIGCOMM' 94, London, UK, pp. 234–244, August–September 1994.

[9] D. B. Johnson, D. A. Maltz, and J. Broch, "Dynamic source routing for multihop wireless ad hoc networks," in Ad Hoc Networking, edited by Charles E. Perkins, Addison-Wesley, Chapter 5, pp 139–172, 2001.

[10] V. Park and S. Corson, "Temporally-ordered routing algorithm (TORA) Version 1 functional specification," IETF Internet Draft, work in progress, draft-ietf-manet-tora-spec-04.txt, 2001.

[11] Y. Lu, W. Wang, Y. Zhong, and B. Bhargava, "Study of distance vector routing protocols for mobile ad hoc networks," Proceedings IEEE International Conference Pervasive Computing and Comm. (PerCom), pp. 187–194, March 2003.

[12] S. J. Lee and M. Gerla, "Dynamic load-aware routing in ad hoc networks," Proceedings IEEE International Conference on Communications, pp. 3206–3210, June 2001.

[13] S. J. Lee and M. Gerla, "Split multipath routing with maximally disjoint paths in ad hoc networks," Proceedings IEEE International Conference on Communications, pp. 3201–3205, June 2001.

[14] M. Marina and S. Das, "On-demand multipath distance vector routing in ad hoc networks," Proceedings IEEE International Conference Network Protocols (ICNP), pp. 14–23, 2001.

[15] Nasipuri and S. R. Das, "On-demand multipath routing for mobile ad hoc networks," Proceedings IEEE International Conference on Computer Communications and Networks (IC3N), October 1999.

[16] NS-2. Network Simulator. http://www.isi.edu/nsnam/ns/.

[17] M. Chowdhury, Y. Tu, and T. Pham, "MPEG traffic picture matrices via TCP over ATM," Proceedings of the International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CSITeA'03), Rio de Janeiro, Brazil, pp. 128–132, June 5–7, 2003.

[18] A. Boukerche, "Performance evaluation of routing protocols for ad hoc wireless networks. Mobile networks and applications," Vol. 9, pp. 333–342, 2004.

[19] Y. Lu, W. Wang,Y. Zhong, and B. Bhargava. "Study of distance vector routing protocols for mobile ad hoc networks," in IEEE International Conference on Communication, Helsinki, Finland, pp. 3201–3205, June 2001.

[20] C. Perkins, E. M. Royer, S. R. Das, and M. Marina, "Performance comparison of two on-demand routing protocols for ad hoc networks," IEEE Personal Communications Magazine, Vol. 8, No. 1, pp. 16–28, February 2001.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Performance Study of a Cross-Layer Based Multipath Routing Protocol for IEEE 802.11e Mobile Ad Hoc Networks

Hang SHEN[1], Guangwei BAI[2], Junyuan WANG[3], Yong JIN[4], Jinjing TAO[5]
Department of Computer Science and Technology, Nanjing University of Technology, Nanjing, China
E-mail:{[1]shen_hang, [4]njutjinyong}@yahoo.com.cn, [2]bai@njut.edu.cn, {[3]yuanwa001, [5]taojinjing_1985}@163.com
Received November 26, 2007; revised April 30, 2008; accepted May 21, 2008

Abstract

Communication over wireless links identifies significant challenges for routing protocols operating. This paper proposes a Cross-layer design based Multipath Routing Protocol (CMRP) for mobile ad hoc networks, by means of the node energy signal from the physical layer. The purpose is to optimize routing decision and path quality. The nodes' mobility behavior is predicted using a notion of "Signal Fading Degree, SFD". Especially, in combination of the IEEE 802.11e standard at the MAC layer, we determine that the IEEE 802.11e makes a significant contribution to performance improvement of CMRP. Performance evaluation of AODV in legacy 802.11 and CMRP in IEEE 802.11e shows that, as a function of speed of node mobility, a tremendous reduction achieved, in metrics such as the average end-to-end delay, route overhead, route discovery frequency, normalized routing load – almost more than 80%, 40%, 40%, and 40%. In the case of varying number of sessions, the reduction for route discovery frequency and normalized routing load are up to 70% and 80%.

Keywords: Wireless Mobile Ad Hoc Networks, Multipath Routing, Cross-Layer Design, IEEE 802.11e

1. Introduction

Technologies such as IEEE 802.11 wireless LANs (WLANs) have revolutionalized the way people think about networks, by offering users freedom from the constraints of physical wires. Mobile users are interested in exploiting the full functionality of the technology at their fingertips, as wireless networks bring closer the "anything, anytime, anywhere" promise of mobile networking [1,2].

Routing in wireless mobile *ad hoc* networks (MANETs) has been an active area of research for many years [3, 4]. A MANET is an autonomous network that can be formed without (necessarily) using a pre-existing infrastructure. The characteristics such as self-organizing make MANETs be prevalent today and be continued growth in popularity. Without centralized administration, individual nodes in MANETs are responsible for dynamically discovering which other nodes they can directly communicate with. A key assumption is that not all nodes can directly communicate with each other, so mobile nodes forward packets for each other, that is, multi-hop, allowing communication among nodes outside wireless transmission range. The node mobility, dynamic topology and the fundamentally limited capacity of the wireless medium, together with wireless transmission effects such as attenuation, multipath propagation and interference, combine to create significant challenges for routing protocols operating.

Firstly, recent research shows, that the single routing protocol reflects some limitations in case of highly dynamic network topology and strictly limited resources. One observation of single routing AODV [5] is that, though the source actually discovers multiple paths during the route discovery process, it chooses only the shortest delay route and discards the rest. Also, frequent route breaks cause the intermediate nodes to drop packets because no alternate path to the destination is available. Therefore, multipath routing algorithms have drawn researchers' attention. The multipath routing allows building multiple paths between a source-destination pair. It can provide benefits such as fault tolerance, load balancing, bandwidth aggregation, and improvement in QoS metrics such as delay [7–25].

Another key issue is cross-layer optimization. For MANETs protocol design, the physical layer must adapt to rapid changes in link characteristics, the MAC layer needs to minimize collisions and allow fair access, the network layer needs to make a routing decision for effective data delivery to the destination, and so on. The cross-layer design is desirable for improving performance in MANETs, since the methodology of layered protocol design does not necessarily lead to an optimum solution for dynamic environment. Under the layered protocol design, MANET routing protocols are unable to retrieve energy and location information from the underlying data link layer and physical layer and, thus, unable to calculate routes based on such information. In this work, we use cross-layer design to refer to protocol design and optimization, that is, make use of the node energy signal from the physical layer to optimize routing decision.

Finally, the IEEE 802.11e standard was developed to offer QoS capabilities to WLANs (e.g. MANETs), offering significant improvements to multimedia traffic [26]. MANETs will also benefit from this new technology since the most widely deployed and used wireless interfaces are IEEE 802.11 based. Currently, relatively little research work has focused on interaction between IEEE 802.11e and multipath routing protocols. In this work, the performance of CMRP gain obtained from IEEE 802.11e is demonstrated, by means of a series of simulation experiments.

Based on cross-layer design, we propose a multipath routing protocol (CMRP), in consideration of IEEE 802.11e technology, to improve dynamic multi-hop routing performance for MANETs. CMRP uses signal strength information to optimize routing decision and path quality. The purpose of this work is to ensure wireless multi-hop network performance improvement. Our simulation results demonstrate that, in combination of the IEEE 802.11e standard at the MAC layer, CMRP provides significant performance improvement in terms of average end-to-end delay, route overhead, route discovery frequency and packet loss as well.

The remainder of this paper is organized as follows. Section 2 discusses related work on current MANETs routing protocols. Section 3 proposes the cross-layer optimized multipath routing protocol and presents the details of its implementation. Section 4 discusses that the performance improvement of CMRP using IEEE 802.11e standard. Section 5 involves thorough analyses and evaluation of the CMRP performance in simulation methodology. Finally, Section 6 concludes the paper.

2. Related Work

Most proposed wireless mobile *ad hoc* routing protocols are unipath protocols, which only use a single path to send packets to the destination. The main idea with multipath routing, which has been originally studied in wired networks, has existed for some time. Recently, many different multipath routing protocols based on AODV or DSR [6] for wireless multi-hop network have been proposed in literature.

As an extension to AODV, M. K. Marina *et al.* proposed a multipath routing algorithm, i.e. AOMDV [7]. The protocol computes multiple loop-free and link-disjoint paths. Loop freedom is guaranteed by using a notion of “advertised hopcount”. Link-disjointness of multiple paths is achieved by using a particular property of flooding. In details of CMRP, we modify both routing selection and routing maintenance based on AODV, in a manner similar to AOMDV. Z. Ye *et al.* proposed AODVM [8], which achieves a framework for reliably routing information. Duplicate RREQ messages are not discarded by intermediate nodes. Instead, all received RREQ packets are recorded in an RREQ table at the intermediate nodes. Caching and Multipath (CHAMP) Routing Protocol reported in [9] uses cooperative packet caching and shortest multipath routing to reduce packet loss due to frequent route breakdowns. X. Li *et al.* propose NDMR [10], which modify and extend AODV to include the path accumulation feature of DSR in route control packets, so that much lower overhead is employed to discover multiple node-disjoint paths.

Derived from DSR, SMR [11] focuses on building and maintaining maximally disjoint paths in order to prevent certain links from becoming congested and to efficiently utilize the available network resources. W. Wei *et al*. propose RMPSR [12], which distributes video packets over two primary routes of two route sets, to support Multiple Description Coding (MDC) application over MANETs. A. Nasipuri *et al.* developed a multipath protocol [13], in consideration of the situation where the destination replies to a selected set of RREQs. Recently there has been increased interest in protocols for wireless networks that rely on cross-layer [14–16]. M. Li *et al.* present a cross-layer multipath routing protocol (EMRP) [17]. By sharing the information among the physical layer, the MAC sublayer and the network layer, EMRP is able to utilize the network resources efficiently. H. Sun *et al*. propose an adaptive QoS routing scheme supported by cross-layer cooperation [18], considering the impacts of node mobility and lower-layer link performance. The multiple QoS requirements are satisfied by adaptively using forward error correction and multipath routing mechanisms, based on the current network status.

Routing protocols for NANETs have traditionally focused on finding paths with minimum hopcount in the last few years. In [3], R. Draves *et al.* find that minimal hopcount paths may provide poor performance because they tend to include wireless links between distant nodes and these long wireless links can be slow or loss, leading to poor throughput. Therefore, the routing algorithm can select better paths by explicitly taking into account the quality of wireless links. We design the schemes for path storage and selection in consideration of this idea.

For interaction between IEEE 802.11e and routing protocols, in [26] Carlos T. Calafate *et al.* exposed results related to the interaction of AODV/DSR and the IEEE 802.11e technology in terms of throughput and normalized routing load in order to assess the improvements resulting from the IEEE 802.11e. In this work, we use IEEE 802.11e as the protocol of MAC sublayer to improve the performance of CMRP.

To the best of our knowledge, currently, relatively little research work has focused on interaction between IEEE 802.11e and multipath routing protocols. This paper is, based on cross-layer design, aimed at a multipath routing protocol for IEEE 802.11e MANETs. We refer to our work as an enhanced version of AODV, focusing on the performance improvement of overall network.

3. A Cross-Layer based Multipath Routing

The main idea in CMRP is to compute multiple power-aware paths during route discovery. It is designed primarily for highly dynamic ad hoc networks where link failures occur frequently. CMRP stores and selects the paths according to signal strength. That is, it stores multiple *SDFs* of path on receiving routing message from the same source node, and selects a path also with largest *SDF* when transmitting data (see Subsection 3.1 for details). As a result, data packets are able to travel along the stable path. Especially, we deduce that the path with the largest *SDF* is also an energy-efficient path, since it can reduce signal attenuation for packets sending.

We describe the CMRP from the following two aspects. At first, a policy to predict the node mobility behavior is suggested. Then, the process of CMRP routing establishment and maintenance are presented on the basis of mobility prediction. At the same time, we discuss the power-saving characteristic based on routing mechanism as mentioned above.

3.1. Node Mobility Prediction

It is well known that the radio signal gets weaker as it propagates. In a simulation environment, the node energy signal strength is able to indicate the distance between the sending node and the receiving node, as well as the quality and stability of the link to certain extent. In a realistic environment, however, an estimation of distance using signal strength may introduce errors, but we still can deduce node mobility behavior and the relative distance between a node and its neighbor via measuring signal fading. For example, we can deduce if the moving nodes lead to the link interruption in a short time. The hopcount in traditional routing protocols does not reflect the nodes' relative location exactly. In a link with weak signal strength, a few hopcounts may lead to numerous packets loss. Therefore, using multipath routing will be not worth the candle if the paths are not chosen appropriately. We store and choose the paths according to the signal strength from the physical layer.

The severe signal fading is one of the characteristics in wireless communication. We use "Signal Fading Degree, *SFD*" to predict the node mobility, i.e., the distance between the sending node and the receiving node. The smaller the *SDF* is (the weaker the signal is) and the further the distance is, and the higher the probability of link interruption is. Formula (1) defines *SDF* of the node, which is transmitted by routing message (see Section 3.2 for details). This Formula gives a measure of the relative stability between two serial nodes in the entire path.

$$SDF_{node} = \frac{RP_{node} - TP_{node}}{TP_{node}} \quad \mathbf{(1)}$$

where *RP* denotes the remaining node energy. *TP* represents a fixed energy consumption of every efficient data packet sending. The accumulation of each node *SDF* hop by hop is the *SDF* of the whole path. *SDF* is used for measuring path reliability, i.e. a larger *SDF* indicates a more reliable link, whereas, a smaller *SDF* indicates a less reliable link. There are two ways to represent the *SDF* of the whole path, i.e. This Formula gives a measure of the relative stability of the path.

$$SDF_{path} = \sum_{i \in path} SDF_i \quad \mathbf{(2)}$$

$$SDF_{path} = \prod_{i \in path} SDF_i \quad \mathbf{(3)}$$

Table 1 shows the structure of the route table entries for CMRP. We add the *SDF* in path list.

Formula (2, 3) gives a measure of the relative stability of the path. We indicate the node on the path, and Formula (3) is a measuring standard for the signal fading degree of the whole path. As shown in Figure 1, there are two paths from the source node *S* to the destination node *D*, with the value on the arrow denoting *SDF*. *SDF* results of the two paths (path_1 and path_2) calculated by Formula (2) are 0.6 and 0.65 respectively. However, the distance between node *B* and node *E* is longer and the interruption probability of path_2 is higher than that of path_1, so the result calculated by Formula (2) is wrong. While the *SDF* result of the two paths calculated by Formula (3) are 0.008 and 0.004 respectively, which can reflect the actual condition more exactly. Therefore, Formula (3) is chosen to calculate.

3.2. Routing Establishment and Maintenance

The multipath routing protocol seeks multiple disjoint

Table 1. Structure of the route table for CMRP.

Main Information Field	Contents
destination	An IP address to which data packets are to be transmitted.
sequence number	A monotonically increasing number maintained by each originating node
advertised hopcount [7]	It is used to maintain multiple loop-free paths.
path list (not more than 3): {(*SDF*-1, Hopcount-1, Nexthop-1, Lasthop-1) (*SDF*-2, Hopcount-2, Nexthop-2, Lasthop-2) (*SDF*-3, Hopcount-3, Nexthop-3, Lasthop-3)…}	Pr means current power signal strength of this node, provided by the physical layer.
lifetime	Expiration time of the route entry

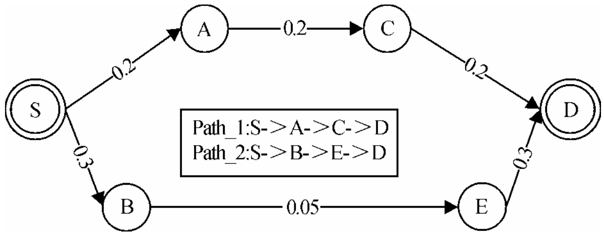


Figure 1. SDF comparison of two paths.

routes between source and destination nodes, which refer to node-disjoint or link-disjoint. We use link-disjoint in the design of CMRP since it can establish more paths than node disjoint, with higher stability. Figure 2, 3 and 4 describe the CMRP routing establishment and maintenance procedure.

As shown in Figure 2, nodes *A*, *B* and *C* receive RREQ from *S*, and then Node *E* receives RREQs from *A* and *B*, one after another. It forwards only the RREQ from *B* and discards the one from *A*. Since the *SDF* of path *E-B-S* is higher than *SDF* of path *E-A-S*, it is selected as the primary path. The other one becomes the alternate path. Destination *D* receives RREQs from nodes *F*, *G* and *H*. Note the route table of node *E* and node *D*.

The numbers in parenthesis indicates the SDF carried by the sending node, the numbers on the right indicates the *SDF* of the link. The updating process is described by formula (1, 2): for example, *SDF* of *S-B* is 0.3, which multiplied by *SDF* initial value 1 is 0.3. The result 0.3 is sent out as *SDF* of node *B* by RREQ. 0.3 multiplied by the *SDF* of path *B-E* is 0.09 as *SDF* of node *E*, which is sent out by RREQ. In a similar calculation way, the *SDF* of the entire path is calculated finally. The path with the largest *SDF* is the primary path.

Figure 3 shows the process of route reply and path store. Destination *D* replies to nodes *F*, *G* and *H*. Node *E* forwards the RREP from node *F* to node *B* and from node *G* to node *B*. Source *S* thus obtains three routes to *D*. *S* selects the path *S-B-E-F-D* as its primary route since it has higher *SDF* than the other paths (see route table).

Source node *S* receives several RREP one after the other (see Figure 4). According to the *SDF* in route table,

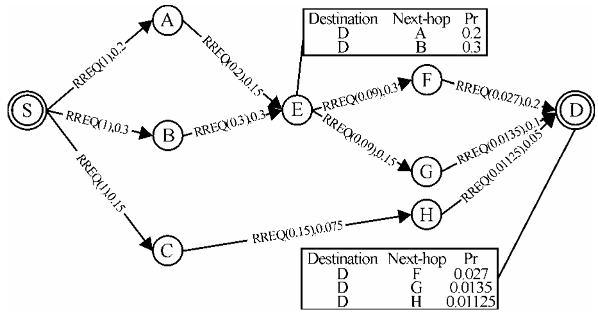


Figure 2. Process of route discovery and path store.

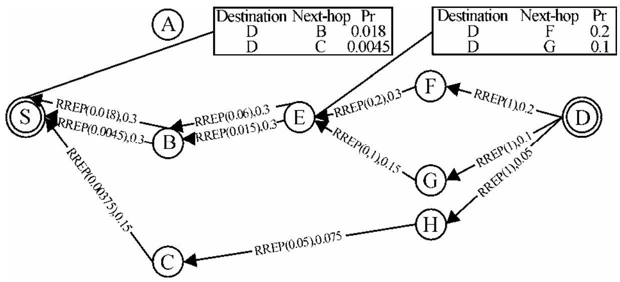


Figure 3. Process of route reply and path store.

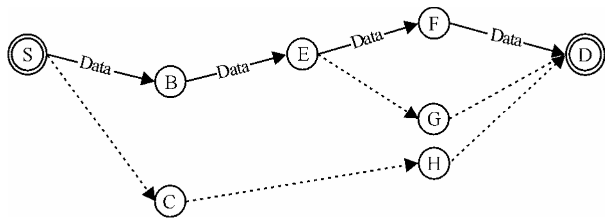


Figure 4. Establishment of multiple routes.

S selects the path_1 as the primary route for sending data packet. Path_2 and path_3 shown are used as alternative transmission paths.

We present the main process of routing computer as Formula (5)～Formula (9):

$$f(no\_route) then \quad sendrequest(SDF_i^d := 1) \tag{4}$$

$$if(rq\_dst_j^d \neq index) then \quad set \ \ rq\_pr_i^d := rq\_SDF_j^d * (SDF_i^d - tpr)/tpr; \tag{5}$$

$$forward(rq_i^d); \tag{6}$$

$$if(rq\_dst_j^d = index) then \quad sendreply(SDF_i^d := rq\_pr_j^d) \tag{7}$$

$$if(rp\_dst_j^d \neq index) then \quad set \ \ rp\_pr_i^d := rp\_SDF_j^d * (SDF_i^d - tpr)/tpr; \tag{8}$$

$$forward(rp_i^d); \tag{9}$$

The main process of loop-freedom is illustrated by Formula (10)～Formula (15):

$$if\,(seq\_num_i^d < seq\_num_j^d)\,then \quad seq\_num_i^d := seq\_num_j^d; \tag{10}$$

$$f\,(i \neq d)\,then \quad iad\_hopc_i^d := \infty;\ path\_list_i^d := NULL; \tag{11}$$

$$insert(pr_j^d, j, ad\_hopc_j^d{+}{+}, last\_hop)\ in\ path\_list_i^d; \tag{12}$$

$$else: ad\_hopc_i^d := 0; \tag{13}$$

$$elseif\,(seq\_num_i^d = seq\_num_j^d)\,\&\&(ad\_hopc_i^d > ad\_hopc_j^d) \tag{14}$$

$$insert(pr_j^d, j, ad\_hopc_j^d{+}{+}, last\_hop) in\, path\_list_i^d; \tag{15}$$

This is used whenever a node *i* receives a route message to a destination d from a neighbor *j*. *rq*, and *rp* stand for route requests and route replies. *SDF* and *tpr* respectively denote remaining node energy when packet is received and fixed transmission power for two-ray ground. The variables *seq_num* and *ad_hopc* represent the sequence number and advertised hopcount.

Thus, CMRP stores and selects the paths according to signal strength, stores multiple *SDFs* of path when receiving routing message from the same source node, and selects paths with strongest signal strength when transmitting data to enhance transmission reliability.

The transmission power is peculiar to wireless ad hoc networks, and is important because typically the nodes involved have a limited power supply, and radio communication consumes a large fraction of this supply [4]. Based on power-aware routing mechanism of CMRP, we deduce that the path with the largest *SDF* is not only a reliable path but also an energy-efficient path, since it can reduce signal attenuation for packets sending.

4. IEEE 802.11e QoS Enhanced WLAN

In this section, we briefly explain the IEEE 802.11e Enhanced Distributed Channel Access (EDCA) since we focus on *ad hoc* mode, and then discuss the significant contribution of IEEE 802.11e provided for enhanced performance of CMRP in MANETs.

4.1. Enhanced Distributed Channel Access

The most widely deployed and used wireless interfaces for IEEE 802.11e are IEEE 802.11 based. As a matter of fact, the IEEE 802.11e standard was developed to offer QoS capabilities to WLAN, offering significative improvements to multimedia traffic. In this work, we determine that the IEEE 802.11e makes a significant contribution to performance improvement of CMRP by means of a series of simulation experiments.

The IEEE 802.11e standard introduces the hybrid coordination function (HCF) which defines two new medium access mechanisms to replace legacy PCF and DCF. These are the HCF controlled channel access (HCCA) and the enhanced distributed channel access (EDCA). The HCCA is used in both periods, while the EDCA is used only during the CP. This new characteristic of HCF obviates the need for a contention-free period (CFP) since it no longer depends on it to provide QoS guarantees. With IEEE 802.11e, the point coordinator is replaced by a hybrid coordinator (HC) which also resides in an AP. A Basic Service Set (BSS) including a HC is referred to as a QBSS. In this paper we focus on *ad hoc* networks and, therefore, we are only interested in 802.11e stations implementing EDCA.

Table 2. User priority to IEEE 802.11e access category mapping.

User Priority	Designation	Access Category
1	BK (Background)	AC_BK
2	BK (Background)	AC_BK
0	BE (Best-effort)	AC_BE
3	EE (Video/Excellent-effort)	AC_BE
4	CL (Video/Controlled Load)	AC_VI
5	VI (Video)	AC_VI
6	VO (Voice)	AC_VO
7	NC (Network Control)	AC_VO

Table 3. IEEE 802.11e MAC parameter values.

Access category	AIFSN	CWmin	CWmax	TXOPLimit(ms)
AC_BK	7	15	1023	0
AC_BE	3	15	1023	0
AC_VI	2	7	15	3.008
AC_VO	2	3	7	1.504

EDCA is designed to provide prioritized QoS by enhancing the contention-based DCF. Before entering the MAC layer, each data packet received from the higher layer is assigned a specific user priority value. How to tag a priority value for each packet is an implementation issue. At the MAC layer, EDCA introduces four different first-in first-out (FIFO) queues, called access categories (ACs). Each data packet from the higher layer along with a specific user priority value should be mapped into a corresponding AC according to the Table 2. Different kinds of applications (e.g., background traffic, best effort traffic, video traffic, and voice traffic) can be directed into different ACs. In Table 3 we can see each AC behaves as a single DCF contending entity with its own contention parameters (CW_{min}, CW_{max}, AIFSN and $TXOP_{Limit}$), which are announced by the QAP periodically in beacon frames. Basically, the smaller the values of CW_{min}[AC], CW_{max}[AC], and AIFS[AC], the shorter the channel access delay for the corresponding AC and the higher the priority for access to the medium.

A new type of IFS is introduced In EDCA, the arbitrary IFS (AIFS), in place of DIFS in DCF. Each AIFS is an IFS interval with arbitrary length as follows:

AIFS [AC] = SIFS + AIFSN [AC] × slot time, where AIFSN [AC] is called the arbitration IFS number. After sensing the medium idle for a time interval of AIFS [AC], each AC calculates its own random backoff time (CW_{min}[AC] ≤ backoff time ≤ CW_{max}[AC]). The purpose of using different contention parameters for different queues is to give a low priority class a longer waiting time than a high-priority class, so the high-priority class is likely to access the medium earlier than the low-priority class. Note that the backoff times of different ACs in one QSTA are randomly generated and may reach zero simultaneously. This can cause an internal collision. In such a case, a virtual scheduler inside every QSTA allows only the highest-priority AC to transmit frames.

4.2. Contribution of IEEE 802.11e

IEEE 802.11e has a great potential to improve CMRP performance in wireless networks. Firstly, IEEE 802.11e allows wireless nodes to occupy channel for a long period of time during Transmission Opportunity (TXOP). This characteristic is able to dramatically decrease channel overheads caused by interception, Inter-frame, backoff and competition; from our perspective, CMRP performance will benefit from the improved path quality together with the extended occupation period of the channel. Secondly, Block ACK mechanism, i.e. it only replies one ack_frame to multiple data packet to decrease the overheads; finally, CFB enables an EDCA to transmit multiple frames once the medium or TXOP is acquired, without contending for the medium for every frame.

We consider that these characteristic mentioned above of MANETs stations to IEEE 802.11e are very important not only for multimedia traffic support, but also to improve the efficiency of the routing mechanism. This deduction will be verified by means of a series simulation experiments later.

5. Performance Evaluation

In this work, we use NS-2 [27] with TKN 802.11e module [28] to evaluate the performance of CMRP, comparing it with AODV in the same MAC sublayer protocol and conventional layered protocol stack, which use Legacy 802.11 in the MAC sublayer. Two simulation experiments are conducted, where the rate of the node motility and the number of sessions are varied in order to analyze and compare the performances of CMRP and AODV in Legacy 802.11 and IEEE 802.11e. The detailed simulation parameter setting is illustrated in Table 4.

The following key metrics are used in different scenarios to evaluate CMRP performance.

- Average End-to-End Delay: It includes all delays caused by buffering during route discovery, queuing at the interface, retransmission at the MAC, propagation and transfer times.
- Total Packets Loss: This includes all possible packets loss such as data packet loss and control packet loss.
- Route Overhead: The total number of control packets transmitted by any node.
- Normalized Routing Load: The total number of control packets divided by the total number of CBR packets received by destination node.
- Route Discovery Frequency: The total number of route discoveries initiated per second.
- Average Hopcounts: Average hopcounts of routes for data sending.

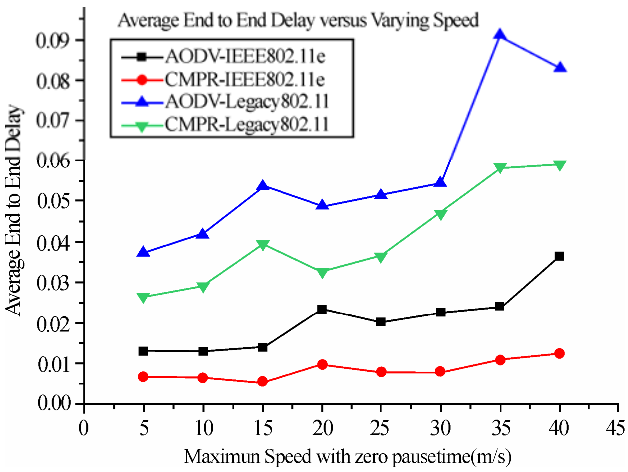
5.1. Performance with Varying Mobility

Figure 5 shows the six performance metrics as a function of mobility in experiment I. The max speed of node mobility is varied from 5m/s to 40 m/s. The number of CBR sessions is 10.

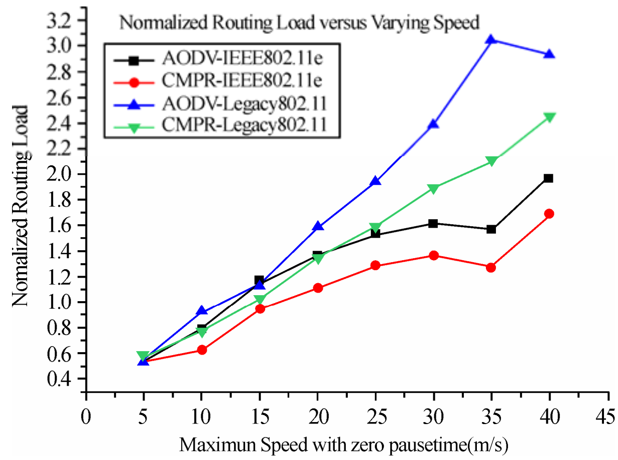
Figure 5(a) shows comparison of average end-to-end delay between the two routing protocols. CMRP with IEEE 802.11e has the shortest delay. Next is AODV with IEEE 802.11e, followed by CMRP with Legacy 802.11 and AODV with Legacy 802.11 respectively. The simulation results demonstrate that a tremendous reduction is achieved, in the average end-to-end delay with both CMPR and AODV in IEEE 802.11e, but that of CMPR decreases much more pronounced, as shown: 80% decreases against AODV in Legacy 802.11, 60% against AODV in IEEE 802.11e. On the other hand, the delay variation of CMRP tends to be much smoother comparing to AODV. CMRP builds multiple link-disjoint routes in the route request process and triggers a new route request process when all the routes are broken. These steps help CMRP maintain multiple routes longer than that of AODV. In traditional multipath routing, the primary path selected may not be always optimal in some cases. Moreover, the reliability of alternative paths often become poor, even broken when needed. By improving primary and alternative path qualities, CMRP is able to suspend link failures. Another important aspect is that IEEE 802.11e provides significant contribution for CMRP performance

Table 4. Simulation environment.

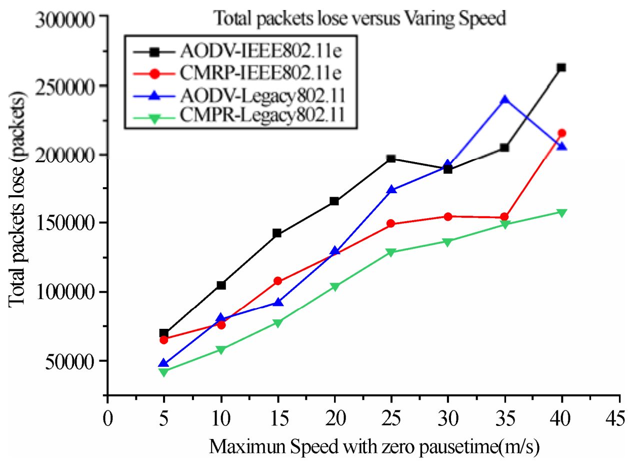
Parameter	Value
Transmission Range	250m
Simulation Time	800s
Topology Size	750m*750m
Number of Mobile Nodes	50
Interface Queue Type	PriQueue
Interface Queue Length	50
Traffic Type	CBR(constant bit rate)
Packet Rate	5 packets/s
Packet Size	512 bytes
Pause Time	0s
Model Mobility	Random Waypoint
Traffic Model	Spread Randomly
Maximum Speed (experiment I)	5m/s - 40m/s
Maximum Speed (experiment II)	10m/s
Number of Sessions (experiment I)	10
Number of Sessions (experiment II)	5 - 25



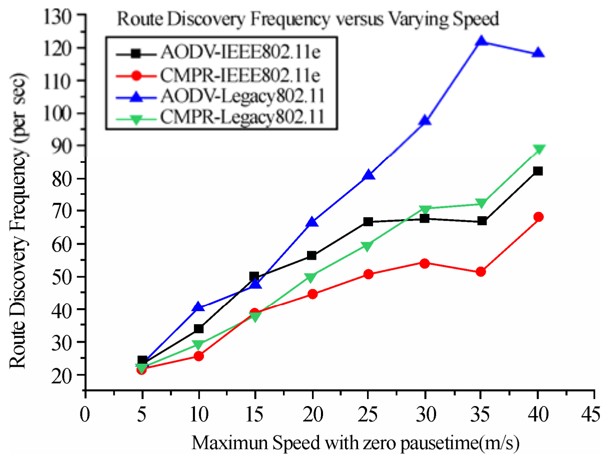
(a) Average end to end delay



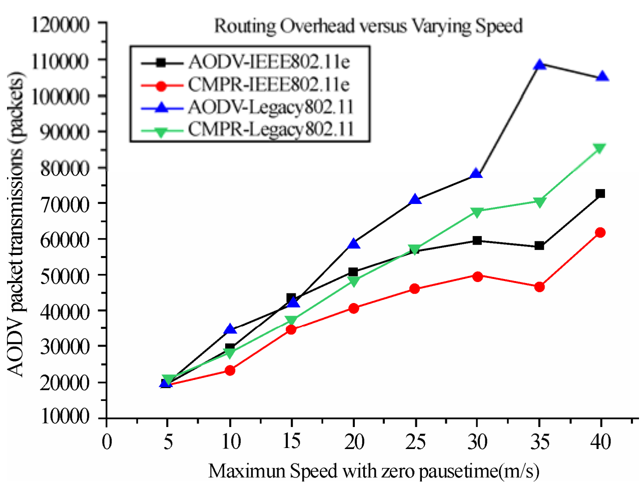
(d) Normalized routing load



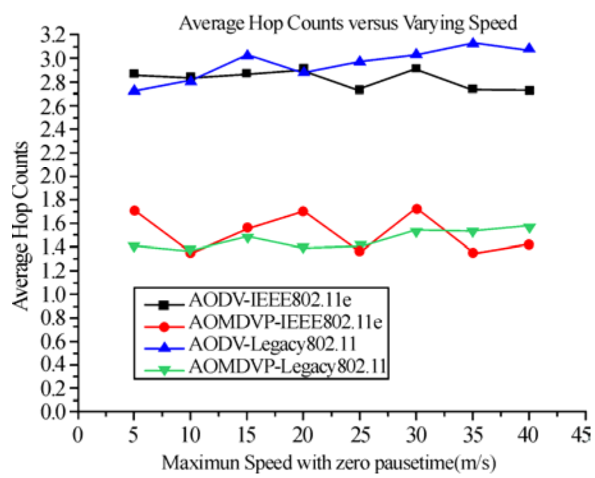
(b) Total packet loss



(e) Route discovery frequency



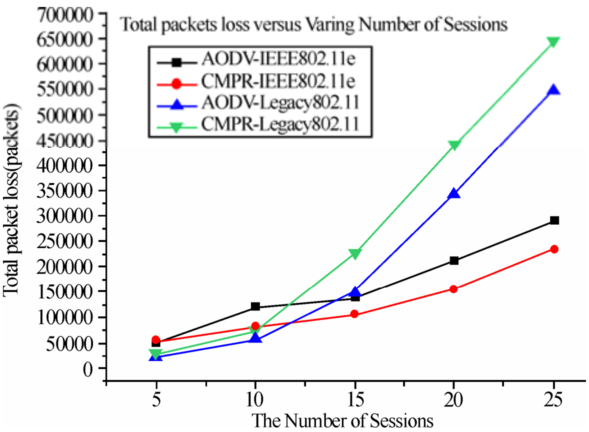
(c) Route overhead



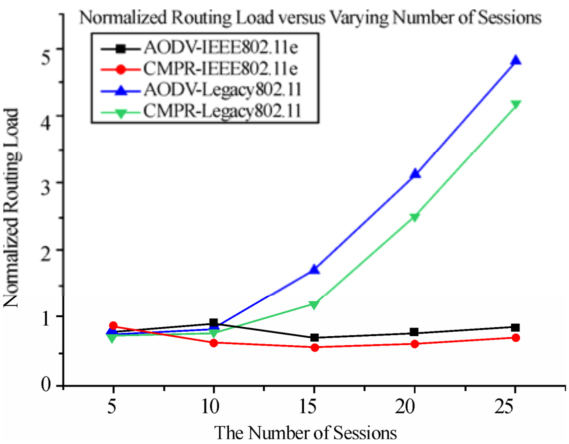
(f) Average hopcounts

Figure 5. Performance parameters with varying node moving speed.

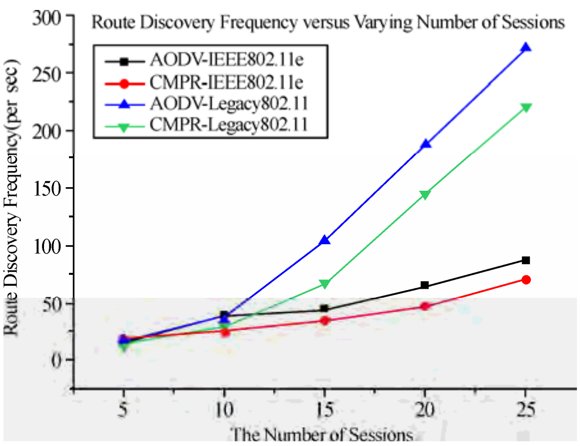
improvement. IEEE 802.11e allows wireless nodes to occupy channel for a long period of time during Transmission Opportunity (TXOP). It dramatically decrease channel overheads caused by interception, Inter-frame, backoff and competition. These approaches are important for enhanced performance of delay.



(a) Total packet loss



(b) Normalized routing load



c) Route discovery frequency

Figure 6. Performance parameters with varying number of session.

The number of packets loss is shown in Figure 5(b). Using CMRP, the simulation result shows that total packet loss decrease much more pronounced in comparison with AODV. CMRP with legacy 802.11 has the least amount of packet loss. Next is CMRP with IEEE 802.11e, followed by AODV with Legacy 802.11 and AODV with IEEE 802.11e respectively. It indicates that CMRP can not only extend path lifetime, but also improve path reliability. However, this performance tendency of two routing protocol is different from the tendency of average end-to-end delay; note that CMRP with legacy 802.11 has the least amount of packet loss. The IEEE 802.11 standard defines two Medium Access Control (MAC) protocols, namely Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), and Request-To-Send/Clear-To-Send (RTS/CTS). The RTS/CTS mechanism, which is included in the legacy 802.11 model, has never been used in the TKN802.11e model of our simulation experiments. Without the four ways handshake mechanism (RTS/CTS/DATA/ACK), the average end-to-end delay achieve improvement to some extent since the channel overhead of RTS/CTS will increases data packet delay. However, without RTS/CTS mechanism, the probability of packet loss has increased in IEEE 802.11e.

The battery power of nodes in MANETs limited, so the route overhead is an important metric for extending overall network lifetime. As shown in Figure 5(c), basically, the number of control packets increases with the node mobility level for both AODV and CMRP. CMRP with IEEE 802.11e has the least amount of route overhead. Next is AODV with IEEE 802.11e, followed by CMRP with Legacy 802.11 and AODV with Legacy 802.11 respectively. Two routing protocol produces a slight difference at a low speed. However, CMRP achieves a remarkable reduction in route overhead at medium and high speed. By constructing multiple paths in one route query round, CMRP increases the average time between RREQ processes, thus effectively reducing the amount of broadcasting messages. Using CMRP, the source node will receive multiple reply messages in one route query round. Although this is a disadvantage for CMRP, the route overhead still descends as a whole.

Figure 5(d) presents the performance of normalized routing load. This metric has a similar tendency with route overhead. CMRP in IEEE 802.11e improves this performance shown at around 50% comparing to AODV in Legacy 802.11. Using CMRP, the route overhead has achieved reduction as possible. This is important for improvement of normalized routing load. On the other hand, the availability of alternate routes reduces the data packets loss and retransmission. This contribution also enhances the performance of normalized routing load. As a whole, CMRP with IEEE 802.11e has the least normalized routing load. Next is AODV with IEEE

802.11e, followed by CMRP with Legacy 802.11 and AODV with Legacy 802.11 respectively.

Figure 2(e) illustrates the simulation result on route discovery frequency. This metric has a similar tendency with two metric as mentioned above. CMRP with IEEE 802.11e has the least frequency of route discovery. CMRP in IEEE 802.11e improves the performance of route discovery frequency at around 50% comparing to AODV in Legacy 802.11. By reducing the amount of broadcasting messages, CMRP achieves remarkable reduction in route overhead. On the other hand, CMRP maintains multiple paths longer than AODV; so that CMRP increases the interval between route query processes and suspends link failures. As expected, CMRP performs better than AODV does for both Legacy 802.11and IEEE 802.11e.

The average hopcounts is shown in Figure 5(f). CMRP descend the average hopcounts at around 50% for both Legacy 802.11 and IEEE 802.11e. CMRP maintains multiple routes longer than that of AODV. This step help CMRP descends the counts of new route discovery, so that the hopcounts of routes for data sending has achieved reduction as possible.

5.2. Performance with Varying Sessions

Figure 6 shows the three performance metrics as a function of varying sessions in experiment II. We vary the number of sessions from 5 to 25 in order to compare performance of CMRP and AODV when offered load increases. The max speed of node mobility is 10 m/s.

As shown in Figure 6(a), the simulation result shows that the number of packet loss for both AODV and CMRP increases as the offered load increases. Two routing protocol perform alike at a low speed. However, CMRP achieves a remarkable improvement in packet loss at medium and high offered load. At a high offered load, CMRP in IEEE 802.11e descend the number of packet loss at around 60% comparing to AODV in Legacy 802.11. As mentioned in Section 5.1, CMRP with legacy 802.11 has the least amount of packet loss (see Figure 5(b)). This simulation result illustrates that the number of packet loss of CMRP in legacy 802.11 will beyond that of CMRP in IEEE 802.11e at medium and high number of sessions. As expected, the variation of AODV and CMRP performance in legacy 802.11 shows a rapidly growing tendency with increase of offered load. As a whole, CMRP with IEEE 802.11e has the least amount of packet loss. Next is AODV with IEEE 802.11e, followed by AODV with Legacy 802.11 and CMRP with Legacy respectively. This result shows CMRP with IEEE 802.11e is able to lower the number of packet loss effectively even at a high offered load.

Figure 6(b) presents the performance of normalized routing load. At very low sessions, both protocols perform alike. This is because link failure rates are very high, compared to offered load. As the number of session is increased beyond the rate of link failures, CMRP begins to indicate its preferred performance since CMRP can provide more reliable route. CMRP with IEEE 802.11e has the least amount of packet loss. At a very high offered load, CMRP in IEEE 802.11e descend the normalized routing load at around 80% comparing to AODV in Legacy 802.11.

Figure 6(c) plots the simulation result on route discovery frequency. This metric has a similar tendency with normalized routing load. CMRP with IEEE 802.11e has the least frequency of route discovery. At a very high offered load, CMRP in IEEE 802.11e descend the normalized routing load at around 70% comparing to AODV in Legacy 802.11. Maintaining multiple reliable paths for CMRP is important for enhanced performance, thus reducing route discovery frequency.

6. Conclusions

In this paper, an improved Cross-layer Multipath Routing Protocol (CMRP) for IEEE 802.11e-based MANETs was proposed. CMRP uses the node energy from the physical layer to make better routing decision and path quality. The nodes' mobility behavior is predicted using a notion of "Signal Fading Degree, SFD". Especially, we determine that the IEEE 802.11e makes a significant contribution to performance improvement of CMRP. The IEEE 802.11e standard was developed to offer QoS capabilities to WLANs, offering significant improvements to multimedia traffic. MANETs will also benefit from this new technology. Our simulation experiment results demonstrate that, in combination of the IEEE 802.11e standard in MAC layer, CMRP provides significant performance improvement in term of average end-to-end delay, packet loss, route overhead, normalized routing load, route discovery frequency, and so on. Our ongoing work focuses, on the one hand, on the more realistic simulation setup to analyze and evaluate the performance of the proposed scheme. On the other hand, we will try to improve the performance of wireless media streaming using reliable multipath routing policy.

7. References

[1] H. Shen and G. W. Bai, "A Cross-layer based multipath routing protocol for IEEE 802.11e mobile ad hoc networks," Proceedings of IEEE WiCOM, Shanghai, China, pp. 1554–1557, September 2007.

[2] G. Bai, K. Oladosu, and C. Williamson, "Performance benchmarking of wireless web servers," Journal of Ad Hoc Networks (Elsevier), 5(3), pp. 392–412, 2007.

[3] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," Proceedings of ACM SIGCOMM, Portland, Oregon, USA, pp. 133–144, August 2004.

[4] A. Sankar and Z. Liu, "Maximum lifetime routing in wireless ad-hoc networks," Proceedings of IEEE INFOCOM, Hong Kong, China, pp. 1089–1097, March 2004.

[5] C. Perkins, E. Belding-Royer, and S. R. Das, "Ad hoc on-demand distance vector (AODV) routing," Request for Comments 3561, MANET Working Group, July 2003, work in progress: http://www.ietf.org/rfc/rfc3561.txt.

[6] D. Johnson, D. Maltz, and Y. Hu, "The dynamic source routing protocol for mobile ad hoc networks," Internet Draft, MANET Working Group, work in progress: http://www.rfc-editor.org/rfc/rfc4728.txt.

[7] M. K. Marina and S. R. Das, "On-demand multipath distance vector routing for ad hoc networks," Proceedings of IEEE ICNP, Riverside, California, pp.14–23, November 2001.

[8] Z. Ye, S. V. Krishnamurthy, and S. K. Tripathi, "A Framework for reliable routing in mobile ad hoc networks," Proceedings of IEEE INFOCOM, San Francisco, CA, USA, pp. 270–280, March 2003.

[9] A. Valera, W. Seah, and S. Rao, "Cooperative packet caching and shortest multipath routing in mobile ad hoc networks," Proceedings of IEEE INFOCOM, San Francisco, CA, USA, pp. 260–269, March 2003.

[10] X. Li and L. Cuthbert, "A reliable node-disjoint multipath routing with low overhead in wireless ad hoc networks," Proceedings of ACM MSWiM, Venezia, Italy, pp. 230–233, October 2004.

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

[12] W. Wei and A. Zakhor, "Robust multipath source routing protocol (RMPSR) for video communication over wireless ad hoc networks," Proceedings of IEEE ICME, Taipei, Taiwan, China, pp. 1379–1382, June 2004.

[13] A. Nasipuri, R. Castaneda, and S. Das, "A performance of multipath routing for on-demand protocols in mobile ad hoc networks," ACM/Kluwer Mobile Networks and Applications 6, pp.339–349, 2001.

[14] V. Srivastava and M. Motani, "Cross-layer design: A survey and the road ahead," IEEE Communications Magazine, 43 (12), pp. 112–119, 2005.

[15] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," IEEE Communications Magazine, 12 (1), pp. 3–11, 2005.

[16] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," IEEE Journal on Selected Areas in Communications, 24(8), pp. 1452–1463, 2006.

[17] M. Li, L. Zhang, and V. O. K. Li, "An energy-aware multipath routing protocol for mobile ad hoc networks," Proceedings of ACM SIGCOMM Asia Workshop, Beijing, China, pp. 166–174, April 2005.

[18] H. Sun and H. D. Hughes, "Adaptive QoS routing by cross-layer cooperation in ad hoc networks," EURASIP Journal on Wireless Communications and Networking 5, pp. 661–671, 2005.

[19] C. Chen, W. Wu, and Z. Li, "Multipath routing modeling in ad hoc networks," Proceedings of IEEE ICC, Seoul, Korea, pp. 2974–2978, May 2005.

[20] R. Banner and A. Orda, "Multipath routing algorithms for congestion minimization," IEEE/ACM Transactions on Networking, 15(2), pp. 413–424, 2007.

[21] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "High-throughput path metric for multi-hop wireless routing," Proceedings of ACM MobiCom, San Diego, California, USA, pp. 134–146, September 2003.

[22] P. P. Pham and S. Perreau, "Performance analysis of reactive shortest path and multi-path routing mechanism with load balance," Proceedings of IEEE INFOCOM, San Francisco, CA, USA, pp. 251–259, March 2003.

[23] Y. Ganjali and A. Keshavarzian, "Load balancing in ad hoc networks: Single-path routing vs. multi-path routing," Proceedings of IEEE INFOCOM, Hong Kong, China, pp. 1120–1125, March 2004.

[24] A. Tsirigos and Z. J. Haas, "Analysis of multipath routing, Part 1: The effect on the packet delivery ratio," IEEE Transactions on Wireless Communications, 3(2), pp. 138–146, 2004.

[25] A. Tsirigos and Z. J. Haas, "Analysis of multipath routing, Part 2: Mitigation of the effects of frequently changing network topologies," IEEE Transactions on Wireless Communications, 3(2), pp. 500–511, 2004.

[26] C. T. Calafate, P. Manzoni, and M. P. Malumbres, "On the interaction between IEEE 802.11e and routing protocols in mobile ad-hoc networks," Proceedings of IEEE Euromicro-PDP, Lugano, Switzerland, pp. 110–117, February 2005.

[27] K. Fall and K. Varadhan, ns notes and documents, The VINT Project, UC Berkeley, LBL, USC/ISI, and Xerox PARC.

[28] S. Wiethölter and C. Hoene, "IEEE 802.11e EDCA and CFB simulation model for ns-2": http://www.tkn.tuberlin.de/research/802.11e_ns2.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www. SciRP.org/journal/ijcns/).

Comparing the Accuracy of Network Utilization Performance between Real Network and Simulation Model for Local Area Network (LAN)

Mohd Nazri ISMAIL[1], Abdullah Mohd ZIN[2]
[1]*Faculty of MIIT, University Kuala Lumpur, Malaysia*
[2]*Faculty of Computer Science, UKM, Malaysia*
Email: mnazrii@miit.unikl.edu.my
Received March 6, 2008; revised August 22, 2008; accepted October 8, 2008

Abstract

This article presents a novel approach for the measurement and estimation of network traffic utilization between network nodes in heterogeneous environment. This research investigates performance evaluation of network interface on heterogeneous services and technologies environment. This study proposes an enhanced equation to evaluate the performance of network interface via Little Law and Queuing theories to improve the evaluation algorithm. To get accuracy results on the performance of simulation model, it measures (verify and validate) data from Local Area Network (real network environment). This project uses network management tool to capture those data and Fluke Optiview device to generate traffic. As a result, this simulation model can provide a good approximation of the real traffic observed in the real network environment. Through laboratory and field experiments, the result shows that the model via simulation is capable of approximating the performance of network utilization and traffic over heterogeneous services and techniques within a minimum error range.

Keywords: Network Utilization, Real Network, LAN

1. Introduction

Considerable research has been conducted to model and quantify the performance of heterogeneous services and technologies (e.g., [1–3]). Accurate measurements and analyses of network characteristics are essential for robust network performance and management. However, no current research specifically focuses on using queuing theory to measure heterogeneous services and technologies performance, which is the object of this research. Queuing theory [4] has been used as an effective tool to model performance in several technical and social contexts. Evaluating the performance of a computer networking usually involves constructing an appropriate model to predict the heterogeneous environment behaviour via simulation model. The heterogeneous environment model is then analyzed and simulated using mathematical techniques. For example, several flow-level network traffic models have been proposed to describe/stimulate [5–7]. These models have been used to study fairness, response times, queue lengths and loss probabilities under different assumptions and using a variety of mathematical techniques. Queuing theory has been widely used to model and analyze the network performance of complex systems involving services, communication systems, computer networks and vehicular traffic. In contrast to other works in the literature (e.g., [8–10]), developed simulation model to measure the performance of heterogeneous environment. Our model can be used to generate representative packet traffic in a live network environment or in a simulated environment.

The significant of this study was to develop a simulation model to measure the performance of network traffic utilization in heterogeneous network environment using Queuing theory. This model could assist network administrators to design and manage heterogeneous network systems. This simulation model can be used in various services and technologies to measure heterogeneous environment. Therefore, this simulation model is designed to: 1) predict the performance of various services (e.g. video, audio, voice and message) in order to aid technology assessment and capacity planning; 2) predict the expected behavior of new services and designs through qualitative or quantitative estimates of network performance; 3) assist network administrator to prepare, propose, plan and design network topology more effective and systematic; and 4) conduct "What-If" analysis for evaluating heterogeneous network environment performance. Moreover, in the future, the integration of data and communication services, almost every "Internet Ready" device will be a communicable device [11]. With the availability of this infrastructure, users are now demanding and expecting more services [12,13]. Convergence is pushing towards an environment that requires new investment in infrastructure and able to support the delivery of rich services (various services), applications and content [5,14]. In addition, more people are using multimedia services such as MMS, WAP, i-mode or push-to-talk. GPRS (General Packet Radio Service) is an overlay on GSM networks that allows this kind of end-to-end IP-based packet traffic from mobile devices to the Internet [15]. Network deployment is growing increasing complex as the industry lashes together a mix of wired and wireless technologies into large-scale heterogeneous network architecture and as user applications and traffic continue to evolve. Faced with this growing complexity, network designers and researchers almost universally use simulation in order to predict the expected performance of complex networks [16]. The successful evolution of the Internet is tightly coupled to the ability to design simple and accurate models [17]. Many factors may contribute to the congestion of network interface, such as a heavy load in the network that usually generates higher traffic. Once the number of requests exceeds the maximum capability of network, many clients will not able to receive responses from the network [18]. Thus, this research is critical to be conducted in order to predict and measure of network traffic utilization in heterogeneous environment.

2. Problem Statements

In the 21 century, a network infrastructure is based on multi-service implementation over convergence of network medium such as ISP, PSTN and GSM [19,20]. Availability of various services has produced multi-traffic in network infrastructure. Therefore, multi-traffic in the network infrastructure has become more complex to observe and analyze [14,21,22]. Today, retrieving and sending information can be done using a variety of technologies such as PC, PDA, fix and mobile phones via the wireless, high speed network, ISDN and ADSL lines that are more prone to heterogeneous environment, but unfortunately the optimal capability of technologies are not fully realized. The main factors of network congestion are related to network design and bandwidth capacity [23]. Nevertheless, few studies have been conducted to evaluate the application of computer network technologies and services over heterogeneous environment especially in Higher Education Institutes. Algorithms for actively measuring network physical and available bandwidths have been researched for many years. Many tools have been developed, and only a few tools have successfully achieved a close estimation of network bandwidths [3]. Therefore, retrieving and sending information over heterogeneous environment using convergence of technologies in Higher Educational Institutes should be analyzed and evaluated via simulation model. This research has setup a pilot test-bed (real network environment) to analyze and measure of network traffic utilization at University of Kuala Lumpur in Malaysia. This study posits several research questions: 1) what is the performance level of the network utilization and traffic; and 2) Is the simulation model for evaluating and measuring the heterogeneous environment performance effective?

3. Methodologies

Whatever modeling paradigm or solution techniques in heterogeneous environment model development are being used, the performance measures extracted from a simulation model must be a good representation of the real network environment. Inevitably, some assumptions must be made about the real network in order to construct the heterogeneous environment model. Figure 1 shows the overall framework of the simulation model. There are four performance techniques to validate the simulation model: 1) graphical representation; 2) tracing; 3) parameter variability; and 4) predictive validation. In addition, there are two techniques to judging how good a model is with respect to the real network: 1) it must ascertain whether the simulation model implements the assumptions correctly (model verification); and 2) assumptions which have been made are reasonable with respect to the real network (model validation). Comparison with a real network is the most reliable and preferred method to validate a simulation model (refer to Figure 2). Assumptions, input values, output values, workloads, configurations and network system behaviour should all be compared with those observed in the real network.

4. Propose Simulation Model Development for Network Utilization and Traffic

Many different types of modeling and simulation applications are used in various disciplines such as acquisition, analysis, education, entertainment, research and training [24]. In the Figure 3, theoretical model is based on a random distribution of service duration. "Request" defines the way clients use the computer network to request services, while, "Response" represents the way

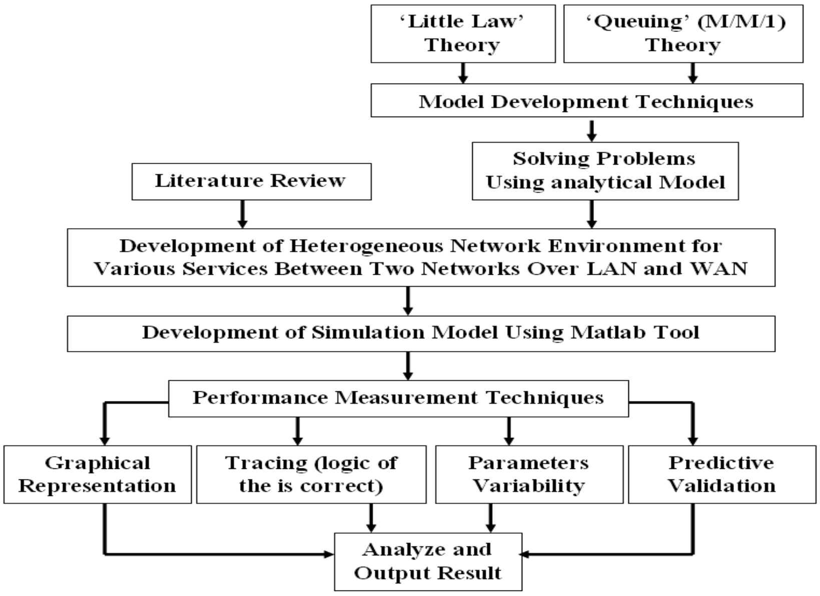


Figure 1. Simulation model development methodology.

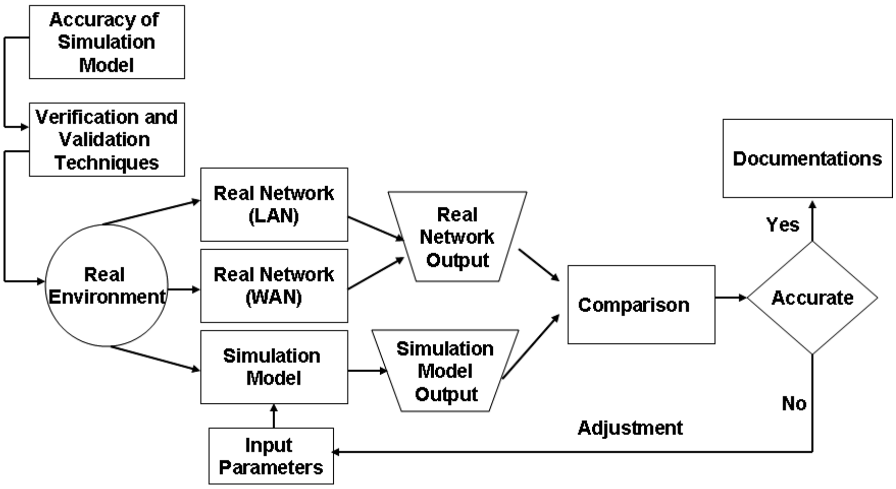


Figure 2. Simulation model verification and validation methodology.

clients receive services from the server. Simulation model is divided as follows: 1) to study physical of real heterogeneous network environment; 2) transform physical of real heterogeneous network environment into logical model; and 3) develop and implement the heterogeneous simulation model.

4.1. Physical Model of Real Heterogeneous Network Environment

Figure 3 shows the network heterogeneous environment in real world. Then, it needs to transform from heterogeneous environment in real world into logical model. The logical model is the phase where mathematical techniques are used to stimulate heterogeneous environment.

4.2. Logical Model of Heterogeneous Network Environment

Figure 4 depicts the open queuing network based on Queuing theory (M/M/1) will use to develop logical model of heterogeneous network environment for network traffic utilization. Queuing theory is robust enough to include many different combinations. Parameters like bandwidth capacity, size of packet services and number of clients are used to "characterize" the application traffic.

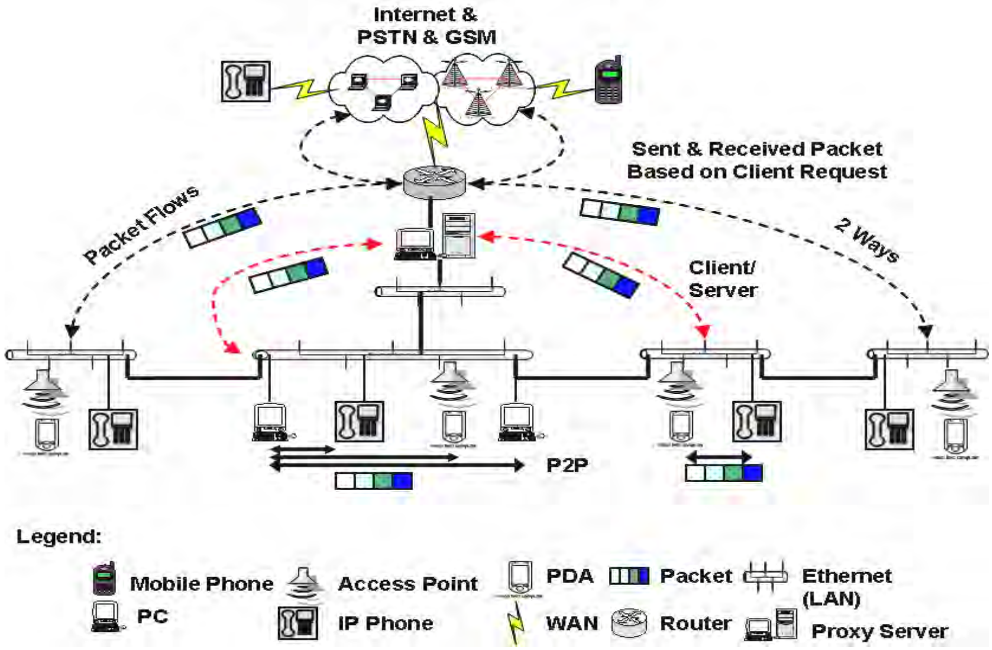


Figure 3. Real heterogeneous network environment at main and branch campus.

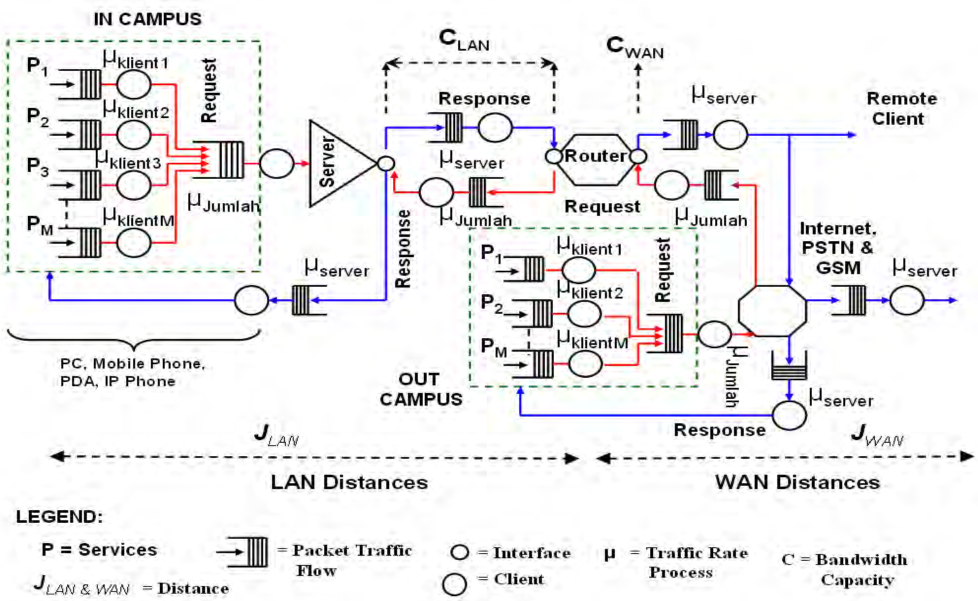


Figure 4. Logical model of heterogeneous environment at main and branch campus.

4.3. Development of Heterogeneous Network Environment Model

This section describes a simple analytical queuing and little law theories that capture the performance characteristics of network utilization and traffic operations. A link refers to a single connection between routers and hosts. The link bandwidth is the rate at which bits can be inserted into the medium. The faster bandwidth the more bits can be placed on the medium in a given time frame [25]. Table 1 shows the parameters that have been used in the model development. In open queuing network, the throughput of the heterogeneous network environment is determined by the input rate in the system. Table 2 summarizes all the parameters used in the model.

The original Queuing theory is defined as an average number of clients in the system (variable name is "N") in Equation (1). Equation (2) is defined as traffic intensity use by clients in the system. Equation (1) and (2) are derived based on logical model that has been designed to meet requirements for heterogeneous network environment. Logical model is derived and formulated in a single service (homogeneous concept) as in Equations (3), (4), (5), (6) and (7). Then, the logical model is derived to the heterogeneous network environment in Equations (8), (9), (10), (11), (12), (13) and (14).

Figure 5 shows how the model has been formulated

$$N = \lambda * T \quad \textbf{(1)}$$

$$\rho = \frac{\lambda}{\mu} < 1; \lambda < \mu \quad \textbf{(2)}$$

from real network environment to simulation model. The main valuable aspects of the simulation study is to explain and understand real world phenomena that are costly to perform in the laboratory or difficult to collect in field experiments. A successful simulation model that is able to provide a sufficiently credible solution that can be used for prediction. Since it is not feasible to construct a simulation model that represents all the details and behaviors of the real network, some assumptions must be made about the real network to construct a simulation model. Therefore, a simulation model is an abstract representation of real network environment.

Table 1. Notations for original queuing and little theories.

Model Parameters	Meaning
N	*Average number of clients in the system*
T	*Average time a client spends in the system (second)*
λ	*Clients arrival rates*
μ	*Service rate in second*
$1/\mu$	*Mean service times*
ρ	*Traffic intensity*

Table 2. Notations for model development.

Model Parameters	Meaning
N	*Average number of traffics on the network*
T	*Average time of clients arrive on the network (second)*
P (P1,P2,P3,...Pm)	*Various of services*
P1	*Client uses single service*
μ_{klient}	*Size of packet service request by client (traffic)*
μ_{server}	*Traffic response from server to clients*
$N_{klient + server}$	*Number of clients in second over single service*
$U_{klient + server}$	*Network traffic utilization usage based on number of clients in second over single*
C (C_{LAN}, C_{WAN})	*Size of Bandwidth on LAN and WAN interface ports*
$U_{hetergenes}$	*Network traffic utilization usage for heterogeneous environment*
$Heter_{klient + server}$	*Number of clients and traffics over heterogeneous environment*
μ_{Jumlah}	*Total size of packet services request by clients (traffic)*
Jum_klient	*Number of clients*
T_{jum}	*Total number of clients access on the network in second*

Client uses single service for accessing network server

$$N = \mu_{klient} * (Jum\_klient * T) \quad \textbf{(3)}$$

$$N_{klient + server} = P_1 * (Jum\_klient * T) \quad \textbf{(4)}$$

$$N_{klient + server} = (Minta + Balas) * (Jum\_klient * T) \quad \textbf{(5)}$$

$$N_{klient + server} = (\mu_{klient} + \mu_{server}) \cdot (Jum\_klient * T) \quad \textbf{(6)}$$

$$U_{\text{klient + server}} = N_{\text{klient + server}} / C \tag{7}$$

Client uses various services for accessing network server in Heterogeneous Environment

$$T_{\text{Jum}} = \text{Jum\_klient} * T \tag{8}$$

$$\text{Heter}_{\text{klient + server}} = (P_1 + P_2 + P_3 + ... + P_m) * T_{\text{Jum}} \tag{9}$$

$$\text{Heter}_{\text{klient + server}} = [(\mu_{\text{klient1}} + \mu_{\text{klient2}} + \mu_{\text{klient3}} + ... + \mu_{\text{klient\_m}}) + \mu_{\text{server}}] * T_{\text{Jum}} \tag{10}$$

$$\text{Heter}_{\text{klient + server}} = [(\mu_{\text{Jumlah}}) + (\mu_{\text{server}})] * T_{\text{Jum}} \tag{11}$$

$$\text{Heter}_{\text{klient + server}} = [(\mu_{\text{Jumlah}}) + (\mu_{\text{server}})] * T_{\text{Jum}} \tag{12}$$

where $P_1+P_2+P_3+\ldots+Pm = U_{\text{klient1}}+U_{\text{klient2}}+U_{\text{klient3}}+\ldots+U_{\text{klientm}} = \mu_{\text{jumlah}}$ **(13)**

$$U_{\text{hetergenes}} = \text{Heter}_{\text{klient + server}} / C \tag{14}$$

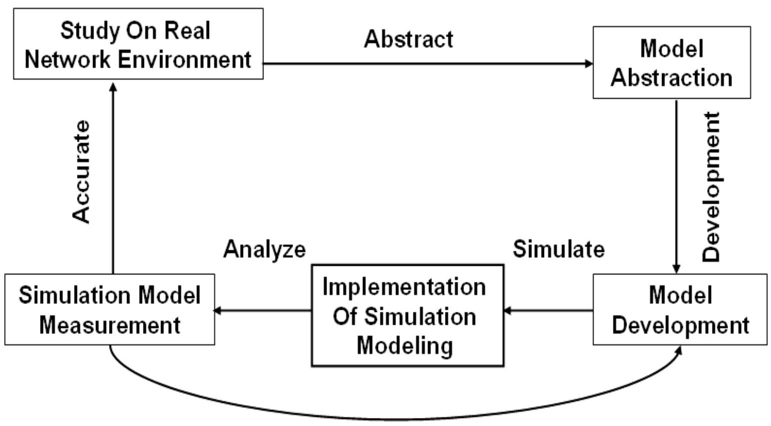


Figure 5. Model and simulation development phases.

5. Verification and Validation of Simulation Model with Real Network Experimental

Lab experiment is based on ideal network in which there is no packet losses, no jitter in delays and network bandwidth is sufficient for all requirements. While, real experiment is based on real network and need to consider as follows: 1) network bandwidth is limited and is not enough for all application and users at the same time; 2) delay due to the network overloads; and 3) packet losses.

5.1. Real Network Setup

This research used a network management application to capture traffic between two networks link in real network environment. Figure 6 shows the experimental setup of real network used in our tests. The real network used switch with Gigabit Ethernet ports, Router ports and Fluke Optiview device can be configured to insert size of packet services and number of clients to generate traffic into the network interface (see Figure 7). By using varying number of clients and size of packet services, Fluke Optiview device is able to simulate network utilization and traffic.

5.2. Real Network Experiment

This research has setup a real network environment of network utilization measurement that generates measurement data to analyze network performance at the main campus. The real network is based on local area network (LAN). The traffics will pump into LAN 100 Mbps (real network) to access network server (see Figure 7). Low bandwidth link affects the size of packet services and number of clients' access to the network server. Therefore, network management application is used to measure traffic and its network utilization performance (see Figure 8). Five sets of experiments were conducted with different scenarios (see Table 3 and

Table 4). Fluke Optiview device is able to generate maximum traffic to 1518 bytes (12144 bits) only in the real network (see Figure 7 and Figure 10). The same input variables have been used in simulation model (see Figure 9 and Table 4) to estimate our data that must be closely resemble to real network environment (see Table 5). This research is concluded that base on our findings, the simulation model is able to predict and estimate network utilization usage for real network environment (see Table 3 and Table 4).

5.3. Comparison of Real Network, Simulation Model and Relative Error Rates

Figure 11 shows a comparison between simulation model, real network and relative error rate using LAN 100 Mbps. The result shows both scenarios use in simulation model and real network are able to predict and measure network traffic utilization. The simulation model provides relatively accurate results when compared to the real network over LAN 100 Mbps. Figure 11 also shows the

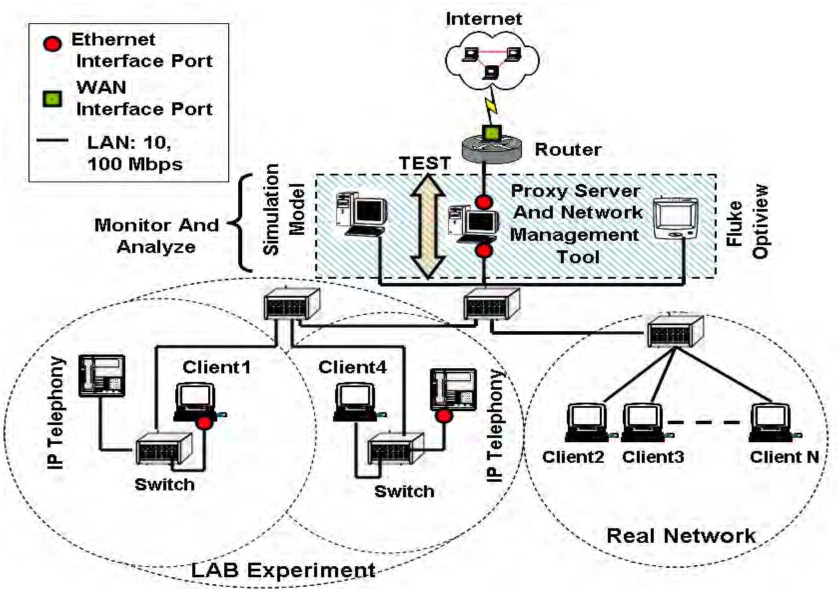


Figure 6. Experimental laboratory for real network environment setup.

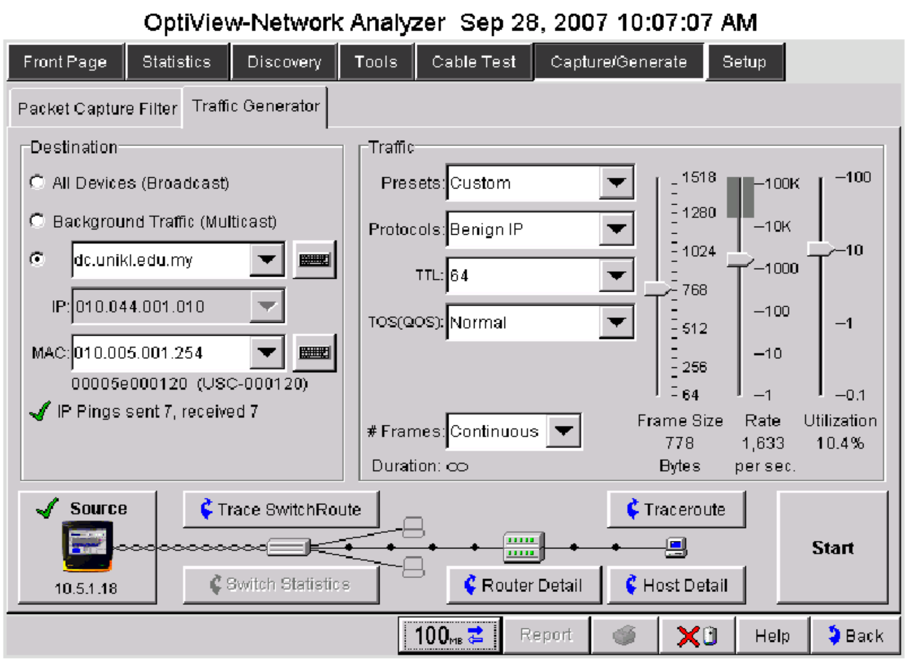


Figure 7. Fluke Optiview engine setting for size of packet services and clients.

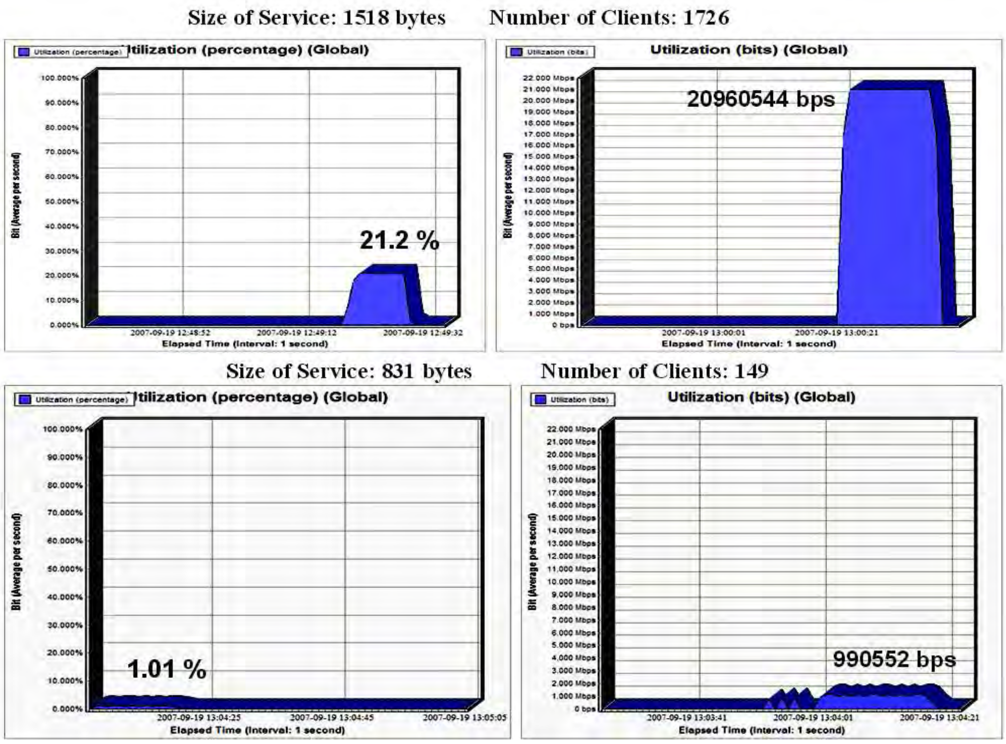


Figure 8. Real network experiment result capture by network management application (100 mbps).

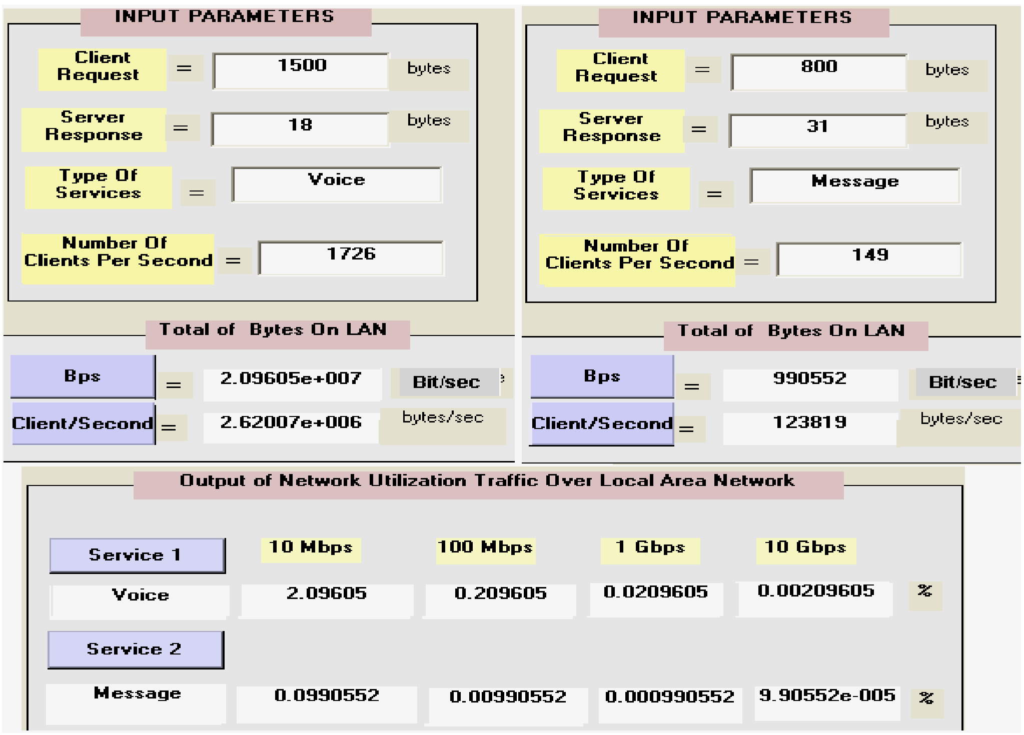


Figure 9. Prediction of network traffic utilization over 100 mbps via simulation model.

comparison of relative error rates between simulation model and real network environment. As a result, this research shows that the simulation model can predict real network experiments with minimum relative error rates. Therefore, from the prediction and estimation result, this simulation model can assist network administrator

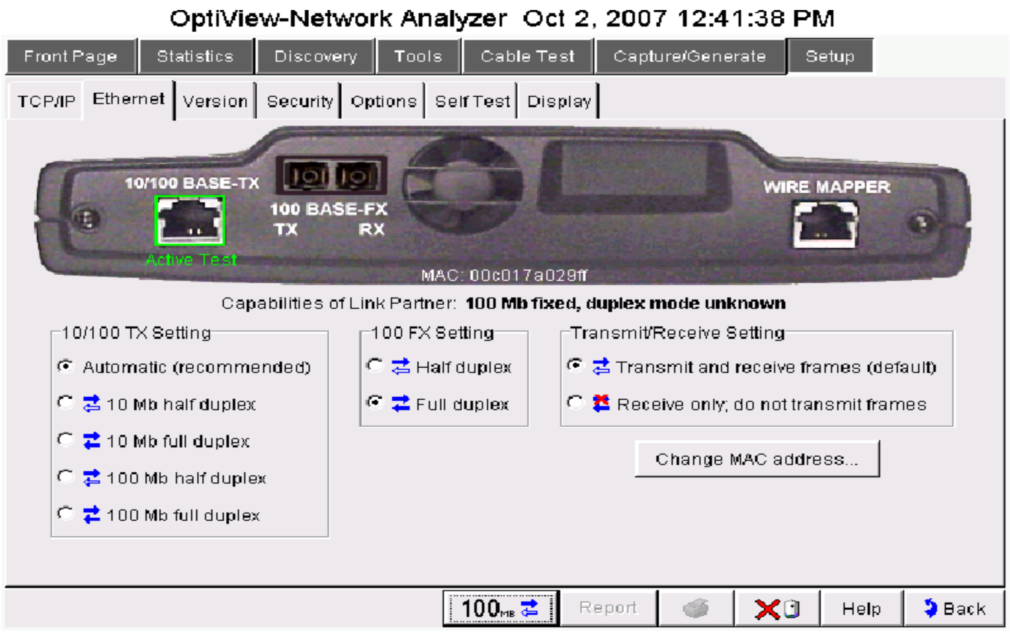


Figure 10. Fluke Optiview device architecture.

Table 3. Network traffic utilization over 100 mbps for real network experiments.

Frame Size (Bytes)	Number of Clients (second)	Traffic in Bytes/Second	Traffic in Bit/Second	Utilization (1-100%)	Utilization (0.1–1%)
512	262	134144	1073152	1.11	0.0111
778	1633	1270474	10163792	10.4	0.104
831	149	123819	990552	1.01	0.0101
1042	3961	4127362	33018896	33.7	0.337
1518	1726	2620068	20960544	21.2	0.212

Table 4. Network Traffic utilization over 100 mbps using simulation model.

Frame Size (Client + Server); (Bytes)	Number of Clients (second)	Traffic in Bytes/Second	Traffic in Bit/Second	Utilization (1-100%)	Utilization (0.1–1%)
512	262	134144	1073152	1.07	0.01073152
778	1633	1270474	10163792	10.1638	0.101638
831	149	123819	990552	0.99	0.0099
1042	3961	4127362	33018896	33.01	0.3301
1518	1726	2620068	20960544	20.96	0.2096

Table 5. Comparison of relative error rates between simulation model and real network environment.

Frame Size (Client + Server); (Bytes)	Number of Clients (second)	Utilization (Simulation Model) (0.1–1%)	Utilization (Real Network) (0.1–1%)	Relative Error Rates
778	812	0.01073152	0.0111	0.000119
831	393	0.101638	0.104	0.002362
1033	393	0.0099	0.0101	0.0002
1518	180	0.3301	0.337	0.0069

to plan, propose and design network topology more systematic and efficiently for heterogeneous network environment.

6. Conclusions and Future Work

This article has shown how an analytical queuing model

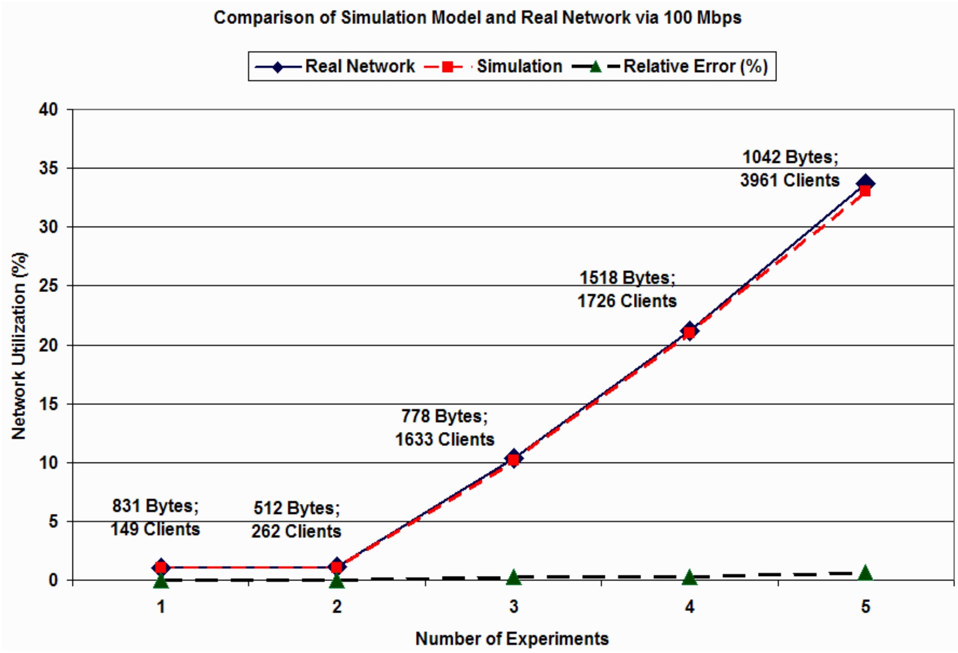


Figure 11. Comparison of simulation model with real network using 100 mbps variable and relative error rate.

can be used to understand the behaviors of heterogeneous environment over LAN experiments. The most apparent aspect is the utilization usage due to size of bandwidth and number of clients. Our simulation model, has demonstrated that it can measure accurately the performance of heterogeneous services and technologies to access network server. Through real network experiments, the simulation model is verified and validated for providing accurate performance information for various services. The simulation-modeling framework described in this study can be used to study other variations, tunings, and similar new ideas for various services and technologies. Network utilization rate will directly affect the network performance. In network management, by monitoring and analyzing network utilization rate, network administrator can monitor the performance of the network, thus to study whether network is normal, optimal or overloaded. Network utilization rate also plays an important role in benchmark setting and network troubleshooting. Future work is to develop a simulation model to analyze bandwidth capacity requirement for various services and technologies in heterogeneous environment.

7. References

[1] S. Kawasaki, Niwa, et al., “A study on formulation of the ubiquitous cloud model”, Mobile Data Management, 2006, MDM 2006, 7th International Conference, pp. 148–148, 2006.

[2] A. Tsalgatidou, G. Athanasopoulos, et al., “Developing scientific workflows from heterogeneous services,” ACM SIGMOD, Vol. 35 (2), pp. 22–28, 2006.

[3] Q. W. Liu, S. L. Zhou, et al., “Cross-layer modeling of adaptive wireless links for QoS support in heterogeneous wired-wireless networks,” Wireless Networks, Vol. 12 (4), Kluwer Academic Publishers, 2006.

[4] R. Puigjaner, “Performance modelling of computer networks,” IFIP/ACM Latin America Networking Conference 2003, La Paz, Bolivia. ACM 2003, October 3–5, 2003.

[5] S. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. Roberts. “Statistical bandwidth sharing: A study of congestion at flow level,” in Proceedings of ACM SIGCOMM ’01, San Diego, CA, August 2001.

[6] G. J. Jin and B. L. Tierney, “System capability effects on algorithms for network bandwidth measurement,” Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, pp. 27–38, 2003.

[7] J. Cong and B. E. Wolfinger, “A unified load generator based on formal load specification and load transformation,” Proceedings of the 1st international conference on Performance evaluation methodologies and tools, ACM International Conference Proceeding Series, Pisa, Italy, Vol. 180 (53), 2006.

[8] S. B. Fredj, T. Bonald, and A. Proutiere, G. Regnie and J. Roberts, “Statistical bandwidth sharing: A study of congestion at flow level,” ACM SIGCOMM, August 2001.

[9] T. Bu and D. Towsley, “Fixed point approximation for TCP behavior in an AQM network”, ACM SIGMETRICS, June 2001.

[10] A. A. Kherani and A. Kumar, “Performance analysis of TCP with nonpersistent sessions”, Workshop on Modeling of Flow and Congestion Control, INRIA, Ecole Normale Superieure, Paris, September 4–6, 2000.

[11] Q. G. Zhao, X. M. Fang, Q. Z. Li, and Z. Y. He, “WNN-based NGN traffic prediction,” ISADS 2005, Proceedings

Autonomous Decentralized Systems, pp. 230–234, 2005.

[12] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski. “Modeling Internet backbone traffic at the flow level”. IEEE Transactions on Signal Processings (Special Issue on Networking), August 2003.

[13] T. Bu and D. Towsley, “Fixed point approximation for TCP behavior in an AQM network,” in Proceedings of ACM SIGMETRICS ’01, San Diego, CA, June 2001.

[14] P. Podhradsky, “Migration scenarios and convergence processes towards NGN (present state and future trends),” Electronics in Marine Proceedings Elmar. 46th International Symposium, pp. 39–46, 2004.

[15] G. Nogueira, “Measurement based validation of GPRS/ EDGE analytical models”.

[16] J. Heidemann, K. Mills, and S. Kumar, “Expanding confidence in network simulation,” IEEE computer, Vol. 15 (5), pp. 58–63, 2001.

[17] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski, “A flow-based model for Internet backbone traffic,” Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement, Marseille, France, pp. 35–47, 2002.

[18] J. kontio and R. Conradi, “Evaluating the performance of a web site via queuing theory,” Software Quality - ECSQ 2002: 7th International Conference, Helsinki, Finland, pp. 63–72, Springer-Verlag Berlin Heideberg, June 9–13, 2002.

[19] K. H. Lee, K. O. Lee, K. C. Park, J. O. Lee, and Y. H. Bang, “Architecture to be deployed on strategies of next-generation networks,” Communications ICC ’03, The IEEE, Vol. 2, pp. 819–822, 2003.

[20] J. Q. Li, Z. H. Sun, 2004, “Internet/Web technology in higher education in China,” Advanced Learning Technologies, Proceedings IEEE International Conference, pp. 993–997, 2004.

[21] X. X. Jiang, F. C. Yang, and H. Zou, “A novel architecture to customer service management for the NGN,” ICCT 2003. International Conference, 1: 123–126, 2003.

[22] B. Thai, R. Wan, A. Seneviratne, and T. Rakotoarivelo, “Integrated personal mobility architecture: A complete personal mobility solution,” Kluwer Academic Publishers, 2003.

[23] J. Curtis and T. McGregor, “Review of bandwidth estimation techniques,” in Proceedings New Zealand Computer Science Research Students’ Conference, Vol. 8, New Zealand, April 2001.

[24] D. R. Gerhan and S. M. Mutula, “Bandwidth bottlenecks at the University of Botswana,” Published by Emerald Group, Vol. 23 (1), pp. 102–117, 2005.

[25] J. Sommers and P. Barford, “Self-configuring network traffic generation,” in Proceedings of ACM Internet Measurement Conference, October 2004.

I. J. Communications, Network and System Sciences, 2008, 4, 285-385
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Towards High Quality VoIP in 3G Networks an Empirical Approach

Andres ARJONA[1], Cedric WESTPHAL[2], Antti YLÄ-JÄÄSKI[3], Martin KRISTENSSON[1], Jukka MANNER[3]
[1] Nokia Siemens Networks, Finland, [2] DoCoMo Labs, USA, [3] Helsinki University of Technology, Finland
Email: {andres.arjona, martin.kristensson}@nsn.com, cwestphal@docomolabs-usa.com, {antti.yla-jaaski, jukka.manner}@hut.fi
Received August 22, 2008; revised October 10, 2008; accepted October 12, 2008

Abstract

Third generation (3G) packet switched WCDMA networks with high-speed downlink packet access (HSPDA) are currently being deployed worldwide to provide wireless broadband connectivity. When introducing HSDPA in 3G networks the end user experience and system capacity with voice over IP applications improve considerably. When later on adding also high-speed packet uplink access (HSUPA), the system capacity and end user experience will improve even further. This paper analyzes with measurements the VoIP quality over current Release 5 HSDPA networks. VoIP is expected to be a widely used application over 3G data services. The results show that even though the introduction of HSDPA significantly reduces the user-to-user voice delay, the performance is satisfactory only for selected devices. Overall, the end user experience is still significantly worse than with circuit switched solutions and is not acceptable. The current limitations with VoIP in HSDPA networks with a too large delay can be improved by using the RLC UNACK mode, potentially decreasing the jitter buffer size and reducing the terminal processing delay. In the longer term, HSUPA and several features in 3GPP Release 7 standards will bring further performance improvements in both user plan latency and system capacity.

Keywords: HSDPA, VoIP, WCDMA, Voice Quality, MOS

1. Introduction

Voice over IP (VoIP) is becoming a widely deployed service in data networks, and it will penetrate from the fixed network domain into wireless network domain. The characteristics of fixed networks and wireless networks are fundamentally different, which will impact the performance of services. In this article we analyze the VoIP service performance in wireless HSDPA and WCDMA networks.

High Speed Download Packet Access (HSDPA) [1] networks are being intensively deployed to provide broadband connectivity to mobile devices, such as handheld terminals and laptops. This broadband wireless access is able to support voice applications over a packet data connection instead of traditional circuit switched calls. With the introduction of multi-radio devices with HSDPA, WCDMA, and WiFi capabilities as well as integrated VoIP clients, ubiquitous connectivity across any of these networks is possible using the same mobile terminal. However, while the mobile terminal and client are the same, performance differs depending on the wireless access in use.

Most of the studies of VoIP over 3G network focus on simulation works. However, there is little data on the performance in actual networks. Since VoIP is expected to become a widely used application, and comes pre-configured in many current handsets, it is of great importance to better understand the performance of such application over 3G networks. We set to answer the following question: is VoIP over 3G network commercially viable with the current state of the arts networks?

This paper studies the quality of VoIP in wireless networks with multi-radio mobile devices both in the lab and in live network environment setups by conducting a methodic performance analysis based on the E-Model [2,3](we will describe the E-Model in more details in Section 2). Likewise, our study will encompass the signaling performance required for VoIP applications.

The key contribution of the paper is to characterize the performance of VoIP over 3G network, and to identify the main differences between HSDPA and WCDMA. We perform a thorough empirical evaluation of VoIP quality and signaling performance with HSDPA and WCDMA. From our evaluation, we will observe that:

- VoIP performance is acceptable in HSDPA networks only for VoIP clients on devices with enough processing power, such as laptops;
- VoIP performance is rarely acceptable in WCDMA networks, even for those high performance clients;
- WCDMA performance can be significantly improved by having retransmissions only at the BTS, not the RNC;
- The delay introduce by the end-user terminal is a critical factor in the performance.

Our study takes into consideration both the performance of the network and also the performance of real embedded VoIP clients. In addition, we validate the results of our study by comparing them to the actual performance in a densely deployed HSDPA network in Finland. Based on the results, we analyze the primary differences in performance between simulations found in the literature, our lab experiences and a live network case study. Finally, we discuss possible features that can improve the performance enough in current and future releases to support VoIP in all handheld devices.

The remainder of the paper is organized as follows. Section 2 describes our research approach, Section 3, 4 and 5 present results from a laboratory setup, a live network scenario, and for VoIP signaling performance respectively. Subsequently, in Section 6 we describe some standardization improvements. In Section 7 we discuss the available related works and finally in Section 8 we draw conclusions.

2. Methodology and Test Environments

Our experiments are composed of measurements in a HSDPA and WCDMA testbed, as well as a live HSDPA network of a Finnish operator. We are interested in measuring both the VoIP service audio quality in both laboratory and live setups and SIP signaling latencies for registering users and setting up calls.

2.1. VoIP Quality Methodology

The evaluation methodology consisted of multiple VoIP tests carried out in a radio interference free environment. These conditions were achieved in a laboratory setup by using an RF room for the BTS and clients [4]. The tests included different wireless accesses technologies and variable combinations of codecs, signal conditions, number of clients and fading profiles among others.

The main evaluation was carried out with two similar tools based on the E-Model [2], which is a ITU-T recommendation for VoIP evaluation. Firstly, with a NSN proprietary tool, which is an implementation similar to the one described in [3], and secondly, with IxChariot, which is a widely used voice evaluation tool [5]. Finally, a third tool based on the PESQ evaluation model was used to determine the average end-to-end delay with real embedded VoIP clients. With such setup, we can evaluate the performance of the different wireless access technologies based on the following test objectives:

- VoIP quality performance with the E-Model;
- Voice quality characterization for different wireless accesses, signal conditions, configurations and fading profiles;
- Benchmark of two voice quality evaluation tools based on the E-Model;
- Estimation of the average end-to-end delay when a real embedded VoIP client is used;
- Effect of simultaneous background traffic during a VoIP call;
- Characterization of delay sources and possible optimizations.

The E-Model is a voice quality evaluation model that is based on network performance metrics. It is based on a mathematical algorithm and provides an "R" performance value based on the sum of four "impairment factors" considered to be cumulative. The algorithm is depicted in Equation (1) where, "Is" is Signal to Noise Ratio, "Id" is delay (ms), "Ief" is packet loss (%), and "A" is expectation factor.

$$R = 100 - Is - Id - Ief + A \qquad \textbf{(1)}$$

In practice, ITU-T proposes to use a simplified version of this algorithm. The simplified algorithm considers that noise cancellation is encountered in the network and also dismisses the expectation factor. The expectation variable is supposed to be used to provide a balance for some environments in which the user expects a degraded quality, such as satellite connections. However, since this variable is merely subjective it is recommended to ignore it. The simplified algorithm is depicted in Equation (2).

$$R = 93.2 - Id - Ief \qquad \textbf{(2)}$$

The R value can be associated with the Mean Opinion Score (MOS) values, which is a subjective grade for voice quality based on studies carried out by ITU-T. However, even though the R-value can match a MOS value, it cannot predict the absolute opinion of an individual user.

In this paper we calculate the MOS scores with two tools based on the E-Model: a Nokia proprietary tool and IxChariot, which is a widely used tool. These tools send dummy packets that resemble VoIP packets. The packet size and transmission intervals are tied to the modeled codec. Based on the received packets, network performance values are calculated and the E-Model algorithm is applied to determine a MOS score. Figure 1 shows an overview of the environment and the E-Model based tools.

In this paper we emphasize the performance of the G.729 codec, which is the only codec supported by all the measuring tools used in this research. G.729 is also similar to AMR-NB. AMR codec is the main building block for a future codec for 3GPP based networks. ITU-T has set out standards for maximum voice quality for several codecs, including G.729 and G.711. However, there is not yet agreement on a standard AMR codec maximum quality definition in relation to the E-Model. Therefore, we can make the assumption that the performance values measured with G.729 codec are representative and are a useful basis for our analysis. In addition, G.711 is not an appropriate codec for wireless networks such as HSDPA due to its high bitrate. However, G.711 is one of the most largely supported codecs, and it is widely used in the Internet. Also, due to legacy equipment it is used in many cases, even though AMR and other lower bitrate codecs (e.g. iLBC) are encouraged. For this reason we study both G.729 and G.711.

2.2. VoIP Signaling Performance Methodology

The evaluation methodology consisted of a variety of VoIP calls using Nokia N95 terminals. We chose this terminal due to its widespread penetration in the market and because it includes an embedded VoIP client by default. This client can also be configured to work with other SIP systems (e.g. Gizmo project). We did not use a 3rd party implementation with Skype because there were no suitable clients for the N95 at the time of the study. We captured SIP packet traces directly from the mobile terminal wireless interface [6]. With such variables we evaluated the different wireless networks available from the following test objectives:

- SIP registration delays
- VoIP call signaling delays (post-dial, answer-signal, and call-release delays)

The two main activities in VoIP calls are: first, a registration to the VoIP server which is required to make and receive calls, and second, the voice call setup itself. The packet captures were carried out with a NSN proprietary tool with a function similar to TCPdump, and analyzed with Wireshark Protocol Analyzer [7].

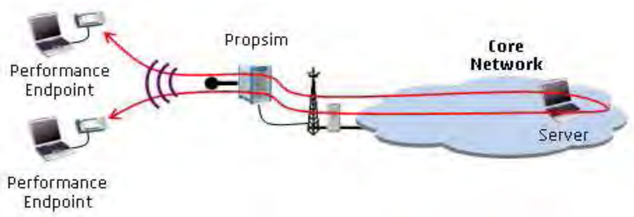


Figure 1. VoIP quality test environment.

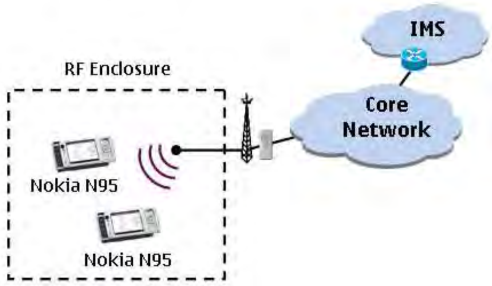


Figure 2. VoIP signaling test environment.

All the calls were carried out with two identical terminals with exactly the same setups, registered to the same VoIP server in the NSN IP Multimedia Subsystem (IMS) and via the same wireless access in an interference free environment. The measured scenarios were HSDPA-to-HSDPA, WCDMA-to-WCDMA, and WiFi-to-WiFi calls. The maximum transfer bitrates were set in the RNC and HLR configurations to model different wireless access scenarios. For WCDMA, maximum uplink and downlink transfer rates were fixed at 64/64 kbps and 128/128 kbps. For HSDPA downlink was 3.6Mbps and the uplink was fixed at 128 kbps. In the case of WiFi, transfer rates were left with default configuration (802.11g and maximum transfer rate). Figure 2 shows the test environment setup.

The core network and IMS system were privately owned and under very low load. The wireless access systems were based on NSN Release 5 equipment for HSDPA and WCDMA tests with default settings. For WiFi tests, we used a Belkin Pre-N Router with default configuration. The core network and IMS system were based on Nokia equipment. The tests executed consisted of multiple iterations of each of the voice call scenarios and registration to the VoIP server. We provide average result values from the measurements. The measurements took place during February-March 2007.

2.3. Live Network Case Study Methodology

The final stage of our study consisted of evaluating voice quality in a live HSDPA network. The network evaluation took place in the Helsinki metropolitan area, and the live network in use was provided by Elisa, Finland's largest 3G operator. The HSDPA coverage in the Helsinki metropolitan area is densely deployed and assumed to be based on NSN equipment similar to the one used in our lab measurements (Release 5 equipment). Therefore, its performance is directly comparable to our previous results. The test objectives for this phase are as follows:

- Characterize the base performance of the network (throughput and round trip time) under different signal conditions
- Evaluate the VoIP quality in different signal conditions

(excellent, medium and poor)

- Evaluate the average VoIP quality in a mobile scenario and determine the signal quality distribution for the test route

Our approach to the live network measurements was modeled in the following way. First, we made a basic network performance evaluation in different radio environments based on signal-to-noise ratio Ec/N0 levels [4].

Ec/N0 values are an objective figure for quality conditions because they take into account both signal strength and the current interference level encountered in the cell. Based on these basic network performance figures, we can evaluate the average performance in terms of maximum downlink and uplink throughput, as well as average round trip time for a particular Ec/N0 range. As a result, we are able to define three signal conditions ranges: 1) Good signal 2) Medium signal, and 3) Poor signal. Second, we evaluate the VoIP quality with the same NSN Proprietary tool used in previous tests under the three different signal conditions. This allows us to get a good metric of what is the quality in a static scenario under specific signal conditions. Third, we evaluate the average VoIP quality under a mobile scenario. The test route chosen crossed a major part of the Helsinki metropolitan area from West to East. The tests were carried out along the route in both directions twice. In addition, we measured the signal levels (Ec/N0) along the whole driving route and carry out statistical distributions for the values.

An obvious limitation of our study is the fact that due to the nature of a live network, we are not able to know or control the other user traffic that could be taking place at the same time. Therefore, we are not able to pinpoint the sources of e.g. a sudden quality drop or reduced bitrate. However, since we carried out multiple tests, our study provides a realistic view of what is the actual performance that could potentially be achieved in the field. The measurements for the live network study took place during July and August 2007.

3. VoIP Quality Analysis and Results

3.1. HSDPA/WCDMA VoIP Performance

The tests to evaluate VoIP quality involved the following variables: signal conditions, wireless access, and fading profiles. Signal conditions were modeled to provide different Ec/N0 levels by using attenuators. However, the results in this paper show that this variable does not make any sustainable difference and therefore, average result values are given instead. The wireless access technologies used were restricted to HSDPA/128, WCDMA 128/128 and WCDMA 64/64. There was no reason to use higher bitrates in this study since VoIP packets require a low bandwidth. Therefore, we emphasize the limits in which VoIP can actually be used with an adequate quality level. The bitrates were fixed and therefore, features that adapt bitrate (by increasing or decreasing) during packet switched connections were not used during the tests. Fading was applied with Propsim C8 fading simulator using Pedestrian-A 3km and Vehicular-A 30km fading profiles. The jitter buffer had a depth of 200ms and first packet play delay of 120ms. That is, all packets are delayed at least 120ms to provide a cushion for possible jitter. These are common settings in VoIP clients for wireless cellular systems. According to Wang et al. [8], a conservative jitter buffer playout delay is about 150ms.

Our results are consistent and show that the achieved quality in the HSDPA system is competitive. Based on ITU-T G.107 [2] quality was in average medium for HSDPA with both measurement tools (NSN Proprietary Tool and IxChariot). The average MOS was roughly 3.7 (see Figure 3 and Figure 4). This is a good figure especially considering that typical PSTN systems provide MOS values around 3.5. In the case of WCDMA, quality differed depending on the bitrate used. WCDMA 128/128 provided low quality and WCDMA 64/64 gave a low/poor quality level. The results also show a difference between the measurement tools. Our proprietary tool was able to differentiate more clearly the quality levels between WCDMA 128/128 and 64/64. However, IxChariot does not recognize much difference between these two bitrates. In any case, both tools show that quality in WCDMA is not optimal and is around MOS 3.0 at its best. WCDMA 64/64 MOS varied between 2.25 and 2.7. ITU states that MOS below 2.5 is not recommended for voice services and that nearly all users will be dissatisfied with such a service. Therefore, we can expect that the end user experience with VoIP WCDMA is not stable and will vary.

Table 1 presents the average end-to-end delays in the experiments (including jitter buffer playout delay). The results also show very similar performance regardless of the signal conditions modeled or the fading profile applied. The reason probably relates to fast power control mechanisms which are able to handle such changes in signal conditions in HSDPA and WCDMA. We recommend that further studies would be performed using noise or traffic generators instead of only modeling signal scenarios with attenuators.

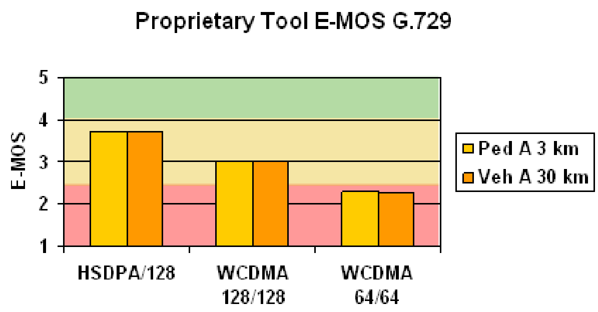


Figure 3. VoIP performance evaluation with proprietary tool.

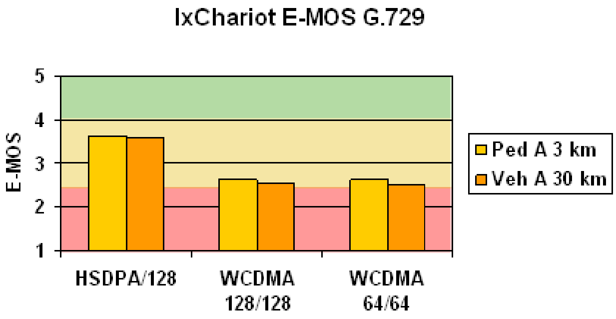


Figure 4. VoIP performance evaluation with IxChariot.

Table 1. Average VoIP end-to-end delays (including jitter buffer).

Access	Proprietary Tool		IxChariot	
	PedA 3km	VehA 30km	PedA 3km	VehA 30km
HSDPA/128	215ms	217ms	223ms	225ms
WCDMA 128/128	295ms	300ms	368ms	381ms
WCDMA 64/64	315ms	355ms	370ms	365ms

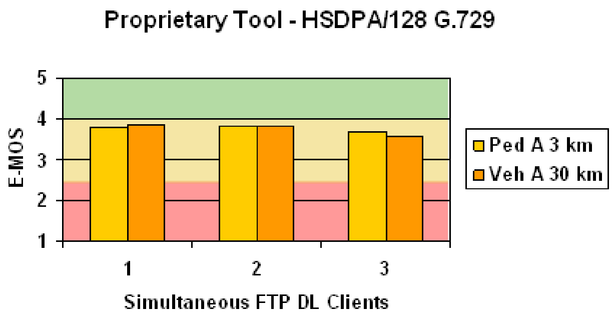


Figure 5. Effect of background traffic.

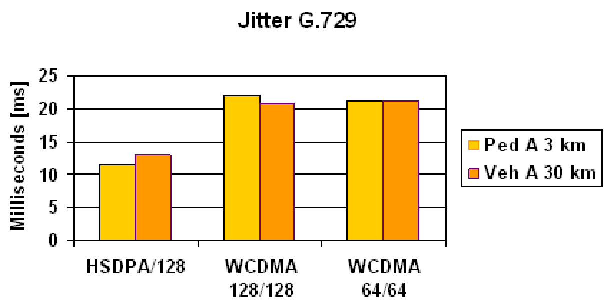


Figure 6. Jitter average.

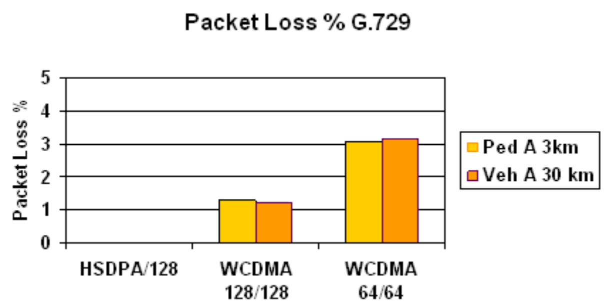


Figure 7. Packet loss percentage.

Finally, we point out that both measurement tools yield quite similar results, with exception of WCDMA 64/64. However, in this case we can observe that our proprietary tool is actually more accurate than IxChariot, especially since IxChariot does not seem to recognize any performance difference between WCDMA 128/128 and 64/64 accesses.

3.2. VoIP Performance with Simultaneous FTP Background Traffic

We also conducted some experiments where we added background traffic. The tests included a small number of simultaneous users running FTP downloads in order to evaluate if they had any effect on the VoIP performance.

As we expected, a limited number of users cannot affect VoIP quality (see Figure 5). The reason is tied to the Round Robin Scheduling used in the system, which divides bandwidth equally among users. With only 4 simultaneous users, each user will be given enough bandwidth on a timely basis (every few milliseconds). In order to measure the effect of background traffic we encourage tests with a much larger number of users, e.g. 15-20 would be required. This is out of the scope of this document. Likewise, testing different scheduling techniques such as Weighted Proportional Fair is of interest. However, there are several simulation based studies [9,10] that study VoIP capacity gains for different scheduling schemes including mixed traffic scenarios. However, note that [11] analytically showed that QoS constraints on VoIP reduce the benefit from the Proportional Fair algorithm over Round Robin scheduling.

3.3. Effect of Jitter and Packet Loss

The next test included experiments with jitter and packet loss. Jitter and packet loss are presented in Figure 6 and Figure 7. From the results we can see the average jitter and packet loss measures for different access networks.

The results show an increased jitter and packet loss for WCDMA 128/128 and 64/64. Further delay analysis shows that this increase is most likely caused by constant RLC retransmissions. RLC retransmissions have an effect on both jitter and packet loss. Every time a RLC retransmission takes place, it will cause a ~200ms delay peak. This peak can potentially fill the jitter buffer causing an overflow, which results in packet loss. Packet loss also affects voice quality. The frequency of RLC retransmissions is dependent of the access in use. Figure 8 shows an example of the RLC retransmissions (200ms delay peaks) for different wireless access technologies.

The performance of these wireless accesses would improve if RLC retransmissions are avoided as much as possible. One possibility is to use the unacknowledged mode (UNACK) feature in the RNC. The principle of operation in HSDPA [1] is such, that the BTS estimates

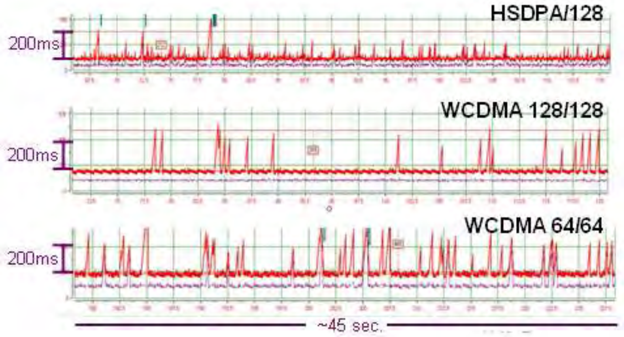


Figure 8. RLC retransmissions.

the channel quality of each user based on the physical layer feedback on the uplink. Subsequently, link adaptation and scheduling takes place at a fast pace. When the packets are first received at the BTS, they are buffered. Then, the BTS transmits the packet; however, it will still keep it in the buffer. The reason being that in case of a failure in the transmission (e.g. decoding failure), a retransmission will take place directly from the BTS without requiring any action from the RNC. This is a powerful advantage since the retransmissions are combined at the terminal. However, if there is a physical layer failure, such as a signaling error, then an RLC retransmission is required, and packets are retransmitted from the RNC (see Figure 9). This obviously results in an increase in delay, which is not beneficial for services like VoIP. While RLC retransmissions are not a very frequent event in HSDPA in static scenarios, they are more likely in mobility scenarios. In contrast, in WCDMA, all retransmissions are RLC retransmissions requiring RNC involvement. In the RLC unacknowledged mode, packets are not retransmitted even if some are lost, for example due to cell change operation [1].

3.4. Codec Performance Evaluation

Even though our study focus was on low bit rate codecs (e.g. AMR or G.729), we also evaluated the performance of the G.711 codec. Using G.711 codec in wireless environments is not encouraged due to its higher bit rate. However, since it is one of the most widely supported codecs, there are cases in which it will be used due to other codec incompatibilities. The performance was measured with a proprietary tool. Tests with WCDMA 64/64 using G.711 failed most of the time or resulted in very long delays of several seconds and are therefore excluded.

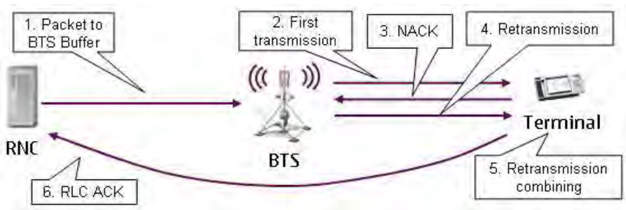


Figure 9. BTS retransmissions handling.

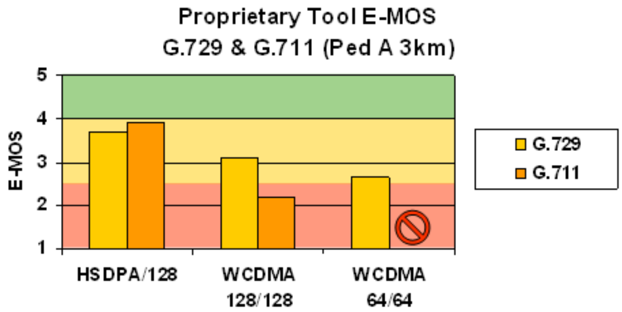


Figure 10. Codec performance evaluation (G.729 and G.711).

Table 2. G.711 codec jitter and packet loss (PedA 3km).

Access	Jitter Average	Packet Loss %
HSDPA/128	13ms	0.34
WCDMA 128/128	19ms	2.78

Figure 10 shows the VoIP quality comparison for both G.729 and G.711 codecs. Table 2 summarizes the jitter average and packet loss encountered when using the G.711 codec.

3.5. Embedded VoIP Client Evaluation

These tests aimed at determining the additional delay resulting from real embedded VoIP clients, such as the one included with the N95. The test setup consisted of establishing a VoIP call using an IMS system with the G.729 codec. Subsequently, we measured the offset delay, that is, the delay between the moment when the original audio sample occurs to the moment the audio sample is reproduced in the other calling end. The tool used for offset measurements was Malden DSLA [12]. Figure 11 shows the measurement environment. The results show that the total offset delay including the VoIP client processing delay is rather high (see Table 3). ITU-T recommends 400ms as the maximum delay for voice services with a reasonable quality. With delay above this limit, conversations are not interactive anymore and result in talker overlaps. Therefore, a voice service with very high delays results in a situation in which most, if not all users are dissatisfied. As a comparison, current circuit switched voice services have a delay of roughly 230–250ms.

With the results we can estimate the client processing delay by subtracting the average end-to-end delay from our tests based on the E-Model, 215ms, 295ms, and 350ms respectively. The result is roughly 210ms additional processing delay when using a real embedded VoIP client. This value differs considerably from the more optimistic processing delay estimations of 50-75ms available in research from [13,14].

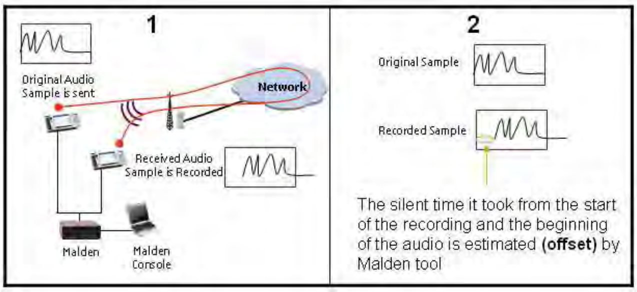


Figure 11. Offset delay measurement environment.

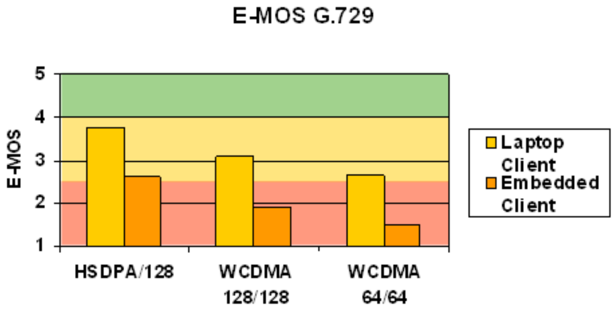


Figure 12. Overall VoIP quality with laptop and embedded handheld clients.

Table 3. Sources of delay (G.729 codec).

	HSDPA /128	WCDMA 128/128	WCDMA 64/64
RTT Delay	85ms	170ms	225ms
Jitter Buffer (100-200ms)	130ms	125ms	125ms
Total E2E Delay	215ms	295ms	350ms
Total E2E Delay, including embedded client delay	425ms	505ms	560ms

3.6. HSDPA/WCDMA Overall Effect on VoIP Performance

End-to-End delay is the main reason for low voice quality. With the total end-to-end total delay average values we can extend the analysis by dividing the sources of delay (Table 3).

With this estimation it is quite clear to understand why VoIP does not perform well in current systems with handheld terminals, and particularly live networks, even when the round trip time (RTT) is low. The final end-to-end delay is just too high. We finalize our VoIP quality analysis by modeling the resulting VoIP quality MOS with the additional embedded VoIP client processing delay based on the E-Model (see formula 2). Figure 12 shows this estimation. The results represent a case of a laptop client versus using an embedded client in a handheld device such as the N95 VoIP client. The figure considers both delay and packet loss impairment factors. It must be noted though, that in a laptop client there will also be an additional processing delay. However, such delay is considerably lower, ~50ms in a worst case scenario [15]. Thus, still ~160ms lower than with the mobile device tested.

Future features such as HSUPA in further 3GPP releases will slightly improve performance. For example, the expected average RTT for HSUPA networks is roughly 65ms (a reduction of 20ms compared with HSDPA). This reduction however does not improve the VoIP quality when using a laptop. That is, the average MOS with a laptop will still be the same. Contrastingly, the expected quality improvement for an embedded client is about 0.2 points in the MOS score. The main reason for the limited quality improvement is that the major sources of delay, and therefore, main impairment factors reducing VoIP quality are not directly related only to the wireless access, but to the VoIP client implementation. However, as we described previously, if some HSUPA features like UNACK mode are enabled in the wireless network, it will be possible to reduce the size of the jitter buffer implementation without compromising the VoIP quality. Furthermore, a reduction in the client processing delay is extremely important in order to seriously improve the VoIP quality in the mobile environment.

4. VoIP Signaling Analysis and Results

In this section we analyze the latencies for VoIP using the Session Initiation Protocol (SIP). This is an important metric because long delays in the call setup seriously harm the overall VoIP experience; people have certain expectations based on the current circuit switched services, and it is crucial to meet those.

4.1. SIP Registration Setup

The signaling [16] and delay measurements for SIP Registration to the VoIP server in the IMS system are depicted in Figure 13. The measurements show that the registration times with HSDPA and WCDMA are about 30% and 50% higher than with WiFi. While this might not seem much, we should remember that SIP registration requires a very limited number of messages. Therefore, as more messages are required, such as with 3GPP based registration, delays will increase.

4.2. VoIP Call Signaling

ITU E.721 [17] recommends values for call setup delays in circuit switched calls. The recommended values for call setup (post-dial delay) are 3s for local, 5s for toll and 8s for international connections, with 6s, 8s, and 11s as 95% values. The “call answer” (answer-signal) delay reflects

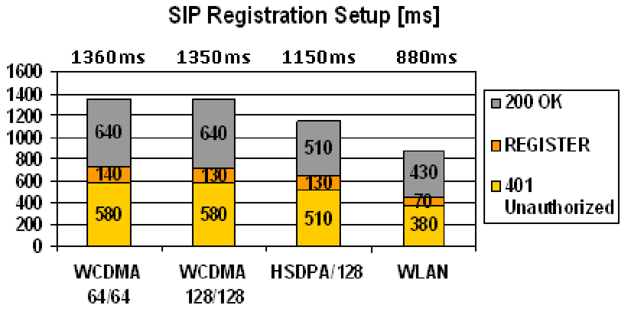


Figure 13. SIP registration signaling delays.

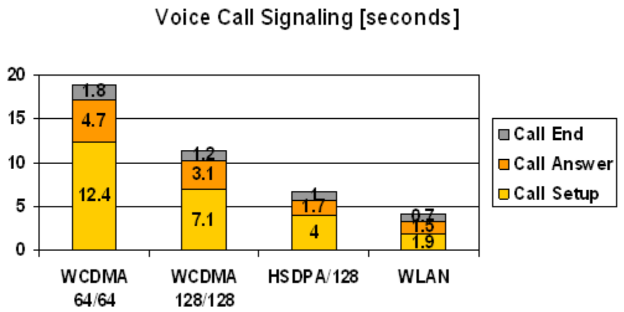


Figure 14. VoIP call signaling delays.

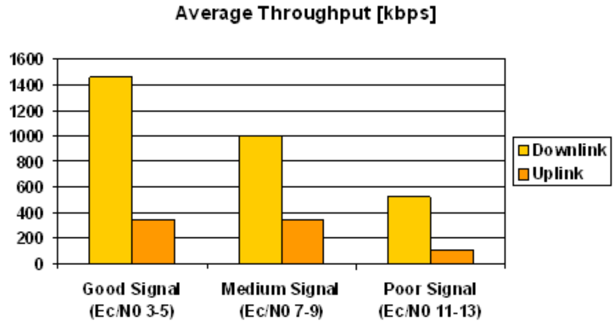


Figure 15. Average throughput in Elisa HSDPA network.

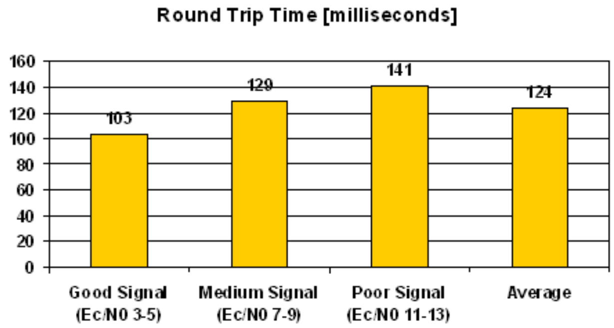


Figure 16. Average round trip time in Elisa HSDPA network.

the time it takes from the moment the receiving end accepts the call until the call is actually established. E.721 recommendation is 0.75s for local, 1.5s for toll, and 2.0s for international connections, with 1.5s, 3.0s, and 5.0s as 95% values. Finally, "call end" (call-release delay) is the time it takes for the call to be terminated [18,19]. The signaling [16] and delay measurements for voice call setups when a PDP context is active and the terminal is registered to the IMS system are depicted in Figure 14. The results show an expected increment in the call signaling delays depending on the access used. Since all of the network elements were located in a private network, the environment could be though of as providing local calls. Our results also show that an embedded mobile VoIP client experiences an increased delay compared to a PC client, such as the one measured by Curcio and Lunden [18] with a WCDMA network.

The setup delays for VoIP calls might be impacted with additional delays in a cellular system in cases were there is no active PDP context, and also due to a required registration to the IMS. The PDP context activation delay was ~3 seconds in our tests. Simulations by Pous et al. [20] propose 2.24 seconds. Based on these values, the always-on enabled calls can be in line with E.721 recommendations. However, when the PDP context is not active, the delay with WCDMA can vary between 11 to 17 seconds, and thus, exceed the recommended values. HSDPA delay in this case is around 8 seconds, which is similar to the recommendation for international calls. However, additional delays from e.g. traversed networks, gateways, and proxies could result in larger total delays than those recommended.

5. Live Network Case Study

5.1. Generic HSDPA Performance

In this section we describe the generic evaluation of the live HSDPA (Release 5 equipment) network performance in Helsinki. The results for throughput and round trip time are depicted in Figure 15 and Figure 16. The measurement results show an increase in round trip time delay when compared to the average values measured in the lab environment (85ms). This means that the VoIP quality (MOS) will be worse than our results in Section 3, and therefore VoIP support will be even more difficult. Throughput was measured via multiple file downloads and uploads from a local server in Finland; while RTT was measured with 32Byte ICMP Echo Request and Reply (ping) packets to the same server.

5.2. VoIP Quality

The VoIP quality in the Elisa HSDPA network is likewise slightly lower than our lab measurements (see Table 4). The mean opinion score was 3.5, 3.5 and 3.3 for good, medium and poor signal conditions. However, we have to consider that once again, the VoIP quality was measured for laptop based VoIP communication. That is, it does not account for the additional processing delay for the terminal VoIP client implementation previously described.

The results in the live case still indicate that VoIP support for a handheld embedded client will be poor. However, these values do take into account the jitter buffer play out delay. The most noticeable difference between the three scenarios is the packet loss ratio, which increases as the signal quality decreases.

5.3. VoIP Quality in Mobility Scenarios

The mobile environment tests were measured from a van driving through a test route at average speeds of 60-80 km/h without stopping. The selected test route crosses the Helsinki metropolitan area from East to West and it is entirely covered by Elisa HSDPA network according to their publicly available coverage map. The test route was about 18.5km and it took approximately 15min to travel. The test route was driven several times to validate the results.

The results show that the average performance is lower than in static scenarios. A mobile scenario obviously brings several additional challenges due to the different cell changes along the test route. The number of cell changes along the route was 28 and were characterized via the changes in scrambling codes used. Table 4 summarizes the VoIP quality results.

Furthermore, in mobility scenarios the amount of RLC retransmissions required is very noticeable. To characterize the retransmissions, we conducted an additional test along the test route in which we sent continuous ping packets of 32B (see Figure 17). The results show a large amount of delay peaks resulting from these retransmissions. Therefore, it further supports our lab measurements and emphasizes the importance of the unacknowledged mode feature. We expect that this mode would potentially take the majority of large delay peaks, and thus, improve VoIP quality. However, if this mode is used, there is a possibility that the packet loss ratio will increase, and for that reason, it is very important to validate future results as well even if the feature is enabled.

In addition, during the mobile tests, we recorded the signal conditions to characterize the signal quality distribution

Table 4. VoIP quality in Elisa HSDPA network (including jitter buffer).

Scenario	Delay Avg.	Jitter Avg.	Packet Loss %	MOS
Good Signal (Ec/N0 -3 to -5)	288ms	19ms	0.4	3.5
Medium Signal (Ec/N0 -7 to -9)	283ms	19ms	1.0	3.5
Poor Signal (Ec/N0 -11 to -13)	266ms	14ms	2.6	3.3
Mobile Environment	331ms	22ms	1.9	3.2

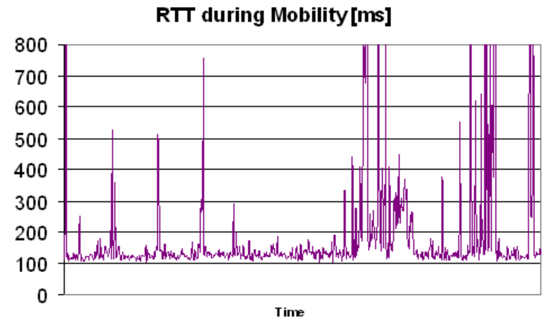


Figure 17. Round trip time during mobility tests.

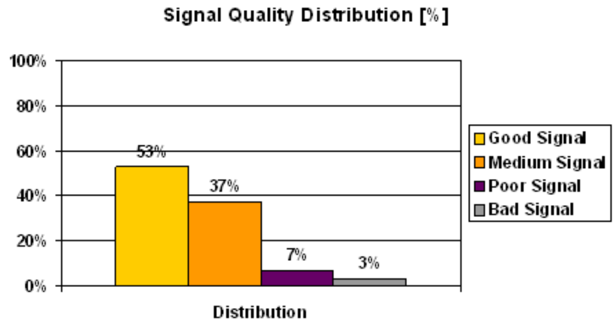


Figure 18. Signal quality distribution.

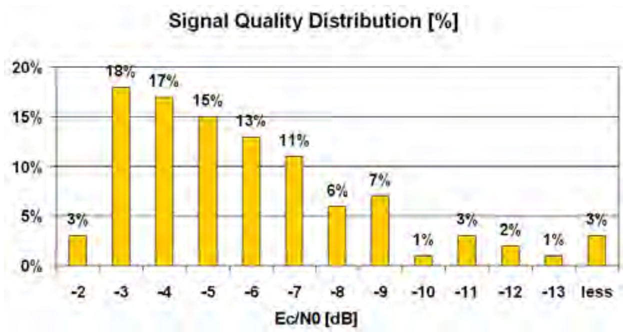


Figure 19. Detailed signal quality distribution.

along the test route. The measurements show that in general, it is highly probable to get a good signal level and that the coverage is well deployed (see Figure 18 and Figure 19).

6. Future Directions

At the current moment, the performance of VoIP in 3G networks is far from optimal. However, with some of the features and improvements in further 3GPP releases, the performance will improve. For instance, Release 6 equipment reduces RTT to roughly 65ms, and even lower with Release 7. Likewise, with Release 7 operators have other choices for deployment prior to full VoIP rollouts. For instance, advances such as Circuit Switched voice over HSPA (CS over HSPA) can improve capacity to similar levels as with VoIP. In this case, traditional voice is carried over packet data. Hence, since VoIP does not provide any significantly better capacity figures over CS over HSPA, operators can delay VoIP deployment can be delayed until adequately performing terminals and networks are available. This however, is only possible if several features are upgraded in several network elements. These improvements occurred while this manuscript was under review. CS over HSPA is expected to be included in 3GPP Release 7 [21].

7. Related Work

Although, there is prior work investigating the VoIP performance in WCDMA and HSDPA systems, it is not very extensive and mostly based on simulations. For instance, some papers [22,23,24] study VoIP performance in WCDMA and provide some baseline results. In addition, other works [25,26] provide some estimated values for processing delays. In these studies, the assumption for the estimations is based on whether the call is towards a landline or a mobile end. Some performance simulations are also available [8,10,13,14,27–29]. However, the simulations only provide a delay budget rather than a description of the end user experience. Contrastingly, our study focuses on end user experience and VoIP quality rather than delay budgets alone. The delay budget values used in simulations vary from 80-150ms for studies ignoring encoding/processing delays and jitter buffer implementations [9,27,28,30,31], to 250-300ms for studies that assume such delays to some extent [8,10, 13,14,29]. In addition, the estimations used in simulations are in general overly optimistic in regards to, e.g. client processing delay. Kim [14] considers the processing/ encoding delay to be 50ms, while Ericson [13] assumes around 75ms. These delay values include the jitter buffer playout delay as well. Therefore, it is noticeable they are too optimistic, especially when compared to our experiment results with actual handsets and VoIP jitter buffer client implementations.

Even though, it is understandable that the exact encoding/processing delays and jitter buffer playout delays are client specific, unless they are modeled accordingly, or at least, to some extent, the differences in performance between simulations and actual deployments will remain very visible. Therefore, simulations results are only comparable to laptop based performance at its best and not to actual handheld performance, which in the end is the primary use case for VoIP services. Other simulation study [28], notices the importance of reducing RLC retransmissions to improve performance in FTP and HTTP browsing. However, the study does not address its importance for VoIP services. Finally, Wager and Sandlund [32] conduct simulations to determine the amount of possible lost frames of VoIP speech in HSDPA mobility scenarios.

In regards to VoIP signaling, SIP call setup delays and signaling performance have been studied previously mostly for Internet scenarios. ITU E.721 [17] recommendation and an IETF Internet Draft [33], provide call setup delays recommendations for circuit switched and Internet Telephony systems respectively. Additionally, Eyers and Schulzrinne [19] provide guidelines for Internet Telephony call setup and signaling transfer delays. In regards to 3GPP based wireless accesses, Kist and Harris [34] provide simulations for transfer delays with 3GPP signaling, while Fathi et al. [35] and Pous et al. [20] modeled signaling performance. Further, Curcio and Lunden [9] provide measurements for a WCDMA setup using laptop clients for local, international and overseas calls. Most of the mentioned research focuses on simulations, and does not consider some end user cases such as calls in wireless environments starting from different states. Additionally, performance with different wireless radio accesses and configurations under the same conditions is not available. Also, the available works do not use an embedded VoIP client in a handheld mobile terminal, which yields different delay values compared to a PC. HSDPA signaling performance has not been evaluated either. Our research aims at covering these items. The importance of evaluating a mobile terminal relies in the fact that the eventual substitution of circuit switched calls in 3GPP networks (HSDPA and WCDMA) for VoIP calls will take place with a handheld mobile device and not with a PC or laptop. Likewise, multi-radio devices can provide ubiquitous access via different wireless access technologies with distinct performance characteristics.

The lack of actual measurement performance values in literature could be mainly due to the unavailability of integrated VoIP clients in the terminals and available HSDPA networks. However, with the introduction of some multi-radio devices with VoIP capabilities (e.g. Nokia N95, Nokia 6110), it is possible to use VoIP applications without a PC.

8. Conclusions

Multiple measurements were carried out to evaluate and characterize the VoIP quality and VoIP signaling performance in HSDPA and WCMDA wireless accesses. The results show that HSDPA access is capable of providing a competitive VoIP quality compared to circuit switched voice. However, WCDMA in 128/128 and 64/64 bitrate configurations can only provide low and poor qualities, the main issues are long delays and packet losses, which occur often due to RLC retransmissions that overflow the jitter buffer capacity. However, the main issue with HSDPA is not only tied directly to the wireless access performance, but to the mobile device capabilities. Our results show that embedded mobile VoIP clients can introduce an increased delay due to processing when compared to laptop performance. This processing includes e.g. encoding/decoding, and other operating system tasks. The additional delay has a considerable voice quality reduction effect. Further, the results from the test cases experimented in a live network resulted in lower performance when compared with similar laboratory measurements. Also, the effect of mobility in regards to VoIP quality degradation is quite noticeable. The degradation is due to handovers during the test route that increase the ratio of RLC retransmissions.

Therefore, the main aspects that can potentially improve VoIP quality performance with the current systems are mainly to reduce the number of RLC retransmissions by using unacknowledged mode, potentially use smaller jitter buffer sizes, and reduce the embedded VoIP client processing delays. High quality VoIP in 3G networks will be possible. However, it is tied to improvements in several areas such as wireless network delay, client implementation, and client processing delay. Finally, a main improvement developed while this manuscript was in process is CS over HSPA, which improves capacity and thus, can allow operators to delay VoIP deployment projects until networks and terminals have better performance.

9. References

[1] H. Holma, and A. Toskala, "HSDPA/HSUPA for UMTS," John Wiley, 2006.

[2] ITU-T, Recommendation G.107 "The E-model, a computational model for use in transmission planning," 2003.

[3] R. Cole and J. Rosenbluth, "Voice over IP performance monitoring," ACM SIGCOMM'01.

[4] A. Arjona, C. Westphal, A. Ylä-Jääski, and M. Kristensson, "Towards high quality VoIP in 3G networks: An empirical study," In Proceedings IEEE AICT'08, Athens Greece, June 8–13, 2008.

[5] IxChariot, http://www.ixiacom.com.

[6] A. Arjona and A. Ylä-Jääski, "VoIP call signaling performance and always-on battery consumption in HSDPA, WCDMA and WiFi," in Proceedings IEEE WiCOM'07, Shanghai China, September 21–23, 2007.

[7] Wireshark Protocol Analyzer, http://www.wireshark.org.

[8] B. Wang, K. Pedersen, T. Kolding, and P. Morgensen, "Performance of VoIP on HSDPA," IEEE Vehicular Technology Conference VTC'05 Spring, Vol. 4, pp. 2335–2339, May 30–June 1, 2005.

[9] P. Lundén, and M. Kuusela, "Enhancing Performance of VoIP over HSDPA", In Proceedings IEEE VTC'07 Spring, April 21-24, 2007.

[10] A. Braga, E. Rodriguez, and F. Cavalcanti, "Packet scheduling for VoIP over HSDPA in mixed traffic scenarios," in Proceedings 17th IEEE PIMRC'06, September 2006.

[11] H. Kim, "Loosing Opportunism: Evaluating Service Integration in an Opportunistic Wireless System," IEEE INFOCOM'07, May, 2007.

[12] Malden DSLA, http://www.malden.co.uk/dsla.

[13] M. Ericson and S. Wänstedt, "Mixed traffic HSDPA scheduling–impact on VoIP capacity," in Proceedings IEEE VTC'07 Spring, April 21–24, 2007.

[14] Y. Kim, "VoIP Service on HSDPA in Mixed Traffic Scenarios," in Proceedings IEEE CIT'06, September 2006.

[15] Cisco, "Understanding delay in packet networks," Document ID: 5125, March 2007.

[16] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session initiation protocol," RFC 3261, IETF June 2002.

[17] ITU, Recommendation E.721, "Network grade service parameters and target values for circuit switched services in the evolving ISDN," 1999.

[18] I. Curcio and M. Lundan, "SIP call setup delay in 3G networks," ISCC'02, Taormina Italy, July 1–4, 2002.

[19] T. Eyers and H. Schulzrinne, "Predicting Internet telephony call setup delay," IPTel2000, Berlin, April 2000.

[20] M. Pous, D. Pesch, G. Foster, and A. Sesmun, "Performance evaluation of a SIP based presence and instant messaging service," 3G 2003, June 25–27, 2003.

[21] Nokia Siemens Networks and Nokia, "Supporting CS over HSPA," 3GPP R2-073487, August 2007.

[22] F. Poppe, D. Vleeschauwer, and G. Petit, "Choosing the UMTS air interface parameters, the voice packet size and the de-jittering delay for a voice-over-IP call between a UMTS and a PSTN party," IEEE INFOCOM, Vol. 2, pp. 805–814, April 2001.

[23] F. Poppe, D. Vleeschauwer, and G. Petit, "Guaranteeing quality of service to packetized voice over the UMTS air interface," 8th International Workshop on Quality of Service, pp. 85–91, June 2000.

[24] R. Cuny and A. Lakaniemi, "VoIP in 3G networks: An end-to-end quality of service analysis," in Proceedings IEEE VTC'03 Spring, April 2003.

[25] ITU-T, Recommendation G.114, "One-way transmission time," 2003.

[26] TIATR-41.1.2, "VoIP end to end delay budget planning for private networks," Cisco 2000.

[27] G. Rittenhouse and H. Zheng, "Providing VoIP service in UMTS-HSDPA with frame aggregation," in Proceedings IEEE ICASSP'05, March 18–23, 2005.

[28] L. Bajzik, L. Korössy, K. Veijalainen, and C. Vulkán, "Cross-layer backpressure to improve HSDPA performance," in Proceedings IEEE PIMRC'06, Helsinki Finland, June 2006.

[29] P. Hosein, "Scheduling of VoIP traffic over a time-shared wireless packet data channel," in Proceedings IEEE ICPWC'05, January 23–25, 2005.

[30] Y. Seo and D. Sung, "Performance of VoIP in HSDPA based on an adaptive power allocation scheme," in Proceedings IEEE WCNC'06.

[31] P. Hosein, "Capacity of packetized voice services over time-shared wireless packet data channels," in Proceedings IEEE INFOCOM'05, March 13–17, 2005.

[32] S. Wager and K. Sandlund, "Performance evaluation of HSDPA mobility for voice over IP," in Proceedings IEEE

Vehicular Technology Conference VTC'07 Spring, April 22–25, 2007.

[33] H. Lin, T. Seth, A. Broscius and C. Huitema, "VoIP Signaling performance Requirements and Expectations," IETF Draft, June 1999.

[34] A. Kist and R. Harris, "SIP Signaling Delay in 3GPP", IFIP Interworking'02, October 13–16, 2002.

[35] H. Fahti, S. Chakraborty, and R. Prasad, "Optimization of SIP session setup delay for VoIP in 3G wireless networks," IEEE Trans. on Mobile Computing, Vol. 5, No. 9, September 2006.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Mathematical Formulations of Signal Propagation in Ultra-Wideband Transceiver Systems under a UWB Channel Environment with an Extension of Frequency Offset Correction

Debarati SEN
G. S. S. School of Telecommunications, Indian Institute of Technology Kharagpur, India
Email: debarati@gssst.iitkgp.ernet.in
Received April 4, 2008; revised August 29, 2008; accepted October 14, 2008

Abstract

This paper analyzes mathematically the crucial aspects of signal processing in a Multi-Band (MB) Orthogonal Frequency Division Multiplexing (OFDM) based system considering Ultra-Wideband (UWB) channel environment. In the process of analysis, it emphasizes the significant features of UWB receiver design in comparison with 'conventional' narrow-band system. The analysis shows that the high dispersive nature of a frequency selective UWB channel effects the design of different signal processing blocks like pre-select filter, low noise amplifier (LNA) and analog-to-digital (A/D) converter in the receiver front end. The characteristic functions of each of these stages are now dominated by the channel characteristics and it needs to be modified accordingly. This analysis is extended further with the study of frequency offset error and its correction. The unbiased Cramer Rao Lower Bound (CRLB) of estimation error is calculated and supported by computer simulation. The performance of an MB-OFDM system with frequency offset correction in terms of Bit-Error-Rate (BER) is also reported.

Keywords: Multi-Band OFDM, Signal Propagation, Transceiver, Ultra-Wideband

1. Introduction

The multi-band orthogonal frequency division multiplexing (MB-OFDM) based transmission scheme for ultra-wideband (UWB) has attracted keen interest of researchers in these days. MB-OFDM has made UWB commercially more viable than 'code-division multiple access'-based and 'impulse-radio-based' UWB systems for short range (＜20m) high bit rate (＞480Mbps) wireless applications.

However, it is essential to understand the crucial aspects of different signal processing stages in a UWB transceiver system that make a UWB receiver design significantly different compared to a narrow-band receiver design. Detailed mathematical studies on a UWB transceiver provides insights and helps in understanding the important aspects of a UWB receiver design which is the main aim of this work. Hence, there is a strong need for carrying on such an analysis for an MB-OFDM UWB system.

In this work, we first analyze an MB-OFDM UWB transmitter-receiver under a realistic UWB channel environment considering perfect timing and frequency offset estimation. Then, the study is extended to evaluate the performance of the system incorporating frequency offset errors through mathematical analysis supported by computer simulations.

The rest of this paper is organized as follows: Section 2 deals with the signal analysis in MB-OFDM transceiver. The ultra-wideband channel is modeled in Section 3. Signal analysis in an MB-OFDM receiver is presented in Section 4. Performance of the system with frequency offset error correction is analyzed in Section 5. It also shows the close match between calculated Cramer Rao Lower Bound (CRLB) of estimated error with simulation result. Section 6 figures out simulation results with some relevant discussion. Important concluding remarks with the summary of the paper are included in Section 7.

2. Signal Analysis in an MB-OFDM Transmitter

2.1. MB-OFDM System Specification

The Federal Communications Commission (FCC) has allotted 7.5GHz wide bandwidth for UWB transmission. The MB-OFDM system divides this unlicensed spectrum into fourteen (14) bands [1]. OFDM symbols are transmitted by frequency hopping over the bands. The block diagrams of a UWB transmitter and receiver considered for this study are shown in Figure 1 and Figure 2.

Let $\{X(k)\}$ be the QPSK modulated complex sequence of symbols to be transmitted using OFDM (= T_6 signal point in transmitter in Figure 1). A complex base band OFDM symbol can be obtained using N point Inverse Fast Fourier Transformation (IFFT) represented as

$$x(t) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j2\pi fkt}; \quad 0 \leqslant t < T$$
$$= 0 \text{ elsewhere} \tag{1}$$

where, f = sub-carrier spacing. Taking samples at $t = nT_S$, Equation (1) can be written as

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j(2\pi/N)nk} \tag{2}$$

where, N = number of samples over duration T. Correspondingly, the *n-th* sample of the *m-th* OFDM symbol can be shown as

$$x(n,m) = \frac{1}{N}\sum_{k=0}^{N-1} X(k,m)e^{j2\frac{\pi k}{N}(n-mN)}; \; 0 \leqslant m \leqslant Z-1 \tag{3}$$

where, 'Z' is the number of OFDM symbols transmitted.

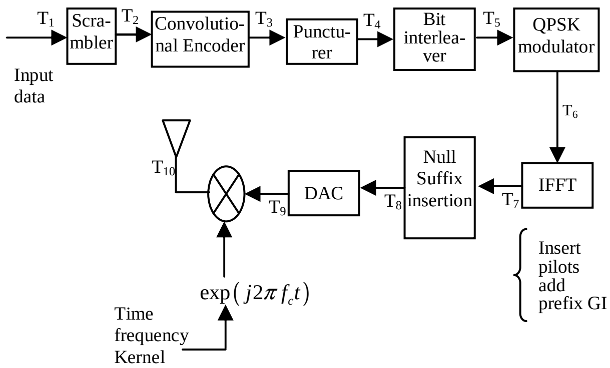


Figure 1. Multiband OFDM UWB transmitter.

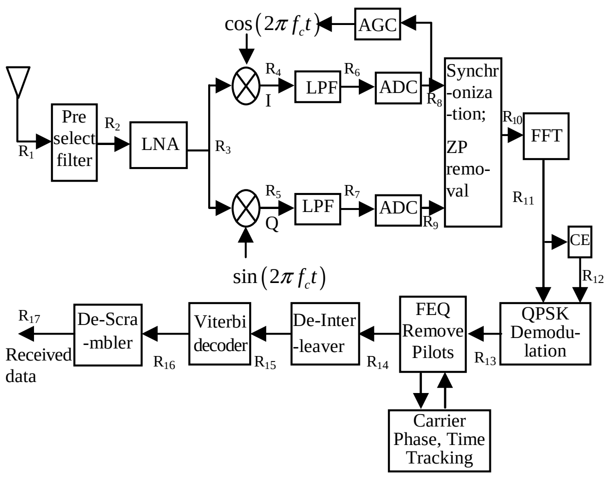


Figure 2. Multiband OFDM UWB receiver.

2.2. Cyclic Prefix Insertion

In MB-OFDM transmission, zero-padded bits are added in place of cyclic prefix to compensate the multipath effect. Let, N_Z number of Zero padded bits (ZP) and N_g number of guard bits (for inter band synchronization in MB-OFDM) are added with 'N' point IFFT output to construct one MB-OFDM symbol. The total number of samples 'N_s' in one OFDM symbol is given as, $N_s = N + N_Z + N_g$. Hence, from Equation (2),

$$x'(n) = 0, \quad \text{for } 0 \leqslant n \leqslant N_z-1$$
$$= x(n-N_z), \quad \text{for } N_z \leqslant n \leqslant N+N_z-1$$
$$= 0, \quad \text{for } N+N_z \leqslant n \leqslant N+N_z+N_g-1 \tag{4}$$

Similarly from Equation (3) (at T_8 in Figure 1),

$$x'(n,m) = \frac{1}{N}\sum_{k=0}^{N-1} X(k,m)e^{j2\frac{\pi k}{N}(n-N_Z-mN_S)}; \begin{cases} 0 \le m \le Z-1 \\ 0 \le n \le N_s \end{cases} \tag{5}$$

2.3. Signal Passing through DAC

After IFFT and zero padding, the signal (5) is passed next through a Digital to Analog Converter (DAC) block. The DAC operation involves two steps: firstly, conversion of the samples $x'(n,m)$ into a sequence of impulses given as

$$x_s(t) = \sum_{n=-\alpha}^{\alpha} x'(n,m)\delta(t-nT_s) \tag{6}$$

Then it is filtered by a reconstruction filter, which is an ideal Low Pass Filter (LPF) having impulse response given as

$$h_r(t) = \sin(\pi t/T_s)/\,\pi t/T_s \tag{7}$$

Hence, the signal at the output of DAC (at T_9 in Figure 1) is

$$T_9 = x'(t) = \sum_{n=-\alpha}^{\alpha} x'(n,m)\frac{\sin\pi(t-nT_s)/T_s}{\pi(t-nT_s)/T_s} \tag{8}$$

The above conversion is performed when sampling of the signal follows the Nyquist sampling theorem.

2.4. Signal Upconversion

Now, the analog real valued signal $x'(t)$ is multiplied by time frequency Kernel $\exp(j2\pi f_c t)$ to convert it to RF signal from base band. Therefore, the transmitted signal (at T_{10} in Figure 1) may be given as

$$T_{10} = x_{tx}(t) = x'(t)\exp(j2\pi f_c t)$$
$$= x'(t)\cos(2\pi f_c t) + jx'(t)\sin(2\pi f_c t) \tag{9}$$

3. The Ultra-Wideband Channel Model

The proposed multipath channel model for UWB [2] by IEEE 802.15.3 channel modeling sub-committee is a modified form of indoor channel model proposed by A. Saleh and R. Valenzuela (S-V) [3] that models multipath arrivals in the form of clusters as well as rays within clusters. S-V model clearly distinguishes between 'cluster arrival rate' and 'ray arrival rate'. The cluster arrival rate 'Λ' and ray arrival rate 'λ' within a cluster is given by Poisson process ($\Lambda \ll \lambda$).

Mathematically, the 'impulse response' of a channel is given as

$$h(t) = \sum_{l=0}^{L-1}\sum_{i=0}^{I-1} \alpha_{i,l}\,\delta(t - T_l - \tau_{i,l}) \qquad \textbf{(10)}$$

where, $\alpha_{i,l}$ = channel coefficient for the *i-th* ray of the *l-th* cluster which is log-normally distributed; T_l = delay of the *l-th* cluster; $\tau_{i,l}$ = delay of the *i-th* ray related to the *l-th* cluster arrival time; $\tau_{0,l}$ = 0, by definition.

The distribution of cluster arrival time and ray arrival time are given as

$$p(T_l/T_{l-1}) = \Lambda \exp[-\Lambda(T_l - T_{l-1})], \quad l > 0 \qquad \textbf{(11)}$$

$$p(T_l/T_{l-1}) = \Lambda \exp[-\Lambda(T_l - T_{l-1})], \quad l > 0 \qquad \textbf{(12)}$$

Channel coefficients are defined as

$$\alpha_{i,l} = p_{i,l}\,\xi_l\,\beta_{i,l} \qquad \textbf{(13)}$$

where, ξ_l = fading associated with cluster; $\beta_{i,l}$ = fading associated with *i-th* ray of *l-th* cluster; $p_{i,l}$ = equiprobable +1/-1 to account for signal inversion due to reflection.

The large bandwidth (7.5GHz) of UWB significantly increases the resolving capability of a UWB receiver. Here, the number of reflections from channel arriving within a short period (0.167nsec. for 6GHz bandwidth and 0.133nsec. for 7.5GHz bandwidth) of impulse is too small. As a result, the central limit theorem is not valid further in UWB systems. Behavior of (averaged) power delay profile is an exponential function of time given as

$$E\left[\left|\xi_l \beta_{i,l}\right|^2\right] = \Omega_0\, e^{\frac{-T_l}{\Gamma}}\, e^{\frac{-\tau_{i,l}}{\gamma}} \qquad \textbf{(14)}$$

where, Γ = cluster decay factor, γ = ray decay factor, Ω_0 = mean energy of the first ray of the first cluster.

However, we shall continue our analysis considering the S-V channel model given by (10) in the next Section.

4. Signal Analysis in an MB-OFDM Receiver

The transmitted signal (9) after passing through a UWB channel is received at the receiver front end given by (at R_1 in Figure 2)

$$R_1(t) = \sum_{l=0}^{L-1}\sum_{i=0}^{I-1} \alpha_{i,l}\, x_{tx}\left[t - T_l - \tau_{i,l}\right] + w(t) \qquad \textbf{(15)}$$

where, $w(t)$ = Additive, white, Gaussian noise.

4.1. Preselect Filter

Preselect filter is the very first block of the receiver front end. Here, the out of band signals are attenuated and signals of desired bands are selected. Functionally, this is a wideband band pass filter (BPF). In practice, transfer function of narrow-band band pass filter is given by

$$h(t) = \mathrm{Re}[\tilde{h}(t)\exp(j2\pi f_c t)] \qquad \textbf{(16)}$$

where, $\tilde{h}(t)$ is complex impulse response of band pass system given by

$$\tilde{h}(t) = h_I(t) + jh_Q(t) \qquad \textbf{(17)}$$

wherein, h_I & h_Q are equivalent LPF coefficients.

The above transfer function of narrow-band BPF is not valid here for UWB. The receiver front end signal passes through BPF with center frequency f_c and bandwidth B, which is wider than or equal to the 10dB B.W. of the transmitted signal. For convenience of analysis, we assume that the BPF passes the received desired signal part perfectly without distortion and the filter bandwidth B is equal to an integer multiple of $1/2T_b$ (T_b = bit duration) such that negligible inter symbol interference and intra symbol interference results. The filtered signal waveform can be expressed (at R2 in Figure 2) as

$$R_2(t) = \sum_{l=0}^{L-1}\sum_{i=0}^{I-1} \alpha_{i,l}\, x_{tx}\left[t - T_l - \tau_{i,l}\right] + \hat{w}(t) \qquad \textbf{(18)}$$

where $\hat{w}(t)$ is filtered noise given by [4]

$$\hat{w}(t) = \sqrt{2}\left[w_c(t)\cos(2\pi f_c t) - w_s(t)\sin(2\pi f_c t)\right] \qquad \textbf{(19)}$$

wherein, $w_c(t)$ and $w_s(t)$ are uncorrelated Gaussian random process with the autocorrelation function at the output of BPF given by

$$R_r(\tau) = E\{\hat{w}(t)\hat{w}(t+\tau)\} = N_0 B \frac{\sin(2\pi B\tau)}{2\pi B\tau}\cos(2\pi f_C \tau) \qquad \textbf{(20)}$$

Power spectral density of each noise component $w_c(t)$ and $w_s(t)$ at the BPF output is given by

$$S_{W_C}(t) = S_{W_S}(t) = \begin{cases} N_0 & |f| < B \\ 0 & otherwise \end{cases} \qquad \textbf{(21)}$$

4.2. Signal Passed through the LNA and Downconverter

The filtered signal is amplified in next step by passing through wideband low noise amplifier with a 3dB flat gain of β. The amplified signal may be given as (at R_3 in Figure 2)

$$R_3(t) = \beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1} \alpha_{i,l}\, x'(t - T_l - \tau_{i,l})\, e^{j2\pi f_C(t - T_l - \tau_{i,l})} + \beta\hat{w}(t) \qquad \textbf{(22)}$$

The amplified signal can also be written as

$$R_3(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\left\{x'(t-T_l-\tau_{i,l})e^{j(w_C t-\varphi_{i,l})}\right\}+\beta\hat{w}(t) \tag{23}$$

where, $\varphi_{i,l} \triangleq w_c\tau'_{i,l}$ and $\tau'_{i,l}$ = delay of *i-th* ray of *l-th* cluster measured with respect to the 1st ray of 1st cluster.

Here, $\varphi_{i,l}$ are essentially independent and uniformly distributed over (0, 2π). In MB-OFDM, the OFDM symbol period is $=1/\Delta f$ =242.42nsec. and sample period is 1.89nsec. The maximum delay spread obtained from measurement and supported by simulation for the worst channel here is 25nsec. (CM4). It is to be noted that the phase component $\varphi \triangleq w_c\tau$ can not be neglected here like narrow-band system as the maximum delay in S-V channel model (25nsec.) is comparable to $2\pi/w_c$. For example, $2\pi/w_c$= 0.3nsec. (for band 1, with f_c=3432MHz) and is 0.09nsec. (for band 3, with f_c =3960MHz).

Expanding the above Equation (23) we get,

$$R_3(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\left\{x'(t-T_l-\tau_{i,l})\cos w_c t+jx'(t)\sin w_c t\right\} e^{-j\varphi_{i,l}}+\beta\hat{w}(t) \tag{24}$$

It is obvious that the phases $\varphi_{i,l}$ are independent (each is a function of its respective path). Assuming that the receiver generates the coherent reference (perfect carrier phase estimate) we get in-phase and q-phase components can be expressed as (at R_4 and R_5 in Figure 2)

$$R_4(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\left\{x'(t-T-\tau_{i,l})\cos^2 w_c t+jx'(t-T_l-\tau_{i,l})\sin 2w_c t\right\} e^{-j\varphi_{i,l}}+\beta\hat{w}(t-T_l-\tau_{i,l})\cos w_c t \tag{25}$$

$$R_5(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\left\{x'(t-T-\tau_{i,l})\cos 2w_c t+jx'(t-T_l-\tau_{i,l})\sin^2 w_c t\right\} e^{-j\varphi_{i,l}}+\beta\hat{w}(t-T_l-\tau_{i,l})\sin w_c t \tag{26}$$

4.3. I-Q Signals at the Low Pass Filter (LPF) Outputs

Both the above down-converted signals after passing through LPFs (which are basically Nyquist filters), give the base band signals expressed as (at R_6 and R_7 in Figure 2)

$$R_6(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}x'(t-T_l-\tau_{i,l})e^{-j\varphi_{i,l}}+\hat{w}_1(t) \tag{27}$$

$$R_7(t)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}x'(t-T_l-\tau_{i,l})e^{-j\varphi_{i,l}}+\hat{w}_2(t) \tag{28}$$

where, $\hat{w}_1(t)=\beta\hat{w}(t-T_l-\tau_{i,l})\cos w_c t$ and $\hat{w}_2(t)=\beta\hat{w}(t-T_l-\tau_{i,l})\sin w_c t$ are Nyquist filtered noise.

4.4. Samples at A/D Converter Output without Frequency Offset

Both the above signals are converted in digital domain taking samples at $t = nT_s$ given as (at R_8 and R_9 in Figure 2)

$$R_8(n)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}x(n-l-i-N_Z)e^{-j\varphi_{i,l}}+\hat{w}_1(n) \tag{29}$$

$$R_9(n)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}x(n-l-i-N_Z)e^{-j\varphi_{i,l}}+\hat{w}_2(n) \tag{30}$$

The above signal samples obtained at the A/D converter output are basically the same samples fed at the DAC input in the transmitter block.

4.5. ZP Removal and FFT Implementation

In the next step, the ZP is removed by overlapping and add method assuming that the transmitted and received signals are perfectly matched. So, from (29) after ZP removal we get

$$R_{10}(n)=\beta\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\frac{1}{N}\sum_{k=0}^{N-1}X(k)e^{j2\pi k/N(n-l-i)}e^{-j\varphi_{i,l}} +\hat{w}_1(n+N_z) \tag{31}$$

Next, the *N*-point FFT is performed to recollect the frequency domain transmitted samples. Now, using $\beta\alpha_{i,l}=\alpha'_{i,l}$,

$$R_{11}(q)=\underbrace{x[q]}_{Signal}\underbrace{\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha'_{i,l}\,e^{\frac{-j2\pi lq}{N}}e^{\frac{-j2\pi iq}{N}}e^{-j\varphi_{i,l}}}_{Ch.\,Coeff} +\underbrace{\sum_{n=0}^{N-1}\hat{w}_1[n+N_z]e^{\frac{-j2\pi nq}{N}}}_{noise} \tag{32}$$

$$\text{where,}\quad x(q)=\sum_{n=0}^{N-1}x(n-l-i)e^{-j2\pi(n-l-i)q/N} \tag{33}$$

It is seen from the above equation that the frequency domain recovered samples are associated with phase errors due to the multipath delay associated with corresponding paths of ray as well as clusters.

4.6. Channel Estimation and Equalization

MB-OFDM channel is estimated from the channel estimate symbols of the preamble. Assuming that the channel is slow fading, its coefficients are considered to be constant during one OFDM symbol. The channel frequency response estimated over one OFDM symbol by the least square estimator is given by

$$\hat{H}(q)=\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha'_{i,l}\,e^{\frac{-j2\pi lq}{N}}e^{\frac{-j2\pi iq}{N}}e^{-j\varphi_{i,l}}+\frac{\sum_{n=0}^{N-1}\hat{w}_1(n)e^{-\frac{j2\pi nq}{N}}}{x(q)} \tag{34}$$

In MB-OFDM, 6 symbols are sent for channel estimation (CE) purpose. Hence final estimated channel frequency response can be obtained by taking the average of estimated channel coefficients over CE symbols available per band (C_b).

$$\hat{H}(final)(q) = \frac{1}{C_b}\sum_{r=0}^{C_b}\hat{H}_r(q) \quad \textbf{(35)}$$

In the above explanation it is clearly vivid that the error in channel estimates i.e. the distortion coming due the addition of noise in path (second part of Equation (34)) and the comparable multi path delay with respect to $2\pi/w_c$ reflected as $e^{-j2\pi(l+i)q/N}$ in the first term of Equation (34). Equalization of data OFDM symbols is done by using one tap equalizer, and hence estimated received samples are given by

$$x_{est}(q) = \frac{x(q)\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha'_{i,l}e^{-j2\pi lq/N}e^{-j2\pi iq/N}e^{-j\varphi_{i,l}} + \sum_{n=0}^{N-1}\hat{w}_1(n)e^{-\frac{j2\pi nq}{N}}}{\hat{H}_{(final)}(q)} \quad \textbf{(36)}$$

where, q = 0, 1, 2, ……………. 128.

4.7. QPSK Demodulation and Detection

The estimated samples $x_{est}(q)$ are fed to QPSK demodulator. At the output of the demodulator the transmitted symbol is recovered back taking decision over the signed amplitude of the $x_{est}(q)$. The QPSK demodulator input symbol $x_{est}(q)$ can be represented as

$$x_{est}(q) = \pm x_{est(inphase)} \pm x_{est(qphase)} \quad \textbf{(37)}$$

5. Signal Analysis in Receiver with Carrier Frequency Offset Correction and CRLB Calculation

Let the normalized carrier frequency offset is expressed in terms of sub carrier spacing f as $f' = \Delta f / f$, where, Δf is the carrier frequency offset. After removal of ZP (32 samples) the signal obtained at the receiver with frequency offset is given by

$$R_{10}(n) = \beta e^{j2\pi f'(n+N_Z)/N} \sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha_{i,l}\frac{1}{N}\sum_{k=0}^{N-1}X(k)e^{j2\pi k(n-l-i)/N}e^{-j\varphi_{i,l}} + \hat{w}_1(n+N_z) \quad \textbf{(38)}$$

After FFT, the signal is presented as

$$R_{11}(q) = e^{j2\pi f N_z/N}x\left((q-f')/N\right)\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\alpha'_{i,l}e^{-j2\pi(l+i)(q-f')/N} e^{-j\varphi_{i,l}} + \sum_{n=0}^{N-1}\hat{w}_1(n+N_z)e^{-j2\pi nq/N} \quad \textbf{(39)}$$

It is observable from Equation (39) that that the performance fall gradually increases with FO (Δf_b) when $|\Delta f_b| < f/2$ while it results in a complete detection error when $|\Delta f_b| > f/2$. In Equation (38) it is observed that f' causes a phase rotation of $2\pi f'(n+N_z)/N$. If it remains uncorrected, it causes both rotation of QPSK constellation points and a spread of the constellation points.

Frequency offset synchronization may be performed in receiver in time domain before performing FFT of the incoming signal either by cyclic extension or by using special training symbols. Synchronization technique based on cyclic extension is not suitable for high rate packet transmission (like MB-OFDM) because - (i) An accurate synchronization needs an averaging over large (>10) nos. of OFDM symbols to attain distinct number of peaks and a reasonable SNR. (ii) For efficient data transmission, synchronization time needs to be as short as possible. In Training symbol based synchronization technique, the incoming signal is correlated with the complex conjugate of the known training signal. From the correlation peaks in the matched filter output signal, symbol frequency offset can be estimated. We continue our analysis with a time domain frequency offset estimation scheme proposed in [5]. Now, considering

$$x_r(n) = \sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\beta\alpha_{i,l}\frac{1}{N}\sum_{k=0}^{N-1}X(k)e^{j2\pi k(n-l-i)/N} \quad \textbf{(40)}$$

The received signal $R_{10}(n)$ for the *n-th* sample of the *m-th* OFDM symbol ($0 \leqslant m \leqslant Z-1$) can be expressed as

$$R_{10}(n,m) = x_r(n,m)e^{j2\pi f'(n+N_Z+mN_S)/N}e^{\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}-j\omega_c\tau'_{i,l}} + \hat{w}_1(n+N_z+mN_s) \quad \textbf{(41)}$$

Now, considering $\sum_{l=0}^{L-1}\sum_{i=0}^{I-1}\tau'_{i,l} = \tau,$

$$R_{10}(n,m) = x_r(n,m)e^{j2\pi\left\{f'(n+N_z+mN_S)/N - f\tau\right\}} + \hat{w}_1(n,m) \quad \textbf{(42)}$$

where, $\hat{w}_1(n,m)$ is the AWGN added with the *n-th* sample of *m-th* OFDM symbol. Based on (42), the *n-th* sample of *m+1-th* OFDM symbol can be given as

$$\begin{aligned} R_{10}(n,m+1) &= x_r(n,m+1)e^{j2\pi\left\{f'(n+N_Z+(m+1)N_S)/N - f\tau\right\}} + \hat{w}_1(n,m+1) \\ &= R_{10(n,m)}e^{j2\pi f N_S/N}\underbrace{-\hat{w}_1(n,m)e^{j2\pi f N_S/N} + \hat{w}_1(n,m+1)}_{\tilde{w}(n,m)} \end{aligned} \quad \textbf{(43)}$$

The phase difference between *n-th* sub-carriers of 2 consecutive OFDM symbols *m* & (*m*+1) is $2\pi\{(f'N_S)/N\}$. Phase offset between 2 consecutive OFDM symbol is obtained by complex conjugate multiplication of *n-th* samples of *m-th* OFDM symbol & (*m*+1)-*th* OFDM symbol. So,

$$Z(n,m) = x_r^*(n,m)x_r(n,(m+1))$$
$$e^{j2\pi\{f'(n+N_Z+mN_S+N_S)/N-f\tau-f'(n+N_Z+mN_S)/N+f\tau\}} + \overline{w}((n,m) \quad \textbf{(44)}$$

where,

$$\overline{w}(n,m) =$$
$$x_r(n,m)e^{-j2\pi\{f'(n+N_Z+mN_S)/N-f\tau\}}\hat{w}_1(n,(m+1)) + \hat{w}_1^*(n,m)$$
$$x_r(n,(m+1))e^{j2\pi\left\{f'\left(n+N_Z+(m+1)N_S\right)/N-f\tau\right\}} + \hat{w}_1^*(n,m)\,\hat{w}_1(n,(m+1)) \quad \textbf{(45)}$$

For preamble, same OFDM symbol is repeated 21 times to constitute the training sequence utilized for synchronization. The frame format is shown in Figure 3. Hence typically,

$$x_r(n,m)x_r(n,(m+1)) = |x_r(n,m)|^2 \quad \textbf{(46)}$$

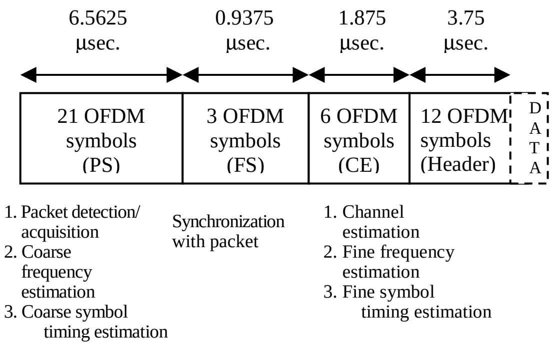


Figure 3. MB-OFDM frame format.

Estimated frequency offset from P number of sample pairs of *m-th* & $(m+1)$ - *th* OFDM symbols,

$$\hat{f}'(m) = f' + \frac{\hat{\varphi}_{noise}(m)}{2\pi N_S / N} \quad \textbf{(47)}$$

where,

$$\hat{\varphi}_{noise}(m) = \arg\sum_{n=0}^{P-1}\overline{w}(n,m) \quad \textbf{(48)}$$

The *n-th* noise sample of *m-th* OFDM symbol is Gaussian distributed with zero mean. The variance of estimated frequency offset from two successively received OFDM symbols can be easily given as

$$Var(\hat{f}'(m)) = \frac{N^2}{(2\pi)^2 N_S^2 P(SNR)} \quad \textbf{(49)}$$

Now, If we continue this offset estimation over 'L' OFDM symbols, comparison will be caused for (L – 1) times.

$$\hat{f}'(L) = f' + \frac{\frac{1}{L-1}\left(\sum_{m=1}^{L-1}\hat{\varphi}_{noise}(m)\right)}{2\pi N_s / N} \quad \textbf{(50)}$$

Correspondingly, the variance calculated,

$$Var\left(\hat{f}'(L)\right) = \frac{N^2}{(2\pi)^2(L-1)N_S^2P(SNR)} \quad \textbf{(51)}$$

5.1. CRLB for Estimated Frequency Offset

The expression for *n-th* pair of *m-th* and $(m+1)$-*th* OFDM symbols can also be written from Equation (43) as

$$R_{10}(n,m+1) = A\cos(2\pi f'N_S / N + \Delta)$$
$$+ jA\sin(2\pi f'N_S / N + \Delta) + \tilde{w}_{in}(n,m) + j\tilde{w}_{qp}(n,m) \quad \textbf{(52)}$$

where, $A = |R_{10}(n,m)|$, $\Delta = \angle R_{10}(n,m)$, $\tilde{w}_{in}(.)$ and $\tilde{w}_{qp}(.)$ are the in-phase and q-phase components of the effective noise of the complex conjugate product $\tilde{w}(n,m)$ respectively obtained from (43).

Now, let us take

$$\left.\begin{aligned} U(n,m) &= A\cos(2\pi f'N_S/N + \Delta) + \tilde{w}_{in}(n,m) \\ V(n,m) &= A\sin(2\pi f'N_S/N + \Delta) + \tilde{w}_{qp}(n,m) \end{aligned}\right\} \quad \textbf{(53)}$$

and

$$\left.\begin{aligned} u(n,m) &= A\cos(2\pi f'N_S/N + \Delta) \\ v(n,m) &= A\sin(2\pi f'N_S/N + \Delta) \end{aligned}\right\} \quad \textbf{(54)}$$

Expressing $R_{10}(n,m+1) = \xi(n,m+1)$, the probability density function of the sample vector $\xi(n,m+1)$ (when unknown parameter vector is θ) is given by

$$f(\xi(n,m+1);\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^P \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{p-1}\{U(n,m)-u(n,m)\}^2\right)$$
$$\times\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^P \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{p-1}\{V(n,m)-v(n,m)\}^2\right) \quad \textbf{(55)}$$

Inverse of Fisher Information Matrix gives the variance of unknown parameter estimates. Unbiased Cramer Rao Lower Bound (CRLB) of estimations obtained from diagonal elements of Inverse Fisher Information Matrix $I(\theta)$ [6] is as

$$I(\theta) = \begin{bmatrix} -E\left[\frac{\partial^2\ln(f(\xi(n,m+1);\theta))}{\partial A^2}\right] & -E\left[\frac{\partial^2\ln(f(\xi(n,m+1);\theta))}{\partial A\partial f'}\right] \\ -E\left[\frac{\partial^2\ln(f(\xi(n,m+1);\theta))}{\partial f\partial A}\right] & -E\left[\frac{\partial^2\ln(f(\xi(n,m+1);\theta))}{\partial f'^2}\right] \end{bmatrix} \quad \textbf{(56)}$$

when frequency offset estimation is carried over 'L' OFDM symbols, the variance is

$$Var(\hat{f}'(L)) = \frac{N^2}{(2\pi N_S)^2 P(L-1)SNR} \quad \textbf{(57)}$$

where SNR = signal to noise ratio.

6. Simulation Results and Discussions

Simulation is carried out to study the performance of MB-OFDM systems with frequency offset correction using Time Frequency Interleaved (TFI) pattern 1 [1] under UWB channel model CM4 and AWGN. Relevant parameters from ECMA-368 standard for MB-OFDM are considered for the simulation study. We have considered 1000 noisy realizations in each of the 100 UWB channels for our simulation. We estimated the channel using Least Square (LS) estimate during channel estimation sequence (25*th*-30*th* OFDM symbol of the frame format) of the preamble.

Coarse frequency offset of the frequency offset estimation scheme [5] is estimated for an AWGN channel using several consecutive OFDM symbols (L = 2, 3 4, 5, and 6) in the preamble for MB-OFDM systems. Figure 4 shows change of variance of the frequency offset estimation error of our scheme vs. SNR for various values of L, viz. 2, and 6. As expected, the variance of frequency offset estimation error decreases with increase in L. However, a lower value of L is desirable in practice to reduce the system computational complexity. Figure 4 also includes the error variance of frequency offset estimate and the CRLB of the error variance for our scheme for L=2 and 6. It is observed that the error variance as per our scheme is very close to the respective CRLB.

Figure 5 presents the BER vs. E_b/N_0 plots for our scheme with L = 2, 3, and 6 in UWB channel CM4 for normalized frequency offset of 0.05. It is observed that, for CM4, E_b/N_0 improves by 8.3dB and 3.6dB with L = 6 compared to L = 2 and L = 3 respectively for BER=10^{-4} [5].

7. Conclusions

To make UWB receiver design techno-economically challenging, understanding of the key aspects of different signal processing stages in an UWB transceiver system is very essential.

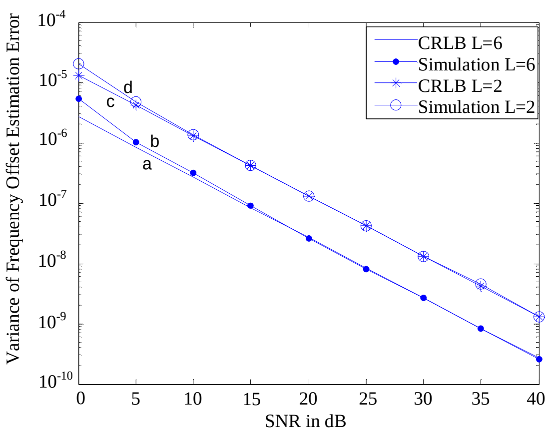


Figure 4. Variance of frequency offset estimation error: a) CRLB calculated with L=6; b) simulation with L=6; c) CRLB calculated with L=2; d) simulation with L=2 in AWGN channel.

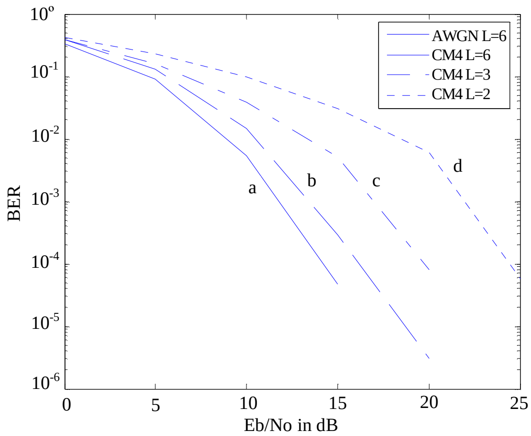


Figure 5. BER vs. Eb/No for: a) AWGN with L=6; b) CM4 with L=6; c) CM4 with L=3 and d) CM4 with L=2.

Considering this, in this work, we have mathematically analyzed the propagation of signal in an MB-OFDM based UWB transceiver system under realistic channel environment. This mathematical analysis is further extended to investigate the effects of synchronization imperfections due to carrier frequency offset in the receiver. The estimation of this frequency offset is carried out following our earlier proposed algorithm [5]. The Cramer Rao Lower Bound of the variance of estimation error is calculated and compared with the computer simulation. Closeness of both the curves as depicted by the graphical representation proves the effectiveness of our estimation algorithm for UWB receivers. A considerable improvement in system performance with higher number of iterations of the estimation algorithm is also reported. However, a higher number of iteration increases the receiver complexity. In practice, a compromise between performance improvement and receiver complexity is essential for effective receiver design.

The significance of this work is to show the fundamental differences of UWB signal processing in comparison with narrow-band system design which is being reported for the first time in literature here to the best of our knowledge. We further believe that, the detailing of the crucial aspects of UWB signal propagation discussed here will help the researchers to develop a clear understanding of this promising technology.

The frequency offset estimation algorithm discussed in this paper can be used in developing the practical UWB receiver. This mathematical frame-work can further be extended for timing imperfection scenario.

8. References

[1] ECMA-368, “High rate ultra wideband PHY and MAC standard,” 2nd Edition, December 2007, available at: http://www.ecma-international.org/publications/standards/Ecma-368.htm

[2] J. Foerster, “Channel modeling sub-committee report final,” IEEE, Document IEEE P802.15-02/490r1-SG3a, February 2003.

[3] A. A. M. Saleh and R.A. Valenzuela, "A statistical model for indoor multipath propagation," IEEE Journal on Selected Areas in Communications, Vol. 5, pp. 128–137, February 1987.

[4] M. K. Simon, S. M. Hinedi, and N. C. Lindsey, "Digital communication techniques signal design and detection," Prentice Hall PTR, 2003.

[5] D. Sen, S. Chakrabarti, and R. V. Raja Kumar, "An efficient frequency offset estimation scheme for multi-band OFDM ultra-wideband systems," IEEE VTC-2008-Spring, Singapore, pp. 973–977, May 11–14, 2008.

[6] S. M. Kay, "Fundamentals of statistical signal processing: Estimation theory," Prentice Hall PTR, 1993.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online November 2008 in SciRes (http://www.SRPublishing.org/journal/ijcns/).

Cluster-Based Design for Two-hop Cellular Networks

Hrishikesh VENKATARAMAN[1], Sinan SINANOVIĆ[2], Harald HAAS[2]
[1]*School of Engineering and Science, Jacobs University Bremen, Germany*
[2]*Institute for Digital Communications, University of Edinburgh, UK*
Email: hrishikesh@ieee.org, {s.sinanovic, h.haas}@ed.ac.uk
Received April 22, 2008; revised September 1, 2008; accepted September 7, 2008

Abstract

Optimal resource allocation with an objective of maximizing the system capacity is an NP-hard problem in multihop cellular networks. Hence, different heuristic algorithms have been developed over the years that would improve the network system capacity. In this paper, a novel cluster-based architecture is proposed for a two-hop cellular network whereby the transmission distance between any communicating pair is restricted to half the cell radius. In this design, a given radio resource is used by two simultaneously communicating pairs in every hexagonal cell, but for only half the time slot period. The characteristic feature of this cluster-based design is that it enables a frequency reuse ratio of *one*. The proposed hierarchical system is analyzed and tested under realistic propagation conditions including lognormal shadowing. It has been observed that the system capacity of a cluster-based design is 2.5 times that obtained from the single-hop cellular system with no relaying. In addition, the cluster-based design achieves higher capacity compared to state-of-the-art two-hop algorithms. This is an important finding since the hierarchical cluster-based approach has fewer degrees of freedom in the selection of the routing path for the end-to-end connection. Practical routing algorithms should be able to benefit from this.

Keywords: Cluster-Based Design, Synchronized Resource Reuse, Interference Model, Spatial Protection Margin

1. Introduction

Some of today's key challenges in the design of wireless systems are to provide high peak data rates as well as to provide a network architecture that allows for an efficient utilization of the scarce spectrum resources while the power consumption of the network is minimized. Currently deployed single-hop cellular networks as a stand-alone technology are handicapped by numerous limitations, viz., inability to cover *dead zones*, high attenuation of signals, high shadowing, inefficient use of energy, etc. A direct evolution of the existing cellular network architecture are multihop hybrid cellular networks, where the communication between the mobile station (MS) and the base station (BS) takes place in multiple hops [1]. One can envisage a multihop hybrid cellular network as a means to enable sharing of information between possibly mobile sensor nodes or gathering of sensed information toward query points on a wireline network [2]. Alternatively, one can view multihop cellular network models as a method to extend the communication coverage and provide higher data rate for an infrastructure-based cellular network [3]. Such a hybrid network model aims at providing global connectivity. At the same time, it seeks to mitigate interference, and to maximize the system capacity of the network while achieving a frequency reuse of *one*. This is the target being aimed for the development of 4th generation, or IMT-Advanced, wireless networks. These systems are primarily based on orthogonal frequency division multiplexing (OFDM), time division multiple access (TDMA), frequency division multiple access (FDMA) and time division duplexing (TDD) [4].

In recent years, there has been extensive research work in the direction of capacity scaling for multihop hybrid wireless networks. It is shown in [5] that having the infrastructure based BS component over the multihop *ad hoc* network drastically increases the connectivity of the network. For a multihop hybrid cellular network with n nodes and m BSs, the results in [6] show that if m grows asymptotically slower than $\sqrt{n}$, the benefit of adding BSs on capacity is insignificant. However, if m grows faster than $\sqrt{n}$, the system capacity increases linearly with the number of BSs providing an effective improvement over the multihop *ad hoc* network. Therefore, in order to achieve non-negligible capacity gain, the investment in the wired infrastructure should be high enough. It has been shown through outage analysis in [1] that an integration of cellular and multihop communication models results in better relaying and avoids traffic congestion. Deploying relays can clearly help improve the performance of the users near the edges of the cell and has the potential to solve the coverage problems for high data rates in macrocells [7]. TDD is the enabling technology for the multihop design [8]. Hence, by having simultaneous transmission by both BSs and relays, capacity gains can also be achieved in the cellular network. However, a multihop hybrid cellular design also requires extra radio resources for relaying hops and is sensitive to the quality of relaying routes. Therefore, multihop hybrid cellular networks require a well-designed radio resource allocation strategy in order to secure performance gains. A hybrid architecture, viz., mobile assisted data forwarding (MADF) is proposed in [9], wherein, a multihop relaying system is overlaid on the existing cellular networks. The main objective of this system is to dynamically divert the traffic load from a *hot cell* (highly loaded cell) to *cooler cells* (lightly loaded) in its neighborhood. Similarly, a multihop cellular network (MCN) architecture is investigated in detail in [10,11,12], wherein, the end-to-end communication is always between the MS and BS, like in a traditional single-hop cellular network. There has been considerable research work in finding different routing techniques for multihop cellular networks, viz., base assisted *ad hoc* routing (BAAR), base-driven multihop bridge routing (BMBP), [13] single-interface multihop cellular network routing protocol (SMRP) [14], for different kinds of traffic patterns. These techniques effectively utilize the *ad hoc* relaying in presence of fixed infrastructure in order to achieve enhanced network capacity. However, it has been shown in [15] that optimum resource allocation in multihop cellular networks, with the objective of throughput maximization with radio resource allocation (TM-RRA) is an NP-hard problem. In fact, the well-known multiple choice knapsack problem, (MCKP), which is proved to be NP-hard [16], is shown to be a restricted version of TM-RRA [15]. Hence, researchers across the scientific community have worked towards designing suboptimal but efficient heuristic algorithms and architecture designs.

In this paper, a multihop cellular network model is designed such that all communication between the source and destination nodes is routed through the BS. All the mobile nodes communicate with the BS in either single-hop or two hops. However, the focus of this paper is on designing a novel cluster-based architecture for a two-hop cellular network. Section 2 presents the system model and the underlying mechanism of the cluster-based design. Section 3 explains how the system capacity is calculated for such a design. A deterministic cluster-based technique is described in Section 4 and a semi-analytical model is presented for calculating the different interferences and the carrier-to-interference ratio, γ. In addition, the capacity bounds for the semi-analytical model and the confidence interval due to lognormal shadowing are also calculated in Section 4. The simulation model and the results of the cluster-based design are presented in Section 5 and Section 6 respectively, along with a performance comparison with other benchmark algorithms. Finally, conclusions are provided in Section 7.

2. System Model

A multi-cellular system, with a BS at the center of each cell, is considered in the network design. The maximum distance between the BS and the edge of the cell is given by, r. There are 19 hexagonal cells in the coverage area. These 19 cells are arranged such that a center cell is surrounded by six cells in the 1st tier and twelve cells in the 2nd tier. A *Protocol Model* [17] is considered in the system design in order to reduce the interference. According to this model, a circular exclusion region is defined around every communicating receiver, such that no other transmitter apart from the desired transmitter communicates in this exclusion region. The radius of this circular exclusion region, r_c, is given by the following equation.

$$r_c = (1+\Delta)d_c \tag{1}$$

In the literature [17], r_c is also sometimes termed as the exclusion range. In Equation (1) above, d_c is the distance between the transmitter and receiver of any communicating pair, and $\Delta \geqslant 0$ is the spatial protection margin, that indicates the ratio of increase of the exclusion range distance to the transmission range distance. Hence, at any time instant in a TDMA system, all receivers of the simultaneously communicating pairs are inherently separated from the unintended transmitters, by at least, the exclusion range distance, i.e., the minimum distance between any receiver and an unintended transmitter is at least $(1+\Delta)$ times the transmission distance of the desired communication pair. It is shown in [18] that under the *Protocol Model*, the system capacity is maximized when the spatial protection margin of the *Protocol Model* is around $\Delta=1.0$. Hence, a spatial protection margin of $\Delta=1.0$ is considered throughout this work, while designing the cluster-based architecture for the two-hop cellular network.

The proposed cluster-based design is based on the formation of multiple clusters in every cell. Each cell is initially divided into two layers, viz., inner layer and outer layer.

1) *Inner Layer:* This is the circular area contiguous to the BS; and the MS in this zone communicate to the BS directly using a single-hop. The distance between the MS and BS in the inner layer is always less than or equal to half the cell edge length, i.e., $r/2$.

2) *Outer Layer:* This circular region is located around the inner zone, and the MSs located in this region communicate to the BS in two hops. This area is further divided into several clusters. The MSs within any of the clusters would communicate to the BS via a cluster-head node, called the gateway (GTW). The GTWs are located on the boundary adjoining the inner and outer layer, i.e., at a distance of $r/2$ from the BS. Since the radius of the hexagonal cell is r, the maximum distance between the GTW and the MSs in the outer layer would be $r/2$. Hence, the maximum transmission distance in the cell, i.e., between BS and GTW (inner layer), or between MS and GTW (outer layer) is $r/2$.

The MSs located in the outer layer are grouped into several clusters. A single cell scenario depicting the schematic of a cluster-based two-hop cellular architecture is shown in Figure 1. There are six circular clusters in each cell. For each of the clusters, a wireless terminal located at the boundary of the inner and outer layer of the cell is selected as a cluster-head node, alternatively known as GTWs. There are six GTWs/cell, each of them located at a distance of $r/2$ from the BS. Cluster-heads GTW_{1a} and GTW_{1b} are diametrically opposite to each other and are separated by a distance of r, i.e., *twice* the transmission distance, $r/2$. The same holds for the cluster-heads GTW_{2a} and GTW_{2b}, and for GTW_{3a} and GTW_{3b}. In practice, the GTWs could be fixed relay stations (RSs), located on the street lamps/roof tops, or, MSs/wireless terminals with their own traffic. In case of fixed RSs/GTWs, they could probably be placed at exactly $r/2$ from the BS.

However, if the MSs are selected as relays, then the exact location of the relay node would depend on the distribution and the density of the mobile terminals. Hence, the selected GTW node could be located at a distance slightly less or greater than $r/2$ from the BS, and also, the GTWs would not be equidistant to each other. There would be a small yet noticeable difference in the system performance due to fixed/mobile GTWs, and is explained later in Section 6. In addition, a deterministic cluster-based design is considered in the semi-analytical model, later in this paper, where the six GTWs in the cell are assumed to be both equidistant to each other, located exactly at a distance of half the cell radius from the BS, and most importantly, the MSs in the outer-layer of the cell are assumed to be at a maximum distance of $r/2$ from the cluster-head GTWs. An important point to be noted is that the number of clusters per cell in the cluster-based design need not be always six. It could be two, four, six, eight, ten or even higher. The only condition is that the number of clusters per cell has to be an even number, due to the basic principle of simultaneous transmission of communication pairs located in the diametrically opposite clusters in the cell. However, in practice there is a limitation that as the number of clusters per cell is increased, the amount of resources that could be given to one cluster decreases. In a very recent work, it is shown [19] that for a cellular network, with six BSs surrounding the central BS, the optimum number of GTWs in each cell that maximizes the system capacity is *six*.

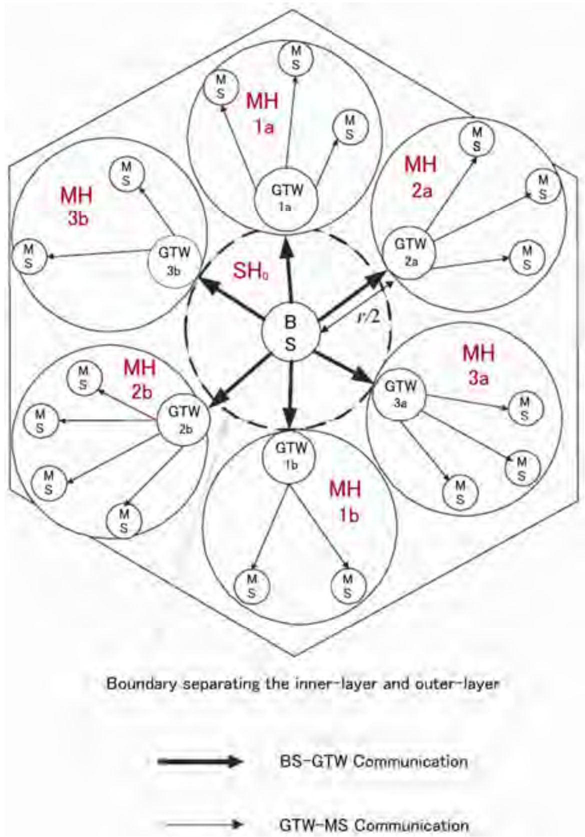


Figure 1. Schematic model of a cluster based two-hop cellular network (downlink).

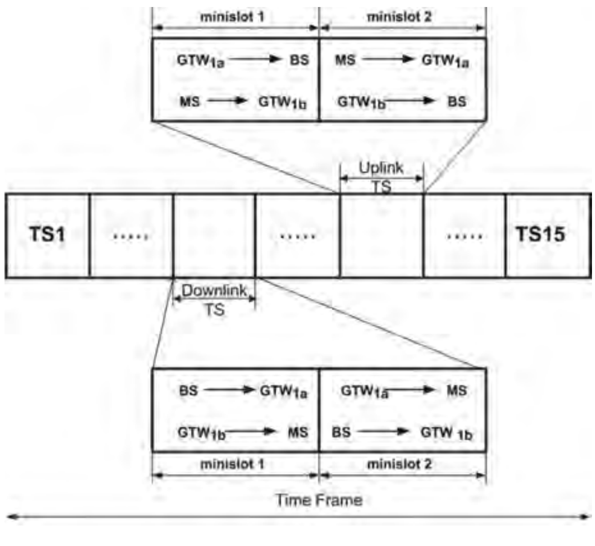


Figure 2. Synchronized resource reuse mechanism for a TDD/TDMA cluster based two-hop cellular network.

This justifies the selection of six GTWs in the cluster-based two-hop cellular design. The variation of system capacity for different number of GTWs per cell is shown in Section 6. In addition, the cell region could be divided such that the radius of the inner layer is τr where 0 <

$\tau < 1$ is the ratio of the inner layer radius to the radius of the cell [20]. In that case, the maximum transmission distance would no longer be restricted to $r/2$. However, the cluster-based design would be still valid, as would be observed later when the GTWs are selected from the MSs.

In order to understand the complete working mechanism, a conceptual model of the cluster-based two-hop cellular network with six clusters/cell and equidistant GTWs at a fixed distance of $r/2$ from the BS is considered, and the underlying principle of the synchronized resource reuse technique is described as follows:

1) As shown in Figure 1, SH_0 is the inner-layer (single-hop region) and MH_{1a}, MH_{1b}, … MH_{3b} are the two-hop clusters in the outer layer. In addition, GTW_{1a}, GTW_{1b} … GTW_{3b} are the respective cluster-heads for MH_{1a}, MH_{1b} … MH_{3b}. Each cluster contains a number of MSs. In case of downlink communication between BS and a wireless terminal located in any of the clusters, the BS would communicate to the cluster-head GTW in the 1st hop, and in the 2nd hop, the GTW would communicate to the MSs associated with the corresponding clusters. Similarly, in the uplink, the MSs in any of the clusters communicate with the BS in two hops, wherein, the MS communicates to the GTW in the 1st hop, and the GTW communicates to the BS in the 2nd hop.

2) A TDD/TDMA scheme is considered for the cluster-based two-hop network. For a multihop system with number of hops per link, $M \geqslant 1$, the signal for any hop can be transmitted only for T/M time slot duration, where T is the TS period. Hence, the TS is divided into two minislots for the two-hop links. However, for a wireless node located in the inner layer (SH_0), the communication between the wireless terminal and the BS would take place in single-hop, for the full duration of one TS.

3) The reusability of the resources is increased by allowing two multihop clusters in any cell to occupy the same TS at the same frequency. As shown in Figure 1, the clusters MH1a and MH1b are located at diametrically opposite sides of the BS. The synchronized TDD frame structure for both uplink and downlink is shown in Figure 2. In the downlink, GTW_{1a} can download to the MS in its cluster in a particular time slot. At the same time instant, the BS could download to GTW_{1b} in the opposite cluster of the same multihop cell. It should be noted that both these simultaneously communicating pairs are outside the exclusion region of each other. Similarly, in the next time slot, the $GTW_{1b} \rightarrow$ MS and BS $\rightarrow GW_{1a}$ communication takes place simultaneously.

4) In the uplink, the transmitters and receivers of the cluster-based model are reversed, as seen in Figure 2, but the governing principle of the resource reuse technique remains the same. It can be therefore noted that the reuse of the resources can be done independently for both uplink and downlink, using the synchronized resource reuse technique. Hence, this cluster-based design remains valid even for asymmetric traffic, as long as the traffic asymmetry remains the same for all the cells.

5) The given TS resource is also allotted to each of the hexagonal cells in the system. As shown in Figure 3, for both uplink and downlink, the transmitters of all the concurrently communicating pairs in the adjacent cells are beyond the exclusion region of the desired receiver in the intended cell. Hence, as shown in Figure 4, a given TS resource is not only used by two simultaneously communicating pairs in any cell, but also, the same TS resource is reused in every cell. However, it is to be noted that a TS resource given to a communicating pair in a two-hop network is only half the time slot period given to an equivalent single-hop network, as shown in Figure 2. Due to using a resource twice within a cell, however, the cluster-based design effectively results in a frequency reuse factor of *one*.

6) The GTWs can be considered to be equidistant and located at approximately $r/2$ from the BS if they are fixed terminals. However, if the GTWs are not fixed, and are selected from the distributed MSs, then, the wireless terminal located at either half the cell radius or closest to half the cell radius (either in the inner-layer or in the outer-layer) is selected as a GTW. Irrespective of whether the selected GTW is in the inner layer or outer layer, the transmission distance of the communicating pair between

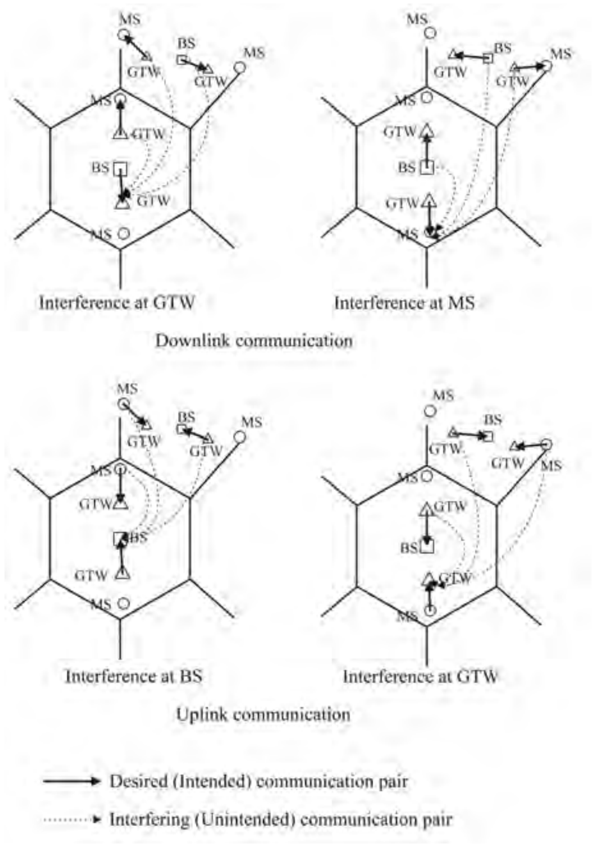


Figure 3. Interference reduction mechanism for cluster based two-hop cellular network using synchronized resource reuse technique.

the BS and GTW would be then different from $r/2$. Correspondingly, the transmission distance of the GTW - MS pair would also vary.

3. Capacity Calculation

All the wireless terminals in any cell are assumed to transmit their signals with the same power, P_T. If d_c is the transmission distance between any communicating pair, then the power received, P_R, using a general propagation model is given by:

$$P_R = P_T - (k_1 + 10\alpha \log_{10}(d_c) + \xi_c) \quad [\text{dB}] \tag{2}$$

where k_1 is a constant that depends on the propagation environment (indoor/urban/suburban), α is the path loss exponent and ξ_c is the shadowing factor across the transceiving pair. In a multi-cell scenario, the given radio resource is utilized by all the cells in the system. The transmitters of all simultaneously communicating pairs in the seven-cell scenario are marked and shown in Figure 4 along with the interference calculation at the receiver gateway (i.e., the receiver of BS → GTW pair) of the center cell (cell 0). The thick arrows in Figure 4, from BS → GTW and GTW → MS in all the cells represent the simultaneous communicating pairs. A reference line (dotted line in the figure) is considered that connects the BS of cell 0, cell 1 and cell 4. The dashed lines from the transmitting BSs and GTWs of cell 2 and cell 3, to the Rx GTW in the center cell indicates the distance of the Rx GTW from other interfering transmitters in cell 2 and cell 3.

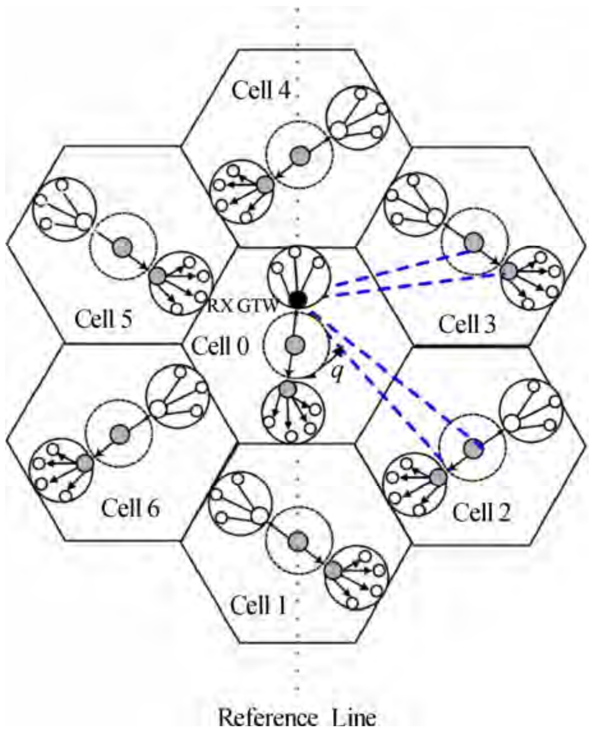


Figure 4. Distance calculation for different interfering entities in the downlink model.

In all, the Rx GTW in cell 0 would experience interference from thirteen interferers: two interferers from each of the six adjacent cells and one transmitting GTW (of GTW → MS pair) from cell 0.

For any communicating pair, the inter-cell interference only across six adjacent cells, i.e., the 1st tier of cells is considered. The transmitting interferers from the 2nd tier of cells are very far from the intended receiver, and hence, the interference generated from these transmitters is assumed to be negligible. Therefore, for any communicating pair in this cluster-based model, there is one interferer from own cell and two interferers from each of the adjacent cells. The carrier-to-interference ratio is therefore calculated as follows:

$$\gamma = \frac{10^{-[k_1+10\alpha\log_{10}(d_c)+\xi_c]}}{\sum_{i=1}^{N_I} 10^{-[k_1+10\alpha\log_{10}(d_i)+\xi_i]}} \tag{3}$$

where d_i is the distance of the desired receiver from the i^{th} interfering entity and N_I is the total number of interfering entities for any receiver in a cluster-based model. ξ_i accounts for shadowing between the desired receiver and the i^{th} interfering transmitter. The capacity in bps/Hz/cell is calculated by finding the system capacity independently over seven cells (center cell and six cells in the 1st tier), as shown in Figure 4, and averaging over them. Each cell in the 1st tier is surrounded by six cells out of which three cells belong to the 2nd tier. The traffic in the twelve cells of the 2nd tier only contribute for the intercell interference calculation for the 1st tier of cells. This 2nd tier of cells is necessary to remove the boundary effects while calculating γ for the 1st tier of cells, and hence, the Shannon bound is not calculated for the twelve cells in the 2nd tier. As shown in Figure 2, the data across each communicating pair is transmitted for only half the time slot period in a two-hop system. As a consequence, the Shannon capacity has to be scaled by a factor of 1/2. Also, in each of the seven cells, there are two simultaneously communicating pairs, and depending on the distance of the interfering transmitters the receivers of these two communicating pairs would have different values of γ. Therefore, the system capacity (of only the two-hop links) is calculated from the Shannon equation as:

$$C = \frac{1}{2N_c} \sum_{i=1}^{N_c} \sum_{j=1}^{N_1} \log_2(\gamma_{ij} + 1) \,\text{bps/Hz/Cell} \tag{4}$$

where γ_{ij} is the carrier-to-interference ratio of the j^{th} communicating pair in the i^{th} cell. N_l is the number of concurrently communicating pairs in the outer layer that use the same radio resource, in any single cell. For a cluster-based design, two pairs located diametrically opposite to each other communicate simultaneously, i.e., $N_l=2$. $N_C=7$ is the number of cells over which the system capacity is calculated. In order to calculate the average per-cell system capacity, the Shannon capacity equation in Equation (4) is summed up over all N_c cells and averaged over them.

4. Semi-Analytical Model

In order to assess the performance of the synchronized resource reuse technique for the cluster-based design, a semianalytical model is developed for a deterministic cluster-based two-hop network. In this deterministic cluster-based model, the GTWs are fixed and located exactly at a distance of $r/2$ from the BS. Also, the GTWs are equidistant from each other. Hence, the six GTWs in the cell represent the six vertices of a regular hexagon, with a side length of $r/2$. In addition, the MSs in the outer layer are assumed to be always located at a distance of $r/2$ from their respective GTWs; and also, the MSs are assumed to be uniformly distributed between [0°,360°] across the outer layer of the cell. Hence, the distance between the BS and GTW, and also the distance between the MS and its corresponding GTW is fixed. This simplifies the analysis for numerically calculating the Shannon capacity of the cluster-based two-hop network. However, the precise location of the GTW in the cell is determined from the angle made by the BS - GTW pair with the reference line, as shown in Figure 5. q_{11} indicates the angle made by the 1st communication pair, BS → GTW, in the intended cell (cell 0 in this case) with the reference line of cell 1, whereas, q_{11} indicates the angle

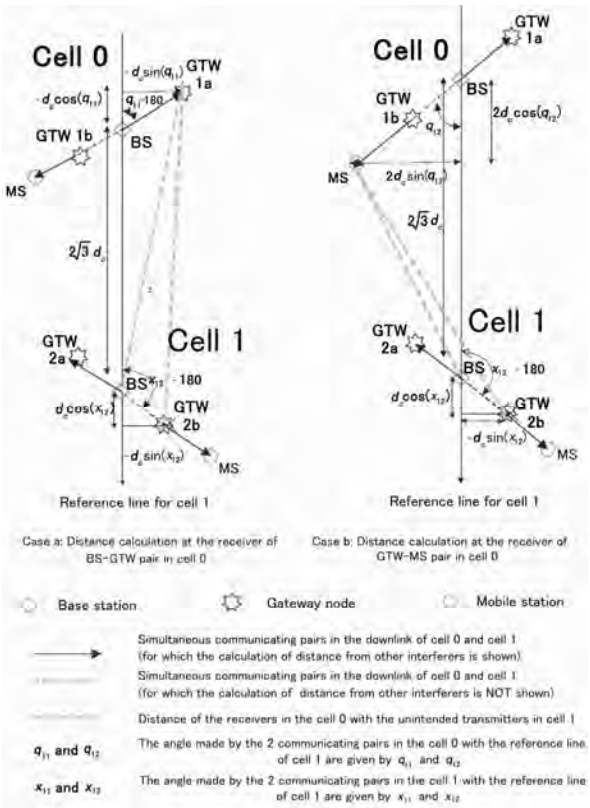


Figure 5. Calculation of distance at the receivers of BS → GTW and GTW → MS pairs (downlink) from the two interfering transmitters of i^{th} cell.

made by the 2nd communication pair, GTW → MS, in the intended cell (cell 0 in this case) with the reference line of cell 1. Similarly, x_{11} and x_{12} indicates the angles made by the 1st interfering communication pair, BS → GTW, and the 2nd interfering communication pair, GTW → MS, (both) in cell 1 with the reference line of cell 1. The precise location of any MS is determined by the angle made by the line joining the MS and its corresponding GTW with the reference line. At the same time, in the deterministic cluster-based model, the trans- mission distance is always the maximum possible value, $r/2$.

In Section 6, the capacity results obtained from the deterministic cluster-based model are compared with the simulation model. For both uplink and downlink, this Semi-analytical model first calculates the distance of the intended receiver from all simultaneously communicating intended transmitters. A downlink schematic for a seven-cell cluster-based two-hop model is shown in Figure 4. It should be noted from Figure 4 that only the locations of the BSs are fixed. The distance between two BSs is $\sqrt{3}\,r$. Since the transmission distance, $d_c=r/2$, for the cluster-based model, the distance between two BSs can also be written as $2\sqrt{3}d_c$. In Figure 4, all the transmitters in the seven cells are shaded with gray background. The black circle in the center cell marks one of the desired receivers which would experience interference from other unintended transmitters. The distance between the black circle (desired receiver) and all the gray colored circles (interfering transmitters from the own cell and all the adjacent cells) marks the distance of the different interfering entities. Hence, as shown in Figure 4, the total interference experienced by a receiver depends on the relative distance between this receiver and all its interfering transmitters.

4.1. Carrier-To-Interference Calculation for Downlink.

In the downlink, the communication takes place from BS → GTW and from GTW → MS. Figure 5 (both, case a and case b) shows the simultaneously communicating pairs in cell 0 and cell 1 in the downlink scenario. As seen in Figure 5, there are two simultaneously communicating pairs per cell, i.e., the BS → GTW pair and GTW → MS pair. The receivers of cell 0 would experience interference not only from its own cell, but also from the simultaneously communicating pairs from other cells. The interference experienced by the communicating pairs in cell 0 are calculated as follows:

1) *BS →* GTW *Communication in the Intended Cell:*

When the gateway in the intended cell is the desired receiver (say, GTW1a in cell 0 in Figure 5), the distance between this gateway and the interfering transmitters of the adjacent cell are calculated as shown in Figure 5(a). There are two cells, cell 0 and cell 1. Using basic trigonometry, the distance of the communicating receiver in cell 0 from the interfering transmitters in cell 1 is computed. As shown in Figure 5(a), the distance of receiving gateway at cell 0, GTW1a, from the BS of cell 1 is given by:

$$d_{\mathrm{BS}_1}{}^2 = (2\sqrt{3}d_c - d_c\cos(q_{11}))^2 + (d_c\sin(q_{11}))^2 \qquad \textbf{(5)}$$

whereas the distance of the unintended transmitting gateway of the cell 1, GTW2b, to the desired gateway receiver in cell 0 is given by:

$$\begin{aligned} d_{\mathrm{GTW}_1}{}^2 &= (2\sqrt{3}d_c + d_c\cos(x_{12}) - d_c\cos(q_{11}))^2 \\ &\quad + (d_c\sin(x_{12}) - d_c\sin(q_{11}))^2 \qquad \textbf{(6)} \end{aligned}$$

The angle, q_{11}, is formed between the line joining the communicating pairs, BS→GTW1a in cell 0 with the reference line of cell 1. Similarly, x_{12} is the angle between the line joining the communicating pairs, GTW2b → MS in cell 1, with the reference line of cell 1. The above equations, Equation (5) and Equation (6) could be generalized to calculate the interference coming from the transmitters of all the six adjacent cells into the desired receiver, i.e., the GTW of the intended cell. By changing the reference line for each of the six adjacent cells, the distance of the interfering transmitters from the i^{th} cell can be calculated as follows:

$$\begin{aligned} d_{\mathrm{BSi}}{}^2 &= (2\sqrt{3}d_c - d_c\cos(\theta_{i1}))^2 \\ &\quad + (d_c\sin(\theta_{i1}))^2 \qquad \textbf{(7)} \\ &= 13d_c{}^2 - 4\sqrt{3}d_c{}^2\cos(\theta_{i1}) \qquad \textbf{(8)} \end{aligned}$$

whereas the distance of the unintended transmitting GTW to the desired GTW receiver is given by:

$$\begin{aligned} d_{\mathrm{GTW}_i}{}^2 &= (2\sqrt{3}d_c + d_c\cos(\phi_{i2}) - d_c\cos(\theta_{i1})) \\ &\quad + (d_c\sin(\phi_{i2}) - d_c\sin(\theta_{i1}))^2 \qquad \textbf{(9)} \\ &= 14d_c{}^2 + 4\sqrt{3}d_c{}^2(\cos(\phi_{i2}) - \cos(\theta_{i1})) \\ &\quad - 2d_c{}^2\cos(\phi_{i2} - \theta_{i1}) \qquad \textbf{(10)} \end{aligned}$$

Here,

$$\theta_{i1} = q_{i1} + 60(i-1) \qquad \textbf{(11)}$$

is the angle in degrees made by the BS→GTW communicating pair in the intended cell with the reference line of the i^{th} cell, and

$$\phi_{i2} = x_{i2} + 60(i-1) \qquad \textbf{(12)}$$

is the angle in degrees made by the GTW → MS in the i^{th} cell with the reference line of the i^{th} cell (Figure 5(a) shows the angle x_{12} made by the GTW2b → MS communicating pair in cell 1, with the reference line of cell 1). It should be noted that all θ and ϕ vary uniformly from [0^{o}, 360^{o}]. In addition, the distance of the intra-cell interfering transmitter is, $d_{\mathrm{owncell}}=2d_c$. The carrier-to-interference value at the receiver of any communication pair is therefore given by:

$$\gamma = \frac{d_c{}^{-\alpha}}{(2d_c)^{-\alpha} + \sum_{i=1}^{6}(d_{(\mathrm{GTW})_i})^{-\alpha} + \sum_{i=1}^{6}(d_{(\mathrm{BS})_i})^{-\alpha}} \qquad \textbf{(13)}$$

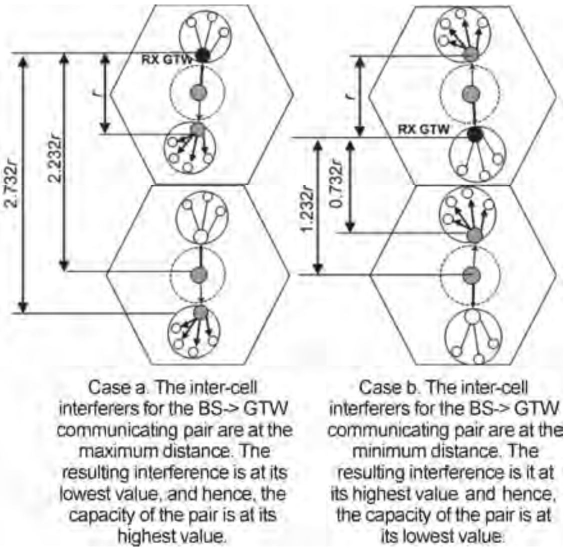


Figure 6. Minimum and maximum distance of the inter-cell interfering entities (downlink) from the receiver of the BS → GTW.

Dividing the numerator and denominator by $d_c{}^{-\alpha}$ results in:

$$\gamma = \frac{1}{2^{-\alpha} + \sum_{i=1}^{6}(\beta_{(\mathrm{GTW})_i})^{-\alpha} + \sum_{i=1}^{6}(\beta_{(\mathrm{BS})_i})^{-\alpha}} \qquad \textbf{(14)}$$

where

$$\beta_{(\mathrm{BS})i}{}^2 = 13 - 4\sqrt{3}\cos(\theta_{i1}) \qquad \textbf{(15)}$$

$$\begin{aligned} \beta_{(\mathrm{GTW})i}{}^2 &= 14 + 4\sqrt{3}\cos((\phi_{i2}) - \cos(\theta_{i1})) \\ &\quad -2\cos(\phi_{i2} - \theta_{i1}) \qquad \textbf{(16)} \end{aligned}$$

If the orientation of the GTW is fixed with respect to the BS, then the interference and the capacity of the BS → GTW pair in the semi-analytical model varies only with the location of all the interfering transmitters from own cell and adjacent cell. Hence, the best and worst case for the capacity of BS → GTW pair in the intended cell can be calculated by considering all the interferers to be located at the maximum and minimum distance from the intended receiving GTW of the BS → GTW pair.

Upper bound for capacity:

As can be seen from Figure 6(a), the own-cell interferer (transmitting GTW from the diametrically opposite cluster) is at a fixed distance of r from the intended receiver (GTW of cell 0). Similarly, the interfering transmitters of the adjacent cell (BS and GTW) are at a distance of $d_{\mathrm{BS}} = \sqrt{3}r + \frac{r}{2} \approx 2.232r$ and $d_{\mathrm{GTW}} = \sqrt{3}r + r \approx 2.732r$ respectively from the intended receiver. It should be observed that these intercell interferers are at a maximum possible distance from the intended receiver; and hence, causes the least interference to the intended communicating pair. The carrier power at the receiver of the BS → GTW pair is given by

$$P_R = P_T - (k_1 + 10\alpha\log10(r/2)) \qquad \textbf{(17)}$$

For a multi-cellular network with r=217m, a propagation model [21] with k_1=37, α=4, and assuming a transmit power of 1W, i.e., 0 dBW, the received power, P_R, would be -118.4 dBW. The BS → GTW pair in the intended cell would experience interference from six adjacent cells. Hence, the upper bound for the capacity of the BS → GTW communicating pair is obtained by substituting the values of these interferences distances into Equation (13) for calculating the resulting interference. The upper bound of the capacity for the BS → GTW communicating pair is then obtained by substituting the resulting interference into Equation (4). The total interference power experienced by the receiving GTW node would be -129.2 dBW. The upper bound for the capacity of BS → GTW pair would be 3.68 bps/Hz.

Lower bound for capacity:

The lower bound for the capacity of GTW → MS pair is obtained by considering the minimum distance of the inter-cell interfering entities. As can be observed from Figure 6(b), the minimum distance of the interfering transmitters (BS and GTW) from the adjacent cell are $d_{\mathrm{BS}} = \sqrt{3}\,r - \frac{r}{2} \approx 1.232r$ and $d_{\mathrm{GTW}} = \sqrt{3}\,r - r \approx 0.732r$ respectively from the intended receiver. The resulting lower bound for the capacity of BS → GTW pair is 0.72 bps/Hz.

2) *GTW → MS Communication Pair in the Intended Cell:*

For the 2nd active communication pair, GTW → MS, in the downlink of cell 0, the maximum distance of the intended receiver, i.e., the MS of the GTW → MS pair, from the BS is *twice* the transmission distance. Therefore, as seen in Figure 5, the MS in cell 0 is located at a distance of $d = 2d_c = r$ from the BS, and is distributed uniformly from [0°, 360°]. The distance of the interfering BS and the interfering GTW from the cell 1 is calculated from basic trigonometry as:

$$d_{\mathrm{BS}_1}{}^2 = (2\sqrt{3}d_c - 2d_c\cos(q_{12})) + (2d_c\sin(q_{12}))^2 \quad \textbf{(18)}$$

$$d_{\mathrm{GTW}_1}{}^2 = (2\sqrt{3}d_c - d_c\cos(x_{12}) - 2d_c\cos(\theta_{12}))^2 + ((d_c\sin(x_{i2}) - 2d_c\sin(\theta_{12}))^2 \quad \textbf{(19)}$$

This equation for calculating the distances of the interfering transmitters from other cells could be written in a generalized form as:

$$d_{\mathrm{BS}_i}{}^2 = (2\sqrt{3}d_c - 2d_c\cos(\theta_{i2}))^2 + (2d_c\sin(\theta_{i2}))^2 \quad \textbf{(20)}$$

$$d_{\mathrm{GTW}_i}{}^2 = (2\sqrt{3}d_c + d_c\cos(\phi_{i2}) - 2d_c\cos(\theta_{i2}))^2 + (d_c\sin(\phi_{i2}) - 2d_c\sin(\theta_{i2}))^2 \quad \textbf{(21)}$$

where, $\theta_{i2} = q_{i2}+60(i-1)$ is the angle in degrees made by the GTW → MS pair in the intended cell with the reference line of the i^{th} cell (Figure 5(b) shows the angle q_{12} between the communicating pair, GTW1b → MS with the reference line of cell 1). Equation (20) is simplified and the corresponding equations for $\beta_{(\mathrm{BS})}$ and $\beta_{(\mathrm{GTW})}$ are given as:

$$\beta_{(\mathrm{BS})_i}{}^2 = 16 - 8\sqrt{3}\cos(\theta_{i2}) \quad \textbf{(22)}$$

$$\beta_{(\mathrm{GTW})_i}{}^2 = 17 + 4\sqrt{3}(2\cos(\theta_{i2}) - \cos(\phi_{i2})) - 4\cos(\phi_{i2} - \theta_{i2}) \quad \textbf{(23)}$$

It should be noted that the equation for γ remains the same as given in Equation (14).

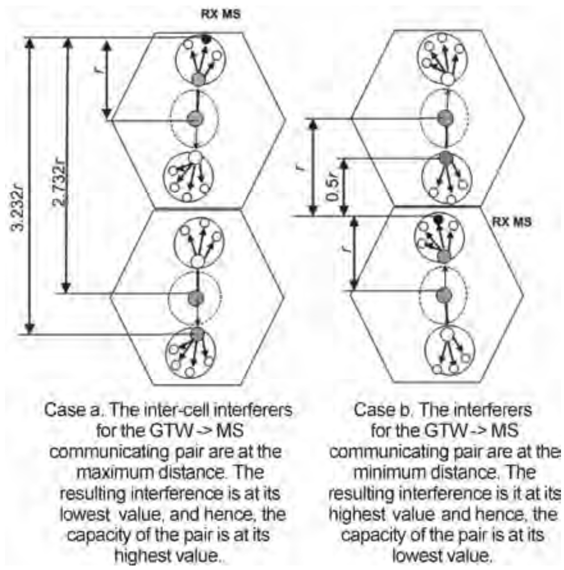


Figure 7. Minimum and maximum distance of the inter-cell interfering entities (downlink) from the receiver of the GTW → MS pair in the intended cell.

Upper bound for capacity:

In the semi-analytical model, the MS in the GTW → MS pair is assumed to be at a maximum distance of $r/2$ from the GTW. The own-cell interferer is the transmitter from the diametrically opposite cluster. Hence, the own cell interferer is at a distance of r from the desired receiver. It can be observed from Figure 7 that the maximum distance of the interferers from an adjacent cell are $d_{\mathrm{BS}} = \sqrt{3}\,r + \frac{3r}{2} \approx 2.732r$ and $d_{\mathrm{GTW}} = \sqrt{3}\,r + \frac{3r}{2} \approx 3.232r$ respectively. Hence, substituting these values into Equation (13) and Equation (4) results in the upper bound for the capacity of the GTW → MS communicating pair, which is 3.91 bps/Hz.

Lower bound for capacity:

Similarly, the lower bound for the capacity of the GTW → MS pair is calculated by knowing the distance of all the interfering entities. The own cell interferer is at a distance of r, whereas the minimum distance of the interferers from any adjacent cell are 0.5r and r, as is shown in Figure 7. Hence, the lower bound of the capacity for the communicating pair is 0.21 bps/Hz.

3) *Capacity Bounds for End-to-End Link:* In a two-hop system, the capacity of the end-to-end link is limited by the lower value of the capacity of any of the two-hops of the link, i.e., if the capacity of the 1st hop of the link is less than the capacity of the 2nd hop of the link, then the capacity of the end-to-end link is limited by the capacity of the 1st hop of the link [22]. Hence, the capacity of the end-to-end link in the downlink scenario in the cluster-based design is limited by the lesser of the capacity values of BS → GTW pair and GTW → MS pair. Hence, the lower and upper bounds of the capacity for the cluster-based two-hop design is limited by the capacity values of the BS → GTW pair, and is equal to 0.21 bps/Hz/cell and 3.68 bps/Hz/cell respectively.

It should be noted at this stage that the upper bound of the GTW → MS pair would be higher than 3.91 bps/Hz/cell if the distance between the GTW and MS is less than $r/2$. However, since the upper bound of the downlink capacity for the cluster-based design is limited by the lesser of the capacity values of BS → GTW and GTW → MS pair, the higher values of the capacity obtained by the GTW → MS pair does not change the capacity result of the cluster-based design.

4) *Effect of Lognormal Shadowing on Capacity Bounds:* In presence of lognormal shadowing, the bounds for the system capacity would vary. The amount of variation would depend on the standard deviation of the lognormal shadowing. Due to the summation of lognormal variables in the calculation of the interference power (as shown in Equation (3)), it is very difficult to find an exact expression that would reflect the effect of lognormal shadowing on the system capacity. Instead, the effect of lognormal shadowing on the capacity bounds is computed in this paper. Lognormal shadowing with zero mean and standard deviation of 4 dB is considered throughout the analysis [23]. The received power for the carrier signal would vary within a value of 2ζ, i.e., 8 dB, for a confidence measure of 95% [24].

Effect on upper bound of capacity:

The upper bound of the capacity in case of downlink communication is 3.68 bps/Hz/cell, achieved for the BS → GTW communication. The corresponding value of the carrier power and total interference power are -118.4 dBW and -129.2 dBW respectively, resulting in a γ of 10.8dB. An 8 dB variation in the signal strength implies that the carrier power would be between -110.4 dBW and -126.4 dBW for 95% of the cases, i.e., a variation of 6.75% on either side of -118.4 dBW. There are thirteen transmitters that would interfere with the intended communicating pair and the resulting interfering power would be a summation of all these interferers. If the same confidence measure of 95% is to remain for both the carrier signal and the resulting interference signal, then a higher confidence measure should be assumed for each of the interfering entities. A confidence measure of 99.6% for the power received from each of the thirteen independent interfering transmitters would result in a confidence measure of 95% ($0.996^{13} = 0.95$) for the resulting interferers [25]. However, a confidence measure of 99.6% for each of the interfering entities implies that the total interference power would vary by 6ζ, i.e., 24 dB. Hence, the total interference power would be between -105.2 dBW and -153.2 dBW for 95% of the cases, i.e., a variation of 18.55% on either side of -129.2 dBW. Therefore, γ, would experience a variation of 25.3% (6.75+18.5) around 10.8 dB (-118.4+129.2). Hence, the system capacity would vary between 2.75 bps/Hz/cell and 4.61 bps/Hz/cell for 90% of the cases, due to lognormal shadowing. It should be noted that the multiplication of the confidence measures for the carrier power (95%) and for the interference power (95%) results in a confidence measure of 90% for the values of γ and the system capacity.

A significant inference that can be derived from the above calculation is that even for a shadowing with a zero mean and 4 dB standard deviation, the upper bound of the system capacity would vary by 25.3% around its mean, for 90% of the cases. If the environment causes a much higher shadowing, the variation in the upper bound of the system capacity would be still higher. This shows the significance of taking the shadowing into considerations while allocating the resources in a wireless network.

Effect on lower bound of capacity:

The lower bound of the capacity in case of downlink communication is 0.21 bps/Hz/cell, achieved for the GTW → MS communication. The corresponding value of the carrier power and total interference power are -118.4 dBW and - 110.3 dBW respectively, resulting in a γ of -8.1 dB. Similar to the upper bound case, a confidence measure of 95% implies a variation of $6\zeta = 24$ dB in the interference power. Hence, the total interference power would be between -86.3 dBW and -134.3 dBW for 95% of the cases, i.e., a variation of 21.75% on either side of -110.3 dBW. Given that the carrier power would experience a variation of 6.75% on either side of -118.4 dB, the value of γ and the system capacity would experience a variation of 28.5% (21.75% + 6.75%) around 0.21 bps/Hz/cell, for 90% of the cases. Hence, for a lognormal shadowing with zero mean and a standard deviation of 4 dB, the lower bound of the system capacity would vary between 0.15 bps/Hz/cell and 0.27 bps/Hz/cell for 90% of the cases. It can be observed that the absolute effect of lognormal shadowing on the lower bound of the capacity value is not a significant issue, since the lower bound of the system capacity is already a very small value.

4.2. Carrier-to-Interference Calculation for Uplink

In the case of an uplink as well, there exist two simultaneously communicating pairs in the cluster-based model: the MS → GTW and the GTW → BS pairs located at the diametrically opposite clusters. In the case of an uplink, both the transmitters in the cluster-based design: the GTW and the MS, are not fixed, whereas, the receiver of one of the communicating pairs, i.e., the BS of GTW → BS pair is located in a fixed position. This results in a slightly modified expression for β in case of uplink, as compared to the downlink scenario:

1) *MS → GTW Communication Pair in the Intended Cell:*

For the MS → GTW pair communication, the expression for β is given by:

$$\beta_{(\mathrm{MS})i}{}^2 = 17 + 4\sqrt{3}(2\cos(\theta_{i2}) - \cos(\phi_{i2})) \\ -4\cos(\phi_{i2} - \theta_{i2}) \quad \mathbf{(24)}$$

$$\beta_{(\mathrm{GTW})i}{}^2 = 14 + 4\sqrt{3}(2\cos(\phi_{i1}) - \cos(\theta_{i2})) \\ -4\cos(\phi_{i1} - \theta_{i2}) \quad \mathbf{(25)}$$

where ϕ_{i2} and θ_{i2} are same, as defined for the BS → GTW communication pair in the semi-analytical model; x_{i2} is the angle made by the GTW_{2b} → MS communicating pair in the i^{th} cell with the reference line of the i^{th} cell; and $\phi_{i1} = x_{i1} + 60(i-1)$, is the GTW → BS communicating pair in the i^{th} cell with the reference line of the i^{th} cell.

2) *GTW → BS Communication Pair in the Intended Cell:*

Similarly, for the GTW → BS communication in the intended cell, the corresponding β values are:

$$\beta_{(\mathrm{MS})_i} = \sqrt{16 + 8\sqrt{3}\cos(\phi_{i2})} \quad \mathbf{(26)}$$

$$\beta_{(\mathrm{GTW})_i} = \sqrt{13 + 4\sqrt{3}\cos(\phi_{i1})} \quad \mathbf{(27)}$$

3) *Capacity Bounds for End-to-End Link:* The lower and the upper capacity bounds for the two communicating pairs in the uplink, GTW → BS pair, and MS → GTW pair can be calculated in a similar manner as is done for downlink. The maximum distance of the inter-cell interferers from the receiver of the GTW → BS pair are 1.732r and 2.732r, resulting in a maximum capacity of 3.32 bps/Hz for the GTW → BS communicating pair. Similarly, the minimum distance of the inter-cell interferers from the receiver of the GTW → BS pair are 0.732r and 1.366r, which results in a minimum capacity value of 0.74 bps/Hz. The minimum and the maximum capacity values for the MS → GTW pair, assuming a constant distance of $r/2$ between the MS and GTW, are 0.88 bps/Hz and 3.71 bps/Hz respectively. Hence, the upper and lower bounds for the system capacity in the uplink scenario are 3.32 bps/Hz/cell and 0.74 bps/Hz/cell respectively.

4) *Effect of Lognormal Shadowing on Capacity Bounds:* Similar to the downlink scenario, the lognormal shadowing causes a variation in the bounds of the system capacity in the uplink scenario as well. The upper bound of the system capacity for the uplink scenario (3.32 bps/Hz/cell) experiences a variation of 25.3% for 90% of the cases in presence of lognormal shadowing of 4 dB standard deviation. Similarly, the lower bound of the system capacity (0.74 bps/Hz/cell) would vary by 27.25% for 90% of the cases.

5. Simulation Model

A simulation model for the cluster-based two-hop cellular network is developed in Matlab. An airport or a campus environment, with a total coverage area of 1 km^2 is considered in the system design. There are 19 cells within a coverage area of 1 km^2. Hence, the distance from the centrally located BS to the edge of the cell, r, is around 130 meters. A propagation model with k_1 = 37 and α = 4 has been considered in the simulation model. 1000 MSs are uniformly distributed around this network coverage area and each cell is designed to have six clusters. All the MSs that are located in the outer layer of the cell are assigned to any one of the six clusters. This assignment is done depending on the closest distance (lowest path loss, in the presence of lognormal shadowing) of the respective MS to the six GTWs in the cell. This results in a system where there are, on an average, eight MSs per cluster. The GTWs are selected from among the MSs. The MSs selected as GTWs are located at nearly half the cell radius. The exact position of the GTW depends on the distribution of the MSs. A TDMA time frame with 16 TSs has been considered in the simulator design. The simulation model calculates the value of γ and the system capacity for seven cells independently and then takes an average over these seven cells. The network is simulated for two different scenarios. In the 1st case, it is assumed that there is no shadowing. Hence, for this case, ζ = 0 in Equation (2). In the 2nd case, a lognormal shadowing with a zero mean and a standard deviation of 4 dB is considered [23].

For different locations of the GTWs and the MSs (with respect to the reference line), the distance of the desired receiver from the interfering transmitters would vary, which in-turn would vary the γ experienced at the receiver of the communicating pair. The synchronized resource reuse technique ensures that all the interfering transmitters are spatially well-separated in distance. The exact value of γ, and thereby the system capacity value, however, depends on the relative distance between the receiver and other transmitting GTWs and MSs. Hence, the system capacity is plotted as cumulative distribution function (cdf) as can be seen from Section 6. The Shannon capacity obtained from the cluster-based two-hop cellular architecture is compared with the following systems:

1) A single-hop cellular network with no relaying:

There are no relays in this design. In every hexagonal cell, the BS and the MS communicate with each other in single hop, irrespective of whether the MS is located in the inner layer or outer layer.

2) A benchmark relaying algorithm for a two-hop cellular network:

The benchmark algorithms for the two-hop cellular design, introduced in [26], provides three efficient methods for finding the wireless terminals that could act as relays in order to maximize the system capacity. These benchmark algorithms could be either distance-based or path loss-based, as explained below. The path loss based algorithms take the random effects, arising due to shadowing, into account. Hence, in the presence of lognormal shadowing, the path loss based algorithms are superior to distance-based algorithms.

5.1. Distance-Based Benchmark Algorithm

In the two-hop design based on benchmark algorithms, the MSs located in the outer layer of the cell communicate to the BS in two hops, as is the case with the cluster-based model. The GTWs/ relay nodes are selected from the mobile nodes available in the network. Suppose, there are *N* possible two-hop routes, between the BS and the MS in the outer layer. Then, the selected route, *rs*, is determined, depending on the transmission distance between the BS and the relay node $d_{c_{n1}}$, and between the relay node and the MS, $d_{c_{n2}}$, for each of the $n \in N$ routes. The three selection schemes of the standard benchmark algorithm for two-hop network [26] are given as follows:

a) shortest total distance (STD) selection scheme:

$$r_s = \min_{\forall n \in N}(d_{c_{n1}} + d_{c_{n2}})$$

b) least longest hop (LLH) selection:

$$r_s = \min_{\forall n \in N}\max(d_{c_{n1}}, d_{c_{n2}}) \quad \text{and}$$

c) shortest relaying hop distance (SRD) selection:

$$r_s = \min_{\forall n \in N}(d_{c_{n2}})$$

5.2. Path Loss-based Benchmark Algorithm

In addition to the distance-based benchmark algorithms, [26] also introduced the path loss-based benchmark algorithms. Let $P_{L_{n1}}$ and $P_{L_{n2}}$ denote the path losses in dB associated with the first hop (BS and relay node), and second hop (relay node and MS of the outer layer) respectively, along the n^{th} route where $n \in N$. Then, the selected route is determined as follows:

a) minimum total path loss (MTP) selection scheme:

$$r_s = \min_{\forall n \in N}(P_{L_{n1}} + P_{L_{n2}})$$

b) least maximum path loss (LMP) selection:

$$r_s = \min_{\forall n \in N}\max(P_{L_{n1}}, P_{L_{n2}}) \quad \text{and}$$

c) minimum relaying hop path loss (MRP) selection:

$$r_s = \min_{\forall n \in N}(P_{L_{n2}})$$

In order to have a fair comparison, the source MSs in case of uplink (or the destination MSs in case of downlink) remain the same in all the methods, viz., the cluster-based two-hop design, the three standard benchmark two-hop schemes, and the single-hop non-relaying technique. Also, an interference avoidance model, with an optimum spatial protection margin of Δ = 1.0, is considered for all the different methods. In addition, it should be noted that, in the simulation model, the increase in the overhead due to additional signaling is not considered in any of the two-hop cellular designs. This increase in the overhead in the two-hop design would cause some reduction in the capacity gain with respect to the single-hop cellular network. However, this paper focuses on the different two-hop schemes, and comparing the performance of the two-hop schemes with the single-hop design is not the main focus of this work. Also, it is expected that, the cluster-based architecture with an intelligent resource allocation technique, would require less or same amount of overhead signaling as compared to the benchmark algorithms, for the two-hop cellular network. Hence, the capacity results obtained in this work for the two-hop networks, viz., the cluster-based design and the three benchmark algorithms are directly comparable. In addition, the MSs in the inner-layer of the cluster-based design are not considered in the simulation, for any of the two-hop methods, as well as for the single-hop design. This is because, in both the cluster-based design and the three standard benchmark algorithms, the wireless terminals located in the inner-layer will communicate with the BS directly in single-hop, as in case of single-hop cellular networks. Hence, all these two-hop methods would use the same amount of radio resource for the inner-layer, and hence, the inner-layer design is not considered in this study.

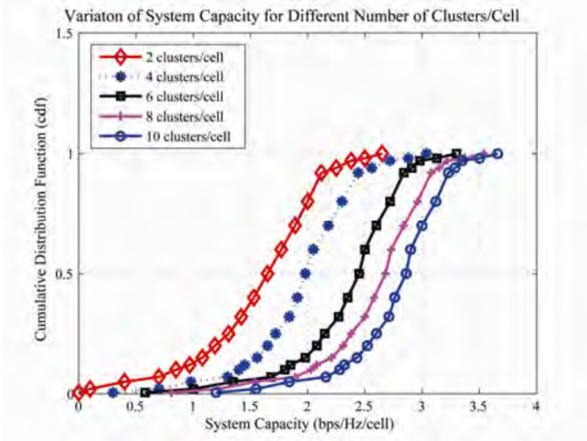


Figure 8. Cumulative distribution function (cdf) of system capacity (average of uplink and downlink) for a two-hop cellular network with different number of clusters/cell.

6. Results

Figure 8 shows the system capacity of the cluster-based two-hop design (average of uplink and downlink results) for different values of clusters per cell. It can be observed that the system capacity shows an increase with an increase in the number of clusters per cell. However, the step size of this increase reduces with an increase in the

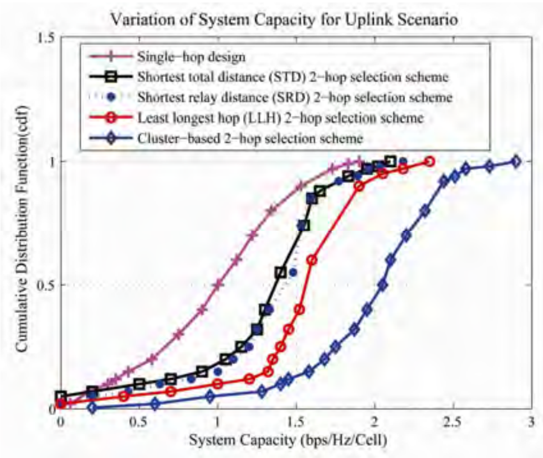


Figure 9. Cumulative distribution function (cdf) of system capacity in the uplink of a multi-cellular network under different transmission schemes (in the absence of lognormal shadowing).

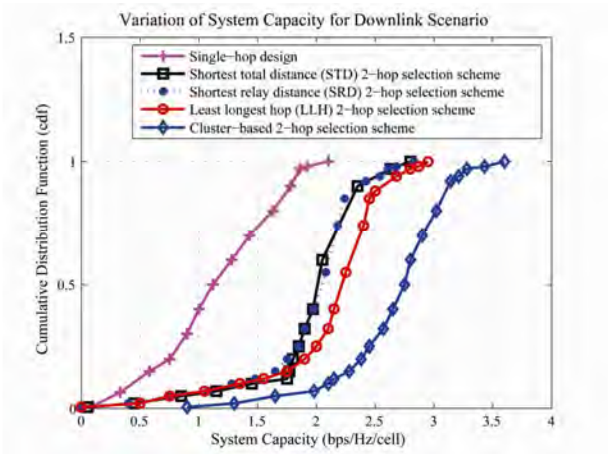


Figure 10. Cumulative distribution function (cdf) of system capacity in the downlink of a multi-cellular network under different transmission schemes (in the absence of lognormal shadowing).

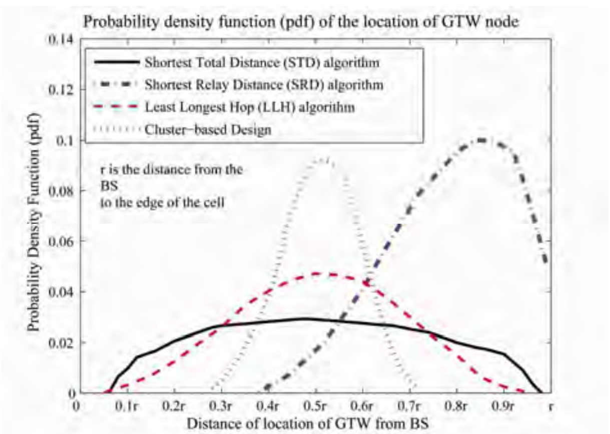


Figure 11. Probability density function (pdf) of the location of GTW for standard benchmark algorithms and the cluster-based design, in the absence of lognormal shadowing.

number of clusters/cell. For example, when the number of clusters per cell is increased from two to four, the expected value of the system capacity is increased from 1.44 bps/Hz/cell to 1.85 bps/Hz/cell, an increase of 0.41 bps/Hz/cell. However, when the number of clusters/cell is increased from four to six, and six to eight, the increase in the expected value of the system capacity is only 0.378 bps/Hz/cell and 0.23 bps/Hz/cell respectively. It should be noted that this increase in the capacity value does not take into account the capacity losses arising due to the increased overhead, as the number of clusters/cell increases. Hence, in all further part of analysis, six clusters/cell are considered, as is also done in [27,28].

Figure 9 and Figure 10 shows the simulation results of the cdf of the system capacity for uplink and downlink scenarios of all different two-hop design methods, in the absence of any lognormal shadowing. The GTWs are selected from among the distributed MSs in the network. Hence, the GTWs are not located at exactly half the cell radius. It is observed in Figure 10 that in case of downlink in the two-hop cluster-based design, the median of the system capacity (2.71 bps/Hz/cell) is 2.5 times that obtained from the single-hop cellular system with no relaying (1.11 bps/Hz/cell). Similarly, the expected value of the cluster-based design is 2.52 bps/Hz/cell, which is more than twice that obtained from the single-hop cellular network value of 1.12 bps/Hz/cell. More significantly, the cluster-based design shows a superior performance over all three standard benchmark techniques for two-hop cellular network. The capacity behavior of the STD and SRD schemes are nearly similar to each other and their expected values are 1.89 bps/Hz/cell and 1.77 bps/Hz/cell respectively. Hence, the expected value of the system capacity in the cluster-based design is 0.63 bps/Hz/cell and 0.75 bps/Hz/cell better than that obtained from STD and SRD algorithms. The LLH technique provides the best performance out of the three standard benchmark techniques. The expected value of the system capacity for the LLH method is 2.05 bps/Hz/cell, which is higher than the expected value of the capacity obtained from the STD and SRD schemes, but less than the expected value obtained for the cluster-based technique by 0.47 bps/Hz/cell. Similarly, in case of an uplink as well, the expected value of the system capacity for the cluster-based design (1.82 bps/Hz/cell) is greater than the LLH method (1.36 bps/Hz/cell) by 0.46 bps/Hz/cell. In the LLH method, the node that has the minimum value among all longest hops (both between, MS and GTW node, and GTW node and BS) among all possible relay nodes is selected as a relay. Hence, a node located in the vicinity of half the cell radius is selected as a relay, which results in more than one pair utilizing the given resource simultaneously, in any cell. It should be noted that this is similar to the cluster-based design introduced in this paper. However, the significant improvement in the system capacity observed in the cluster-based design is due to the synchronized resource reuse technique proposed in this work that ensures a reuse of the radio resource in every cell. The cluster-based design provides

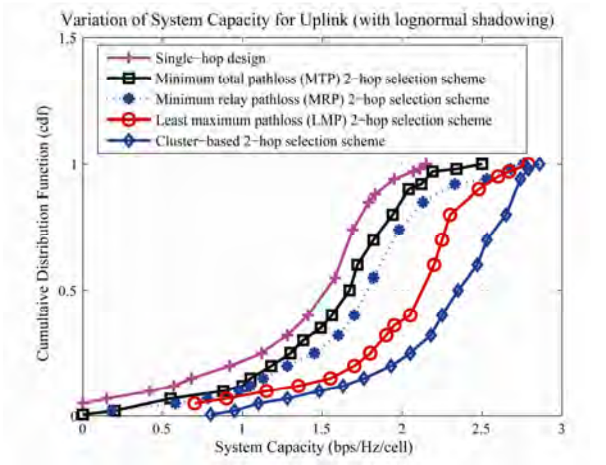


Figure 12. Cumulative distribution function (cdf) of system capacity in uplink, with a lognormal shadowing of zero mean and a standard deviation of 4 dB.

a maximum improvement of 0.8 bps/Hz/cell over the LLH method; and up to 1.4 bps/Hz/cell improvement over the STD algorithm. Figure 11 shows the probability density function (pdf) of the mobile GTWs in the cluster-based design and that of the mobile relay nodes for the benchmark two-hop algorithms. It can be observed from Figure 11 that the pdf of the STD algorithm is almost a straight line, in the range from 0.28*r* to 0.74*r*. The SRD algorithm selects the relay that is more towards the cell edge, than at the center of the cell. Hence, the pdf of the SRD algorithm has a non-zero value only after 0.38*r*. A significant observation that can be made from Figure 11 is that, in case of the LLH method, the mean of the pdf is at 0.5*r*, same as that of the cluster-based design. Hence, the LLH method outperforms STD and SRD benchmark algorithms. However, the variance of the LLH method is 0.28*r*, which is twice more than that of the cluster-based design, which has a variance of 0.13*r*. This implies that in case of LLH method, there is a greater probability of relay nodes being not located in the vicinity of *r*/2, which results in only one pair in the cell being able to utilize the given resource. It should be noted that if the GTWs are selected from among the MS, then the GTW selection would depend on the distribution of MS. But still, the pattern of the GTWs would remain the same for the different benchmark methods and cluster-based design. Hence, the cluster-based two-hop model, with resource reuse in every cell, gives the best performance in terms of system capacity, as compared to the single-hop non-relaying scenario and the benchmark algorithms for the two-hop cellular network.

The performance of the cluster-based design with MSs as GTWs is then compared with the path loss-based benchmark algorithms, in the presence of lognormal shadowing. It is observed from Figure 12 and Figure 13 that, even in the presence of lognormal shadowing, the performance of the cluster-based two-hop network is superior to all three benchmark techniques, for both uplink and downlink. For example, the expected value of the system capacity for the cluster-based design is 2.18 bps/Hz/cell in case of downlink, and is 0.25 bps/Hz/cell better than LMP (the best performing algorithm among all three benchmark algorithms). Similarly, in case of uplink, the expected value of the cluster-based design is 1.98 bps/Hz/cell, and is 0.09 bps/Hz/cell better than the LMP technique. It should however be noted that, in the presence of lognormal shadowing, the performance of the LMP scheme comes close to the performance of the cluster-based design. This is because, the relays are selected not on the basis of distance measurement, but on the basis of path loss measurement, which vary with lognormal shadowing. The presence of lognormal shadowing results in MSs that are far from half the cell radius, *r*/2, to be selected as GTWs. As seen from the pdf of the GTW location in Figure 14, in the presence of lognormal shadowing, there is a non-zero probability for a node located beyond 0.8*r*, to be selected as a GTW. In the absence of lognormal shadowing, the distribution of the GTW is almost symmetric with a mean value of 0.5*r*. However, in the presence of lognormal shadowing, the pdf of the GTW selection exhibits a long tail, resulting in an expected value of 0.58*r*. This results in a situation in the cluster-based design, where the exclusion region of a communicating pair in one cell extends to the other cell, and hence, prevents the simultaneous communication of another pair in the adjacent cell. This, in turn, results in a reduction in the gain in the system capacity.

Figure 15 and Figure 16 compare the results of the cluster-based design independently for uplink and downlink, when fixed GTWs are used instead of mobile GTWs. When there is no lognormal shadowing, the expected value of the system capacity, in case of fixed GTWs, is 2.09 bps/Hz/cell (uplink) and 2.76 bps/Hz/cell (downlink) and is greater than the expected value of that obtained from the mobile GTWs by 0.27 bps/Hz/cell and 0.24 bps/Hz/cell for uplink and downlink respectively. The results observe a similar pattern even in the presence

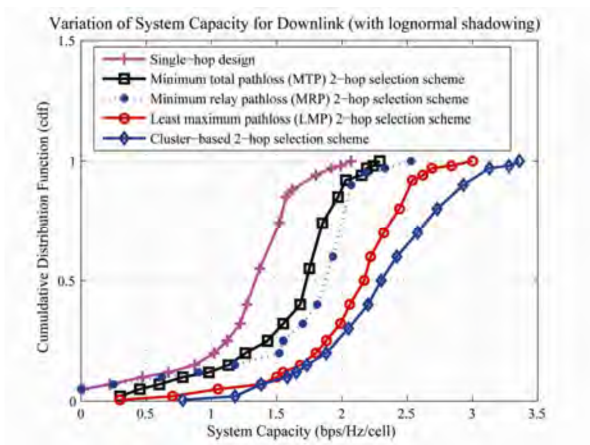


Figure 13. Cumulative distribution function (cdf) of system capacity in downlink, with a lognormal shadowing of zero mean and a standard deviation of 4 dB.

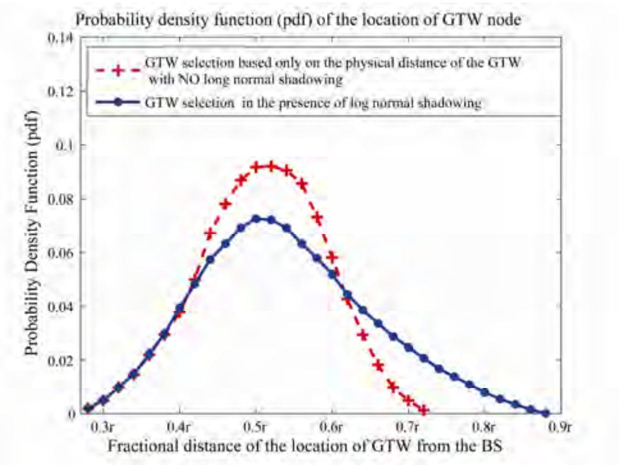


Figure 14. Probability density function (pdf) of the location of GTW in a cell in a multi-cellular network.

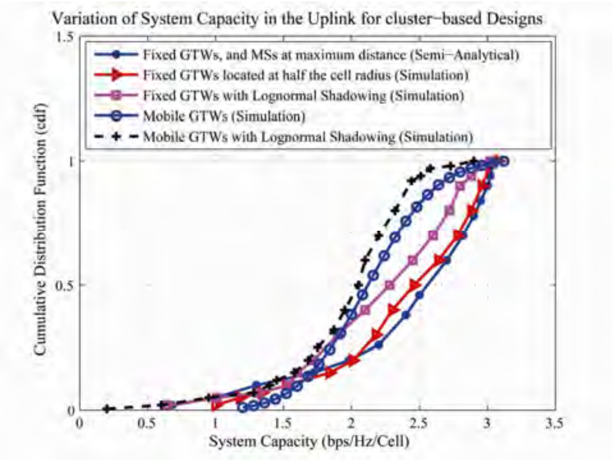


Figure 15. Cumulative distribution function (cdf) of system capacity for different cluster-based scenarios, in case of uplink transmission.

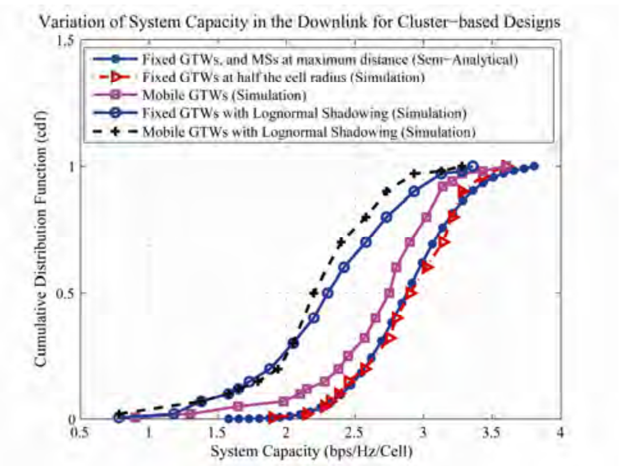


Figure 16. Cumulative distribution function (cdf) of system capacity for different cluster-based scenarios, in case of downlink transmission.

of lognormal shadowing. For example, the expected value of the system capacity in case of uplink is 2.24 bps/Hz/cell for fixed GTWs, and is greater than that of the expected value of the system capacity for mobile GTWs (1.98 bps/Hz/cell) by 0.26 bps/Hz/cell. In addition, it can be observed from Figure 15 (uplink) and Figure 16 (downlink), that the cdf obtained from the semi-analytical model shows a close match with that obtained from the simulation results for cluster-based design with fixed GTWs. The expected values of the system capacity for the semi-analytical model are 2.16 bps/Hz/cell (uplink) bps/Hz/cell for fixed GTWs, and is greater than that of the expected value of the system capacity for mobile GTWs (1.98 bps/Hz/cell) by 0.26 bps/Hz/cell. In addition, it can be observed from Figure 15 (uplink) and Figure 16 (downlink), that the cdf obtained from the semi-analytical model shows a close match with that obtained from the simulation results for cluster-based design with fixed GTWs. The expected values of the system capacity for the semi-analytical model are 2.16 bps/Hz/cell (uplink) and 2.69 bps/Hz/cell (downlink), and is very close to that obtained from the simulation results for fixed GTWs: 2.09 bps/Hz/cell for uplink and 2.76 bps/Hz/cell for downlink respectively. This is primarily because, in case of semi-analytical model, not only all the GTWs are fixed at a distance of *r/*2 from the center of the cell, but also all the MSs in the cluster are at the maximum distance of *r/*2 from the cluster-head GTWs, whereas in the simulation model for fixed GTWs, only the GTWs are fixed at a distance of *r/*2. Hence, the distance between the cluster-head GTW and the MSs could be any value less than or equal to *r/*2.

In a significant observation, the upper and lower bounds for the system capacity obtained from the simulation results is quite close to those obtained from the semi-analytical model where all the MSs were assumed to be distributed at a location of *r/*2 from the GTW. In downlink mode, the lower bound for capacity under the semi-analytical model is 0.21 bps/Hz/cell whereas that obtained from the simulations is 0.22 bps/Hz/cell. The upper bound for the system capacity under the semi-analytical model is 3.68 bps/Hz/cell whereas that obtained from the simulations is 2.91 bps/Hz/cell. Similarly, in the uplink mode, the lower and upper bound for the system capacity under the semi-analytical model is 0.74 bps/Hz/cell and 3.32 bps/Hz/cell respectively; whereas that obtained from the simulations is 0.82 bps/Hz/cell and 2.91 bps/Hz/cell respectively. Hence, it can be concluded that the lower and upper bound results obtained from the simulation model closely match with the results of the semi-analytical model, thereby validating the performance of cluster-based two-hop model.

7. Summary and Conclusions

In this paper, a novel resource allocation mechanism has been proposed for a two-hop cellular network. The new scheme, known as *cluster-based architecture*, is designed using a synchronized resource reuse technique. As per this design, each hexagonal cell is divided into two layers, the inner layer and outer layer. All MSs in the inner layer communicate with the BS in single hop, whereas, the MSs in the outer layer communicate with the BS in two hops, using a GTW terminal as relay node. This architecture design, developed under an interference avoidance *Protocol Model*, results in a frequency reuse ratio of *one*, whereby, the given resource is used twice in every cell in the system, but for only half the duration of a time slot period. This work, first, shows that the system spectral efficiency of a cellular network can be increased significantly by allowing two-hop communication. Second, it has been found that a hierarchical, co-ordinated approach which essentially means to limit the degrees of freedom for forming the two-hop links does not lower the capacity, but in fact gives higher capacity than compared to state-of-the-art two-hop algorithms. This means that the complexity of the routing problem in such two-hop communication systems can be significantly reduced while the system performance does not have to be compromised.

8. References

[1] H. Wu, C. Qao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems," IEEE Journal on Selected Areas in Communication, Vol. 19, No. 10, pp. 2105–2115, October 2001.

[2] A. Zemlinanov and G. Veciana, "Capacity of ad hoc wireless networks with infrastructure support," IEEE Journal on Selected Areas in Communications, Vol. 23, No. 3, pp. 657–667, March 2005.

[3] H. Venkataraman, H. Haas, S. Yun, Y. Lee, and S. McLaughlin, "Performance analysis of hybrid wireless networks," in Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC'05), Berlin, Germany, Vol. 3, pp. 1742–1746, 11–14 September 2005.

[4] H. W. Chen and M. Guizani, "Next generation wireless systems and networks," in John Wiley & Sons, 2006.

[5] O. Dousse, P. Thiran, and M. Hasler, "Connectivity in ad hoc and hybrid networks," in Proceedings of IEEE International Conference on Computer Communications (INFOCOM), Vol. 2, pp. 1079–1088, 23–27 June 2002.

[6] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in Proceedings of IEEE International Conference on Computer Communications (INFOCOM), San Francisco, USA, Vol. 2, pp. 1543–1552, 30 March–3 April 2003.

[7] H. Vishwanathan and S. Mukherjee, "Performance of cellular networks with relays and centralized scheduling," in Proceedings of IEEE Vehicular Technology Conference (VTC Fall'03), Orlando, Florida, USA, Vol. 3, pp. 1923–1928, 6–9 October 2003.

[8] H. Haas and S. McLaughlin, Eds., "Next generation mobile access technologies: Implementing TDD," Cambridge University Press, ISBN: 13: 9780521826228, pp. 420, January 2008.

[9] X. Wu, S. H. G. Chan, and B. Mukherjee, "MADF: A novel approach to add an adhoc overlay on a fixed cellular infrastructure," in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'00), Chicago, USA, pp. 549–554, 23–28 September 2000.

[10] Y. D. Lin and Y. Hsu, "Multihop cellular: A new architecture for wireless communications," in Proceedings of IEEE International Conference on Computer Communications (INFOCOM) Tel Aviv, Israel, pp. 1273–1282, 26–30 March 2000.

[11] R. Ananthapadmanabha, B. S. Manoj, and C. S. R. Murthy, "Multihop cellular networks: The architecture and routing protocol," in Proceedings of IEEE International Symposium on Personal Indoor Mobile Radio Communications (PIMRC'01), San Diego, USA, Vol. 2, pp. 78–82, 30 September–3 October 2001.

[12] H. Li, D. Yu, and H. Chen, "New approach to multihop - cellular based multihop network," in Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC'03), Beijing, China, Vol. 2, pp. 1629–1633, 7–11 September 2003.

[13] Y. D. Lin, Y. C. Hsu, K. W. Oyang, T. C. Tsai, and D. S. Yang, "Multihop wireless IEEE 802.11 LANs: A prototype implementation," Journal of Communications and Networks, Vol. 2, No. 4, December 2000.

[14] V. Sekar, B. S. Manoj, and C. S. R. Murthy, "Routing for a single interface MCN architecture and pricing schemes for data traffic in multi-hop cellular networks," in Proceedings of IEEE International Conference on Communications (ICC'03), Alaska, USA, Vol. 2, pp. 969–973, 11–15 May 2003.

[15] Y. Liu, R. Hoshyar, X. Yang, and R. Tafazolli, "Integrated radio resource allocation for multihop cellular networks with fixed relay stations," IEEE Journal on Selected Areas in Communications, Vol. 24, No. 11, pp. 2137–2146, November 2006.

[16] S. Mertello and P. Torth, "Knapsack Problems, Algorithms and Computer Implementation," NewYork: Wiley, 1990.

[17] P. Gupta and P. R. Kumar, "The capacity of wireless networks," IEEE Transactions on Information Theory, Vol. 46, No. 2, pp. 388–404, February 2000.

[18] H. Venkataraman, S. Sinanovic, and H. Haas, "Variation of spatial protection margin in multihop wireless networks," in Proceedings of IEEE International Symposium on Personal Indoor Mobile Radio Communications (PIMRC'07), Athens, Greece, 3–6 September 2007.

[19] H. Venkataraman, S. Nainwal, and P. Shrivastava, "Optimum number of gateways in cluster-based two-hop cellular networks," AEU Journal of Electronics and Communications, Submitted and under review, http://nextgenwireless.daiict.ac.in/Sub_Papers/AEU_May_2008_Hrishikesh_Venkataraman.pdf, 2008.

[20] S. D. Z. Dawy and I. Oikonomidis, "Coverage and capacity enhancement of CDMA cellular systems via multihop transmission," in Proceedings of IEEE Global Communication Conference (GLOBECOM), San Fransisco, USA, pp. 1147–1151, 1–5 December 2003.

[21] 3rd Generation Partnership Project (3GPP), Technical Specification Group Radio Access Network, "Selection procedures for the choice of radio transmission technologies

of the UMTS," 3GPP TR 30.03U, May 1998. Retrieved November 30, 2006, from: http://www.3gpp.org/ftp/Specs/htmlinfo/3003U.htm.

[22] J. Cho and Z. Haas, "On the throughput enhancement of the downstream channel in cellular radio networks through multihop relaying," IEEE Journal on Selected Areas in Communications, Vol. 2, No. 7, pp. 1206–1219, September 2004.

[23] IST-2003-507581 WINNER, "D7.6 winner system concept description," Retrieved July 15, 2007, from https://www.istwinner.org/DeliverableDocuments, October 2005.

[24] R. A. Fisher, "Statistical methods and scientific inference," in Oliver and Boyd, Edinburgh, 1956.

[25] I. Hacking, "Logic of statistical inference," in Cambridge University Press, Cambridge, 1965.

[26] V. Sreng, H. Yanikomeroglu, and D. D. Falconer, "Relay selection strategies in cellular networks with peer-to-peer relaying," in Proceedings of IEEE Vehicular Technology Conference (VTC Fall'03), Orlando, USA, Vol. 3, pp. 1949–1953, 4–9 October 2003.

[27] P. li, M. Rong, T. Liu, and D. Yu, "Interference modeling and analysis in two-hop cellular network with fixed relays in FDD mode," in Proceedings of IEEE Wireless Communications, Networking and Mobile Computing (WiMob), Montreal, Canada, pp. 177–182, 22–24 August 2005.

[28] T. Liu, M. Rong, H. Shi, D. Yu, Y. Xue, and E. Schulz, "Reuse partitioning in fixed two-hop cellular relaying network," in Proceedings of IEEE Wireless Communications Networking Conference (WCNC), Las Vegas, USA, pp. 177–182, 3–6 April 2006.

I. J. Communications, Network and System Sciences, **2008, 4, 285-385**
Published Online August 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

TABLE OF CONTENTS

Volume 1 **November 2008**

ISSN: 1913-3715
9771913371005 04