Mobile Networks

Edited by Jesus Hamilton Ortiz

MOBILE NETWORKS

Edited by **Jesús Hamilton Ortiz**

Mobile Networks
Edited by Jesús Hamilton Ortiz

Published by InTech
Janeza Trdine 9, 51000 Rijeka, Croatia

Publishing Process Manager Jana Sertic
Technical Editor Teodora Smiljanic
Cover Designer InTech Design Team

First published April, 2012
Printed in Croatia

A free online edition of this book is available at www.intechopen.com
Additional hard copies can be obtained from orders@intechopen.com

Mobile Networks, Edited by Jesús Hamilton Ortiz
p. cm.
ISBN 978-953-51-0593-0

Contents

Preface

The growth in the use of mobile networks has come mainly with the third generation systems and voice traffic. With the current third generation and the arrival of the 4G, the number of mobile users in the world will exceed the number of landlines users. Audio and video streaming have had a significant increase, parallel to the requirements of bandwidth and quality of service demanded by those applications. Mobile networks require that the applications and protocols that have worked successfully in fixed networks can be used with the same level of quality in mobile scenarios.

One of the main differences between fixed and mobile networks lies in the dynamic nature of the latter. The constant movement of mobile devices has a clear impact in the quality of service that can be achieved (delay or loss of packets during a handover from one cell to another). The migration of mechanisms initially meant for fixed networks to mobile networks may cause problems related to topology and mobility factors. Other difficulties may appear when we want to move mechanisms designed for infrastructure and wired networks to ad-hoc or mobile networks in general. These are some of the drawbacks:

Problems related to topology:

One of the great remaining difficulties from the first generation to the fourth generation of mobile devices occurs when there is a handover, either from one cell to another or from one access network to another. This circumstance clearly affects the quality of service in diverse ways: delay of packet transfers, increase of the jitter of audio and video streaming or even damage or loss of packets. There are different types of handovers that produces diverse signalling loads in the access network. A handover involves a route variation in order to reach the mobile terminal. To provide a good level of QoS in mobile environments, a minimal handover delay is always welcome to ensure the smallest traffic interruption during a transfer.

Problems related to mobility: macromobility and micromobility.

- Macromobility: Mobile terminal activity between different access networks or domains (inter-domain).
- Micromobility: Mobile terminal activity inside one access network only (intra-domain).

Although the two types of handovers occur under both circumstances, intra-domain handovers will be a priority due to their higher frequency of signalling load and packet transfers. One of the greatest difficulties in reducing the mobility impact in the terminals when there is a handover is that the protocols or mechanisms to provide quality of service are designed and limited to a certain kind of fixed or mobile networks or at macromobility level. Using these existing mechanisms of QoS involves adapting the dynamic characteristics of the mobile devices. There are cases such as ad-hoc networks that have special mobility specifications, making migration a complex challenge.

Until the third generation of mobile networks, the need to ensure reliable handovers was still an important issue. On the eve of a new generation of access networks (4G) and increased connectivity between networks of different characteristics commonly called hybrid (satellite, ad-hoc, sensors, wired, WIMAX, LAN, etc.), it is necessary to transfer mechanisms of mobility to future generations of networks. In order to achieve this, it is essential to carry out a comprehensive evaluation of the performance of current protocols and the diverse topologies to suit the new mobility conditions.

Dr Jesús Hamilton Ortiz
School of Computer Engineering,
University of Castilla La Mancha, Ciudad Real
Spain

1

Mechamisms to Provide Quality of Service on 4G New Generation Networks

Jesús Hamilton Ortiz, Bazil Taja Ahmed,
David Santibáñez and Alejandro Ortiz
University of Castilla y la Mancha
Spain

1. Introduction

1.1 New generation nerworks (4G)

Currently, the 3rd Generation Partnership Project forum (3GPP) is working to complete the standard that aims to ensure the competitiveness of UMTS in the future. As a result of this work, in 2004 the Long Term Evolution project (LTE) arises, which is expected to become the 4G standard. We can find the requirements for 4G standardization in recent works like "Release 10" and "Advanced LTE".

On the other hand, the System Architecture Evolution (SAE) is a project that seeks to define a new core component of the all-IP network called Evolved Packet Core (EPC). We can consider IPv6/MPLS as part of the development of the LTE standard included in the all-IP concept to meet some requirements of LTE, such as end-to-end quality of service (MPLS, Diffserv, IntServ). SAE allows interoperability with existing technologies in both the core and access networks.

The figure 1 shows the relation between 2G, 3G & LTE technologies and the packet core that is intended to evolve with SAE.

Due to the increasing demand of QoS by users, it is necessary to adopt mechanisms to ensure the requirements of LTE/SAE. As is well known, an all-IP network provides the so-called Best Effort quality of service. For this reason, in order to provide QoS to the LTE/SAE network's core and to the access networks, we propose the implementation of IPv6 (extensions/MPLS into the Evolved Packet Core (EPC).

1.2 Requirements of LTE/SAE

Some of the most important requirements of LTE/SAE are:

- Low cost per bit.
- Increase of the services provided: more services at lower cost to improve the user's experience.
- Flexible use of existing and new frequency bands.
- Simplified architecture.

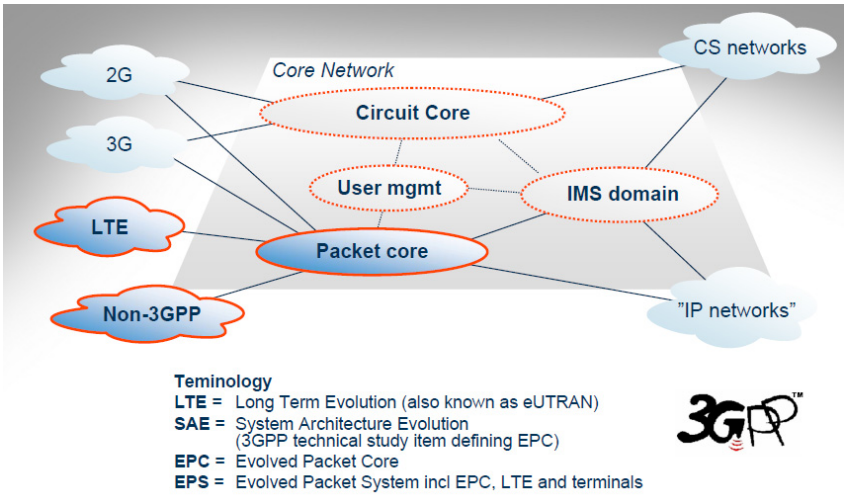


Fig. 1. LTE/SAE

- Reasonable energy consumption.

In addition to the requirements mentioned above, there would be other important requirements as part of the standard, such as throughput optimising, latency reduction and end-to-end QoS for both the core and the access networks. In order to improve these conditions, we have considered the handover a priority. One of the key elements in the all-IP concept is the MPLS protocol as a fundamental part of all IPv6/MPLS architecture to provide quality of service to access networks and core network since it will be compatible with other architectures in the next generation mobile networks.

1.3 MPLS

In 1996, companies like Nokia, Cisco, IBM and Toshiba, among others, introduced proprietary solutions to the problem of multilayer switching. This was not only a solution to integrate ATM with IP, but offered brand new services. Unfortunately, these solutions were not compatible despite the large number of aspects in common. MPLS (Multi-Protocol Label Switch) came up from the work of the IETF in 1997 to standardise the proprietary multilayer switching technologies mentioned above. The main feature of MPLS is the combination of layer 3 routing and the simplicity of level 2 switching.

Another important feature of MPLS is that it provides a good balance between connection-oriented technologies to improve non-IP connection-oriented mechanisms (they can only deliver a Best Effort level of service). On the other hand, MPLS adds labels to the packets, so no routing is based on layer 3 addresses but in label switching. This allows interoperability between IP and ATM networks. It also increases the speed of the packets traversing the network because they do not run complex routing algorithms at every hop; they are forwarded considering the packet's label only. This labelling system is also very useful to classify the incoming traffic according to its higher or lower QoS requirements contracted or required.

Since MPLS is a standard solution, it also reduces the operational complexity between IP networks and gives IP advanced, routing capabilities in order to use traffic-engineering techniques that were only possible on ATM.

1.4 IPv6 extensions

The extensions of the IPv6 protocol were designed to migrate IPv6 to mobile environments. There are several extensions of IPv6 designed with this purpose. We have chosen the following IPv6 extensions: HMIPv6, FHMIPv6 and FHAHMIPv6, The first and second extensions were designed to be used at micro-level mobility, because the signalling, lad at this level is higher, With regards to the FHAMIPv6 protocol, this is a protocol that we have designed to provide hierarchical addresses support in an ad hoc network.

1.5 IPv6/MPLS on LTE/SAE

So far, we have briefly described what LTE/SAE consists of, the current requirements that have to be met to become the 4G standard and the most relevant concepts related to MPLS. Let us now look into the importance of supporting the LTE/SAE core with IP/MPLS.

The use of MPLS on LTE allows reusing much of 2G and 3G technologies, which means a low cost per bit. In addition, MPLS can handle the IP requirements for the wide range of services it supports. MPLS also supports any topology, including star, tree and mesh. On the other hand, IPv6/MPLS can give IP advanced traffic engineering, ensuring that traffic is properly prioritised according to its characteristics (voice, data, video, etc.) and the routes through the network are set up to prevent link failures. The use of differentiated services is also an important feature of MPLS, since Forwarding Equivalent Class (FEC) can perform different treatments to the services provided by IP, including an eventual integration with Diffserv. This contributes to provide a better quality of service (QoS).

In addition, because MPLS creates virtual circuits before starting the data transmission and uses special labelling, it is possible to deliver a better level of security when packets experience higher rates of transmission and processing, since the forwarding is performed according to the label without routing algorithms. This is another important aspect of IPv6/MPLS in order to meet the requirements related to the throughput. Finally, MPLS promotes the simplification of the integration architecture of IP and ATM and improves the users' QoS experience providing redundant paths to different FECs to prevent packet loss. The following figure shows how the transition to IPv6/MPLS will be as part of LTE.

Service providers and network operators want to ensure that their Radio Access Network (RAN) is able to support current technologies such as GSM and UMTS and new technologies such as LTE and WIMAX. At the same time, future broadband requirements must be met in an efficient and effective way. That is why service providers are choosing solutions based on IPv6/MPLS. This technology can fulfil current and future needs while reducing costs.

It is important to point out that the standard WIMAX and advanced WIMAX or mobile WIMAX, which is part of the evolution of IEEE (802.11, 802.16, etc.), complies with the requirements for 4G standard. WIMAX (802.16) can operate in both the core and access networks with IPv6/MPLS.

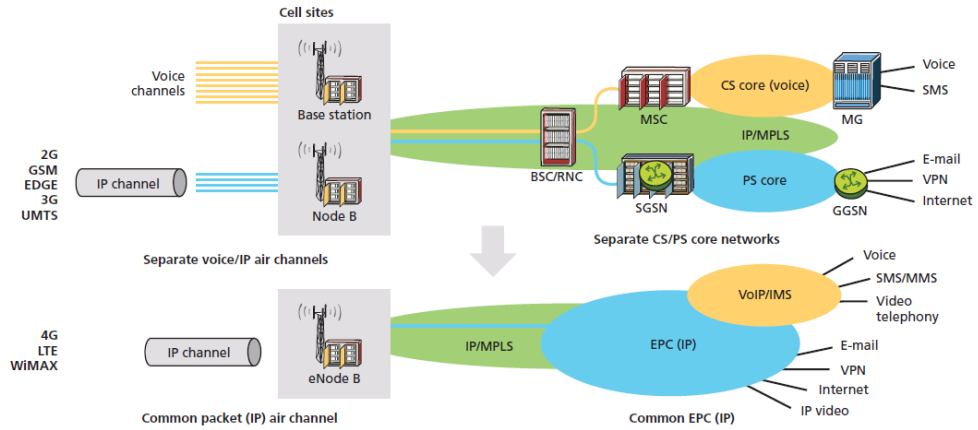


Fig. 2. LTE to IP/MPLS and EPC

Currently, there is competition for the dominant 4G standard. Advanced LTE has a higher market share than advanced WIMAX because it is part of the evolution of GSM and UMTS networks and represents 80% of the worldwide market. However, WIMAX today has a significant market share in the United States. We believe that both LTE and WIMAX meet standard requirements and are compatible with the architectures proposed for an all IPv6/MPLS approach both in access networks as the core of the network.

This chapter is focusing in the integration of mobility protocol (IPv6 extensions) and the protocol of quality of services (MPLS). The RSVP protocol has been used as signalization protocol. The metrics of quality of services tested are: Delay, jitter, throughput, the send and received packets, these metrics were chosen because they are the most sensitives in a handover. The integrations tested in this chapter were: HMIPv6/MPLS, FHM IPv6/MPLS, FHAMIPv6/AODV and FHAMIPv6/MPLS. In order to achieve these integration was necessary modify the source codes and adapt the simulator versions (NS-2). In order to integrated protocols performance as a new protocol.

2. HMIPv6/MPLS integration

In response to the demands of multimedia services on existing mobile systems, cellular areas will have a smaller radius in order to support higher throughput, ensuring acceptable error rates. Having small cells means that the MN can cross borders more frequently and signalling capacity will increase rapidly. In this section, we will integrate HMIPv6 and MPLS. The architecture used is the proposal by Robert Hsieh. We use us the scenario base (R. Hsieh) and then increasing the number nodes and flow of traffic in order to analyse the scalability of this integration. We analyse the relationship between the different metrics and the number nodes, the main idea is that in an handover the metrics of quality of service will be optimized or by default it's were not degraded. The metrics used were: delay, jitter, send and received packets and throughput. This metrics were chosen because they are most sensitive in a handover.

In other work, was evaluating the HMIPv6/MPLS integration, this works were tested in different scenarios [2,],[8],[9],[10] the integrations were HMIPv6/MPLS/RSVP and the

simulation scenario was made in a LAN and WAN networks. In these integrations, the RSVP protocol was used as signalling protocol while hierarchical MPLS nodes were used to achieve interoperability of HMIPv6 and MPLS.

The results obtained in [2],[8],[9],[10] showed that this interoperability is a good alternative to provide QoS in LAN and WLAN networks. In order to better the load signalisation in a handover, in case of Binding Update the HMIPv6/MPLS was used as preliminary work with the idea of future integration FHMIPv6/MPLS and FHMIPv6/MPLS in Ad hoc networks.

2.1 HMIPv6/MPLS integration in a scenario with CBR

2.1.1 Simulation scenario

The scenario simulated is shown (R. Hsieh) in figure 3. The MN is in the area of HA. The traffic used was CBR because is most sensitive in audio/video application. The Bandwidth configuration and delay of each link go as follows:

Link	Delay	Bandwidth
CN-LSR1	2ms	100Mb
LSR1-HA	2ms	100Mb
LSR1-MAP	50ms	100Mb
MAP-LSR2	2ms	10Mb
MAP-LSR3	2ms	10Mb
LSR2-PAR	2ms	1Mb
LSR3-NAR	2ms	1Mb

Table 1. Simulation scenarios

The traffic used was CBR, since it allows audio and video simulation in real time. These applications have a high demand of QoS.

The figure 3 shows the topology of the simulated network. MPLS is the core of the network and is constituted by the following nodes: 1 (MAP), 2 (LSR1), 3 (LSR2), 4 (LSR3), 7 (LER1 for MPLS and PAR for HMIPv6) and 8 (LER2 for MPLS and NAR for HMIPv6); the tag distribution protocol used by MPLS is RSVP. Finally number 6 is the MN.

Every link shows two of their characteristics: bandwidth (in megabits or kilobits) and delay (in milliseconds).

The figure 3 shows the topology of the simulated network. MPLS is the core of the network and is constituted by the following nodes: 1 (MAP), 2 (LSR1), 3 (LSR2), 4 (LSR3), 7 (LER1 for MPLS and PAR for HMIPv6) and 8 (LER2 for MPLS and NAR for HMIPv6); the tag distribution protocol used by MPLS is RSVP. Finally number 6 is the MN.

Every link shows two of their characteristics: bandwidth (in megabits or kilobits) and delay (in milliseconds).

A few seconds later MN moves toward area PAR/LER as the figure 4 illustrate, finally the MN moves to area NAR/LER as the figure 5 illustrates.

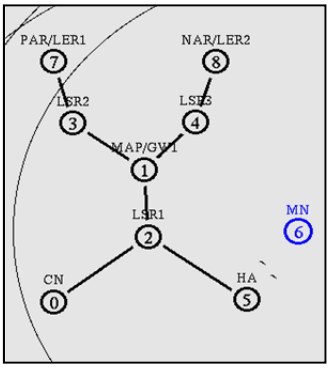


Fig. 3. Scenario of HMIPv6/MPLS simulation

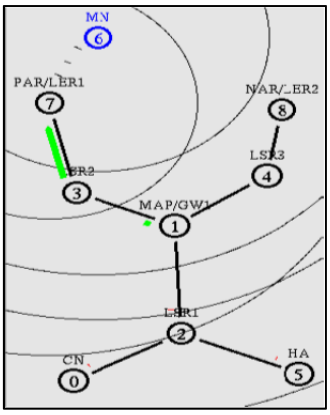


Fig. 4. MN moves the area PAR/LER

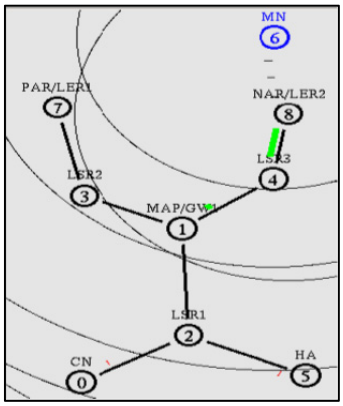


Fig. 5. MN moves the area NAR/LER

Finally, the MN moves to area NAR/LER as the figure 5 illustrates.

2.1.2 Description of simulation

Initially, the MN is located in the area of the HA. 2 seconds after the start of the simulation, the HA moves towards the area of the PAR at 100 m/s, arriving at t=3.5 s approximately. At t=5 s, the CN begins sending CBR traffic to the MN following the route CN→LSR1→HA→LSR1→MAP→LSR2→PAR-MN as shown in figure 3. Then, at t=10 s, the MN starts moving to the area of the NAR at 10 m/s. At the same time, the handover takes places at around t=13.12 s and the MN receives one of the first packets from the NAR. Afterwards, the MN places in the area of the NAR at around t=17 s. Finally, at t=19 s, the CN stops sending traffic flow towards the MN.

Simulation scenarios

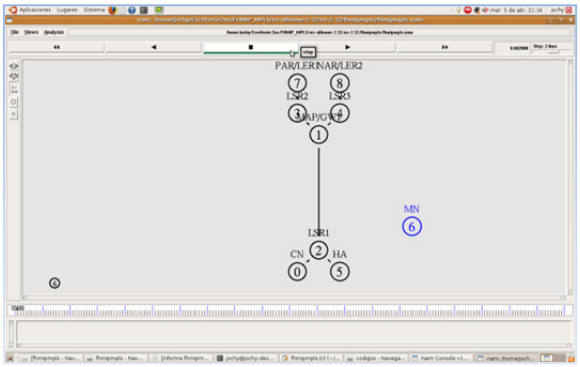


Fig. 6. Scenario with 9 nodes Figure

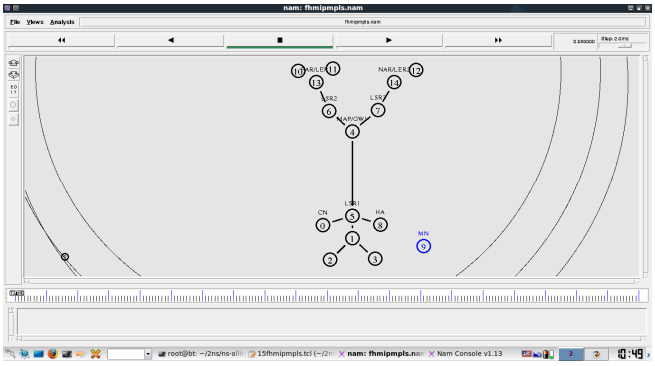


Fig. 7. Scenario with 15

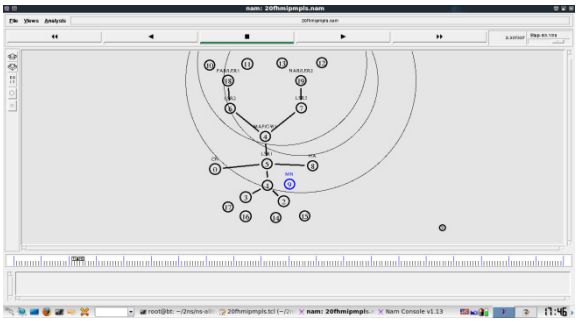


Fig. 8. Scenario with 20 nodes

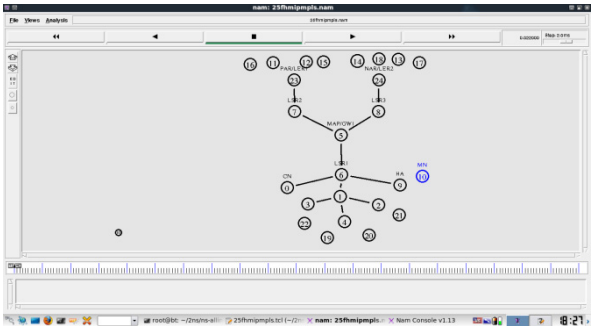


Fig. 9. Scenario with 25 nodes

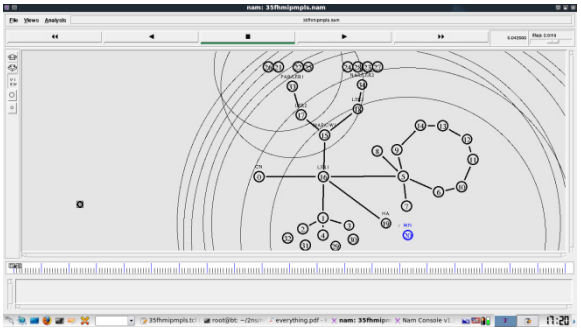


Fig. 10. Scenario with 30 nodes

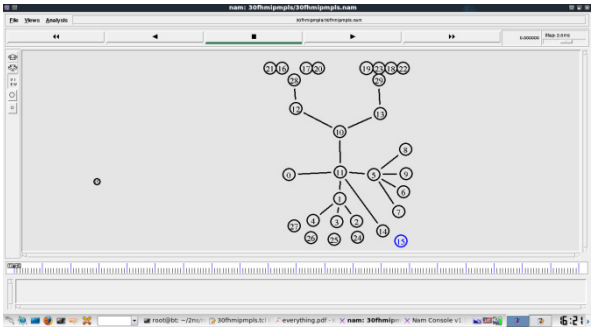


Fig. 11. Scenario with 35 nodes

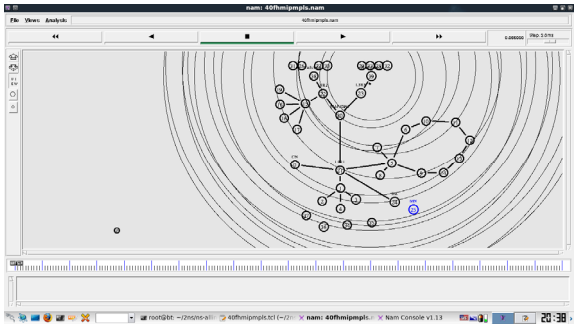


Fig. 12. Scenario with 40 nodes.

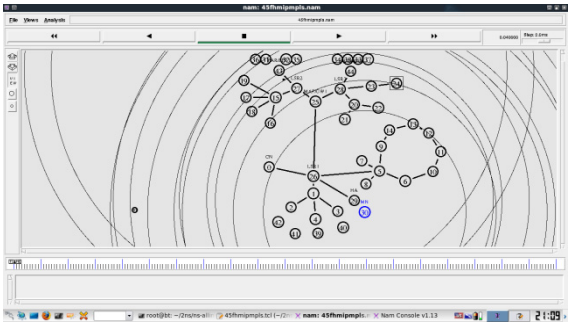


Fig. 13. Scenario with 45 nodes.

2.2 Scalability

The objective of this simulation with different scenarios was to analyse QoS metrics in HMIPv6/MPLS integration with CBR traffic and the scalability. The table2 show the different scenarios simulated. The first scenario was proposal by R.Hsieh, the other scenarios were increasing the number nodes in order to test the scalability, the table show the results of different metrics analysed.

Scenario\ Metrics	Delay(ms)	Jitter(ms)	Throughput (Kbps)	Sends Packets	Received Packets	Lost Packets (%)
9 Nodes	66,85	0,46	424,80	3734	3734	0%
15 Nodes	226,66	1,97	356,11	3734	3151	15,61%
20 Nodes	254,70	1,84	359,60	3734	3204	14,20%
25 Nodes	317,92	4,0	277,60	3734	2476	33,70%
30 Nodes	310,73	3,60	288,27	3734	2571	31,15%
35 Nodes	312,95	3,80	279,96	3734	2497	33,13%
40 Nodes	331,80	4,31	268,22	3734	2392	35,94%
45 Nodes	341,70	4,60	261,90	3467	2156	37,81%

Table 2. Results of different scenarios HMIPv6/MPLS integration

The table 2 shows the results of HMIPv6/MPLS integration. The metrics analysed were: delay, jitter, throughput, sent packets, received packets and lost packets. From the results obtained, we can affirm that, in general, the delay increases as the number of nodes increases. The jitter grows significantly when there are more than 25 nodes. The throughput shows a slight variation, but it does not follow a particular pattern. Sent packets, normally, remain constant; the received packets, generally, decrease as the number of nodes grows and the number of lost packets increases significantly when there are more than 25 nodes.

The figure 14 shows the results of the following metrics obtained of the table 2. In this manner can visualize the behaviour of delay, throughput, send and received packets against the quantity of number of nodes.

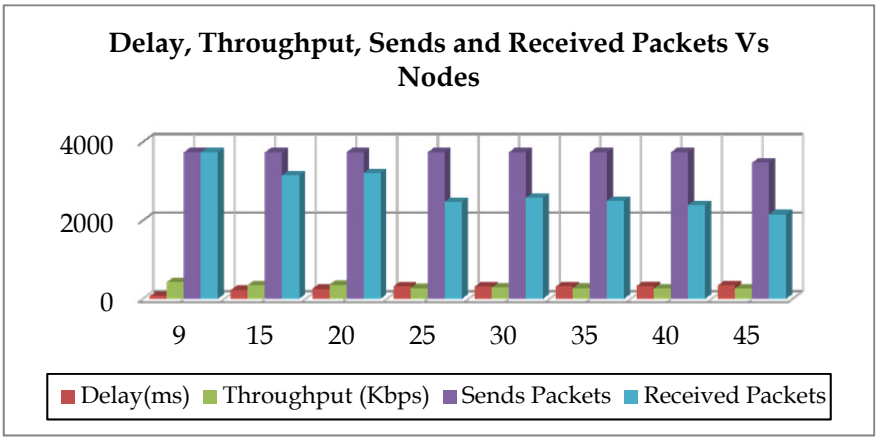


Fig. 14. Delay, throughput, send an received packets Vs. Number nodes

In order to extend the different results obtained in the simulations, the functions (figure15) show the behaviour of the different simulation scenarios. With these functions we could predict what will happen with the metrics (Delay, Throughput, Send and Received Packets) against the number of nodes. In this manner, we could predict what happen when the number of nodes and flow are traffic is increased.

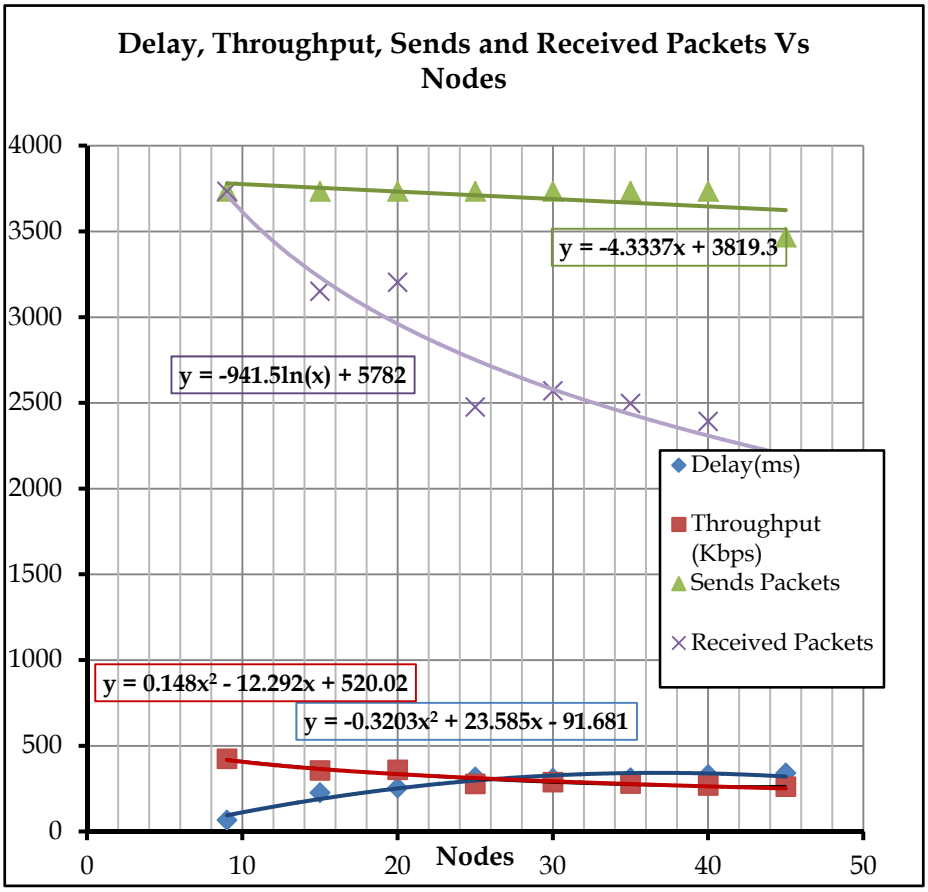


Fig. 15. The functions show the scalability of Delay, Throughput, send and received packets.

The figure 16 shows the results of the following metrics obtained of the table 2. In this manner can visualize the behaviour of Jitter and Lost Packets with different number nodes.

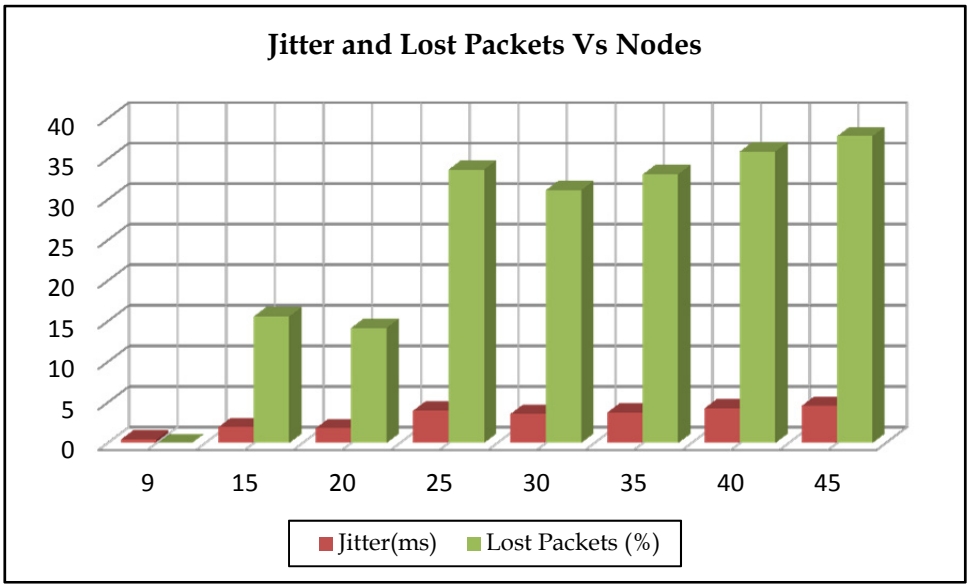


Fig. 16. Jitter and Lost Packets vs. Number nodes

In order to extend the different results obtained in the simulations, the function (figure17) shows the behaviour for different scenarios of simulation. With this functions (Jitter, Lost

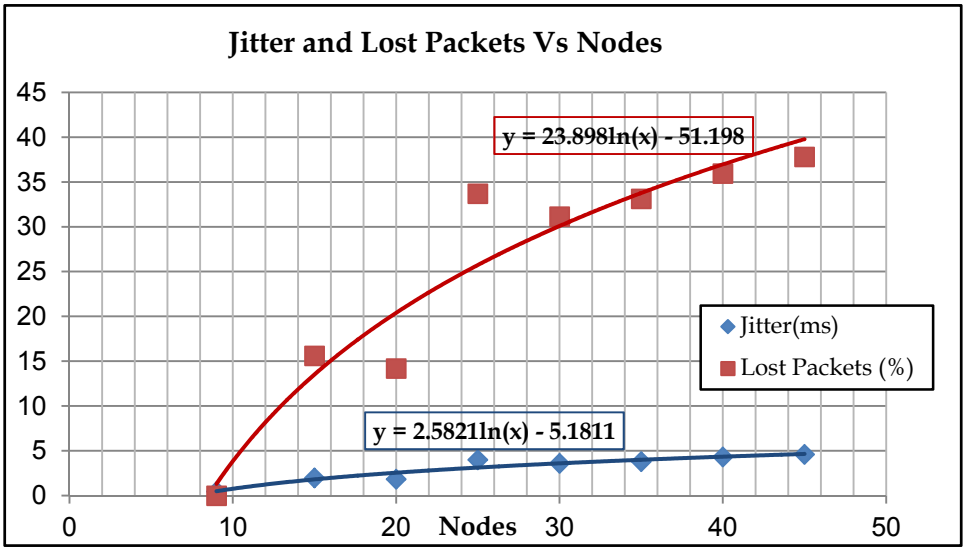


Fig. 17. The functions show the scalability of Jitter and Lost Packets vs. Number nodes

Packets) against the number of nodes. In this manner, we could predict what happen when the number of nodes and flow are traffic is increased.

2.3 Conclusions

In this case, we performed the HMIPv6/MPLS scenario simulation using CBR as test traffic. Various QoS metrics were analysed, such as delay, which on average was 66,82 ms; the jitter, which was rather variable, and throughput, which reached 446,0 Kbps on average. On the other hand, in the course of the simulation, 3,74 packets were sent and 207 were lost; that represents 5,54% of all packets. Therefore, we conclude that the simulation scenario showed very good values of delay and throughput, acceptable packet loss and very irregular jitter figures, so that, in order to achieve good levels of QoS, the performance of jitter has to be improved. A similar scenario with FHMIPv6 instead of HMIPv6 could solve this point.

3. FHMIPV6/MPLS integration

One of the major problems encountered in the integration HMIPv6/MPLS is the amount of signaling load in Binding Update (BU). Especially in case of a handover. At the time of BU can cause problems of safety and quality of services. With respect to security can be sent or received malicious messages, relative to the quality of services, excessive signaling load can significantly degrade the QoS metrics evaluated.

For this reason, we propose FHMIPv6/MPLS integration as a mechanism that will avoid both these problems. FHMIPv6 has a process of pre and post registration which keeps the communication between the mobile node and access router. FHMIPv6 has a process of pre and post registration which solves the problem observed in HMIPv6/MPLS integration. This we can say based on the work of R. Hsieh. FHMIPv6/MPLS integration has been made in the same manner as HMIPv6/MPLS integration. This integration allows us to compare which is better.

Is important mentioned, Fast Handover for Mobile IPv6 (FMIP) is a mobile IP extension that allows the MN to set up a new CoA before a change of network happens. This is possible because it anticipates the change of the router of access when an imminent change of point of access is detected. This anticipation is important because it minimises the latency during the handover, when the MN is not able to receive packets.

FHMIPv6 had been initially proposed by Robert Hsieh [hsieh03] as a way of integrating Fast Handover and HMIPv6 and shows why this integration is a better option than HMIPv6 alone.

3.1 Scenario of simulation

The scenario simulated is shown in figure18. The MN is in the area of HA. Bandwidth configuration and delay of each link are shown below in table3.

The traffic used was CBR, since it allows audio and video simulation in real time. These applications have a high demand of QoS.

Figure18 shows the topology of the simulated network. MPLS is the core of the network and is constituted by the following nodes: 1 (MAP), 2 (LSR1), 3 (LSR2), 4 (LSR3), 7 (LER1 for

Link	Delay	Bandwidth
CN-LSR1	2ms	100Mb
LSR1-HA	2ms	100Mb
LSR1-MAP	50ms	100Mb
MAP-LSR2	2ms	10Mb
MAP-LSR3	2ms	10Mb
LSR2-PAR	2ms	1Mb
LSR3-NAR	2ms	1Mb

Table 3. Bandwidth and delay configuration

MPLS and PAR for F-HMIPv6) and 8 (LER2 for MPLS and NAR for F-HMIPv6); the tag distribution protocol used by MPLS is RSVP.

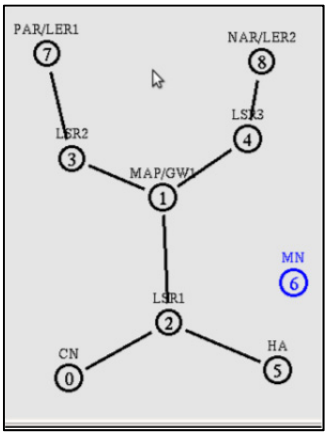


Fig. 18. Scenario FHMIPv6/MPLS Integration

Every link shows two of their characteristics: bandwidth (in megabits or kilobits) and delay (in milliseconds). A few seconds later, the MN moves towards the area of PAR, as figure 19 proves.

Finally, the MN moves to the area of NAR (figure20).

3.2 Description of simulation

Initially, the MN is located in the area of the HA. 2 seconds after the start of the simulation, the HA moves towards the area of the PAR at 100 m/s, arriving at t=3,5 s approximately. At t=5 s, the CN begins sending CBR traffic to the MN following the route CN→LSR1→HA→LSR1→MAP→LSR2→PAR→MN as shown in figure 19. Then, at t=10 s, the MN starts moving to the area of the NAR at 10 m/s. At the same time, the handover takes places at around t=13,12 s and the MN receives one of the first packets from the NAR at t=13,14 s approximately. Afterwards, the MN places in the area of the NAR at around t=17 s. Finally, at t=19 s, the CN stops sending traffic flow towards the MN.

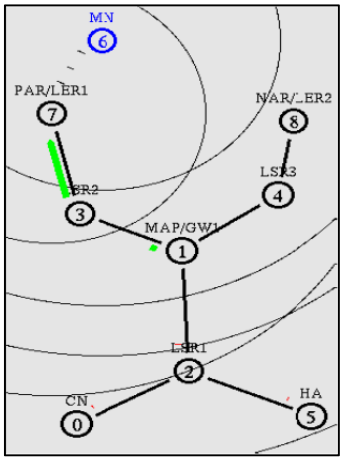


Fig. 19. The MN moves towards the area of PAR

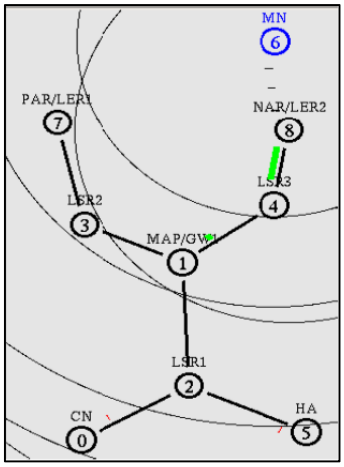


Fig. 20. The MN moves to the area of NAR

3.3 Scalability

The objective of this simulation with different scenarios was to analyse QoS metrics in HMIPv6/MPLS integration with CBR traffic and the scalability. The table4 show the different scenarios simulated. The first scenario was proposal by R.Hsieh, the other scenarios were increasing the number nodes in order to test the scalability, the table show the results of different metrics analysed.

The table4 shows that the delay, jitter, throughput and packet loss rate vary slightly as the topology and the network flow increase. Therefore, we can conclude that the FHMIPv6/MPLS integration keeps the quality of service (QoS) high, despite the growth of the network and the traffic flow.

Nodes	Delay(ms)	Jitter(ms)	Throuhgput (Kbps)	Send Packets	Received Packets	Lost Packets (%)
9	67,16	0,47	446,05	3734	0	0
15	278,82	2,41	334,37	3734	2871	23,11
20	255,9	2,03	372,54	3734	3158	15,43
25	314,41	4	286,64	3734	2435	34,8
30	315,12	3,6	303,89	3734	2582	30,85
35	313,62	4,04	286,96	3734	2437	34,73
40	305,83	4,03	281,91	3734	2395	35,86
45	309,3	4,28	274,85	3467	2168	31,96

Table 4. FHMIPv6/MPLS Integration

The (figure 21) shows the results of the following metrics obtained of the table 2. In this manner can visualize the behavior of: delay, throughput, send and received packets against the quantity of number of nodes.

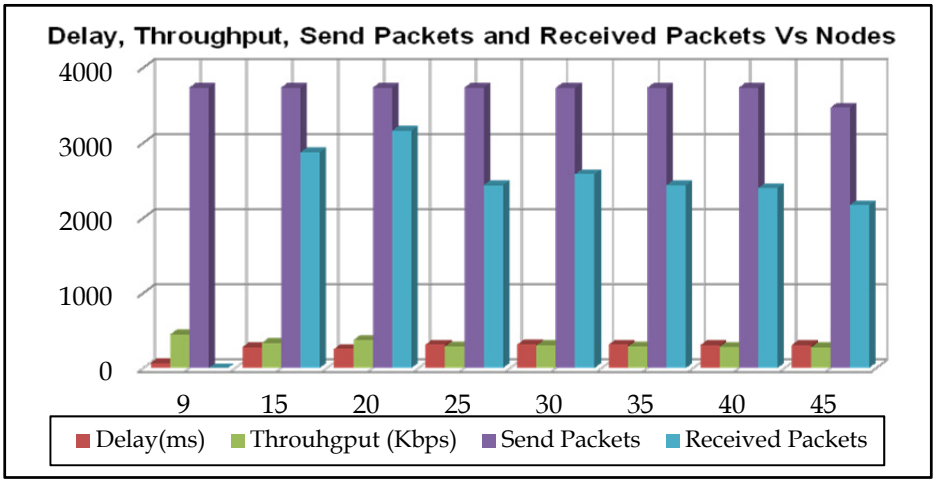


Fig. 21. Delay, Throughput, Send and Received Packets vs. number nodes

In order to extend the different results obtained in the simulations, the function (figure22) shows the behavior for different scenarios of simulation. With this functions can know what happened with the metrics (Delay, Throughput, Send and Received Packets) and the number nodes. In this manner we could predict what happens when the number of nodes and flow of the traffic are increased.

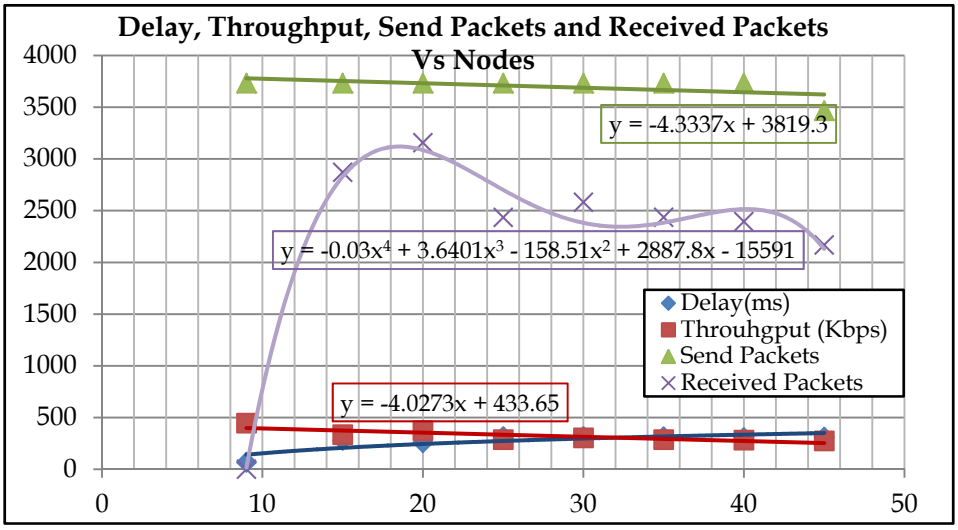


Fig. 22. The functions shows the delay, throughput, send and received packets vs. number nodes

The (figure23) shows the results of the following metrics obtained of the table 2. In this manner can visualize the behavior of jitter and lost packets nodes and the number nodes.

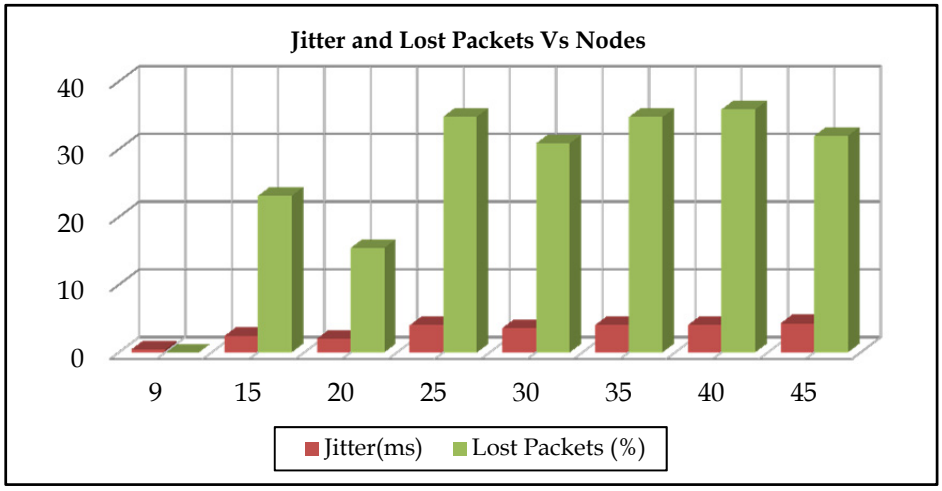


Fig. 23. The functions shows the jitter and lost packets vs. number nodes

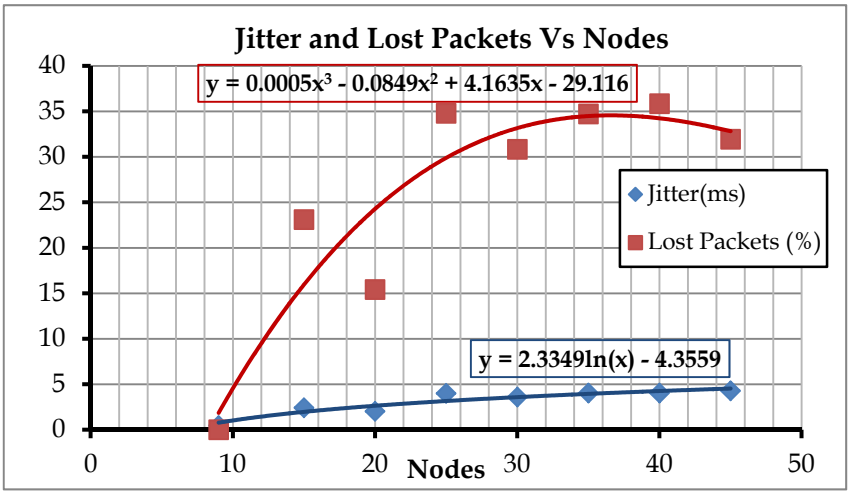


Fig. 24. The functions shows the jitter and Lost packets vs number nodes

In order to extend the different results obtained in the simulations, the function (figure24) shows the behavior for different scenarios of simulation. With this functions can know what happened with the metrics (Delay, Throughput, Send and Received Packets) and the number nodes. In this manner we could predict what happens when the number of nodes and flow of the traffic is increased.

4. FHAMIPv6/AODV integration

FHAMIPv6/AODV present the integration of protocol Fast Hierarchical Ad-Hoc Mobile IPv6 (FHAMIPv6) and the Ad hoc On Demand Distance Vector (AODV). The integration shows the effects of FHAMIPv6/AODV about the QoS. The simulation was realized in NS-2 version 2.32. The traffic used is TCP. We analyze the delay, jitter and throughput in an end to end communication. The metrics from the ACN perspective are presented. The integration FHAMIPv6/AODV is a work advance of the integration FHAMIPv6/ MPLS/ AODV in order to provide quality of service in MANET networks. We can consider FHAMIPv6/AODV and the following integration FHAMIPv6/MPLS as part of the development of LTE standard included in the all-IP concept that allows us to meet some requirements of LTE.

From the table 5 highlight the link AN1 - MAP/GW1 has a bandwidth and delay than the rest, because it represents a connection to the Internet.

The (figure25). Shows that initially (6) AMN is in the area of the (5) AHA in communication with the (0) ACN, also we can see that the core consists of MPLS nodes MAP/GW1, LSR2, LSR3, PAR/LER1, NAR/LER2.

Where MAP/GW1 node performs the functions of default gateway, nodes and LSR3 LSR2 are used simultaneously as routers, LSPs and FHAMIPv6 intermediate nodes.

Link	Bandwith(Mbps)	Delay(ms)
AN1-- MAP/GW1	100	50
MAP/GW1 – LSR2	10	2
MAP/GW1 – LSR3	10	2
LSR2 -- PAR/LER1	1	2
LSR3 -- NAR/LER2	1	2

Table 5. Characteristics of the links FHAMIPv6/AODV

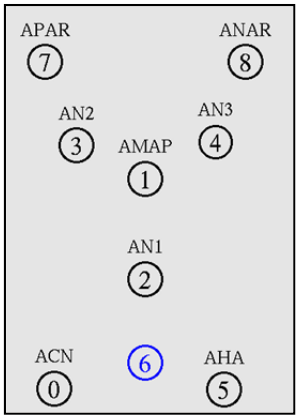


Fig. 25. Illustrates the simulation scenario (base)

Nodes can also be NAR/LER2 PAR/LER1 and have functions MPLS edge router and access router FHAMIPv6.

On the other hand, operates as a node AN1 intermediate FHAMIPv6 but no MPLS features, while ACN and AHA are the CN and HA, respectively, at last, and AMN is the mobile node MN.

4.1 Description of simulation

The AMN (blue node in (figure23) is initially located in the area of the ACN. Here, communication between these nodes occurs with no intermediary elements.

In the 1,3th s, ACN starts to transmit TCP packets towards the AMN. They are transmitted with an average delay of 4,99s. Until the 5th s, communication flows normally. After the 5th s, the AMN starts to move towards the APAR. While this is happening, communication with the ACN is not affected until the 5,43th s, when it is out of the ACN rank. From that mentioned instant until the 6,53th s, the AMN does not receive any packets from the ACN. In the 6,27th s, the AMN locates next to APAR. Around this time (and in many other moments) certain UDP signalling is shown in the network. This signalling corresponds to the AODV signalling packets. That routing protocol takes almost 250 ms to learn the new AMN position. It is only in the 6,53rd s that the AMN resumes the session with the ACN. From that instant until the 14,6th s, communication results as follows:

ACN→AN1→AMAP→AN2→APAR→AMN

In that moment, the AMN begins moving towards the ANAR and finishes in the 15,0th s. In the 15,08rd s the AMN receives the first packet from the ANAR. From then on, this will be the route that will allow the AMN access to the FHAMIP network. Simulation ends after 20 seconds of starting.

4.2 Scalability

The figure26 shows the results of the following metrics obtained of the table6. In this manner can visualize the behavior of: delay, throughput, send and received packets against quantity of number nodes.

Nodes	Delay(ms)	Jitter(ms)	Throughput(Kbps)	Send Packets	Received Packets	Lost Packets (%)
15	320,93	52,76	136,32	655	615	6,11
20	308,03	46,92	136,9	658	617	6,23
25	330,34	58,05	126,04	608	571	6,09
30	314,55	59,67	99,5	460	450	2,17
35	263,93	50,65	111,7	539	505	6,31
40	214,36	62,8	75,71	354	332	6,21
45	309,0	104,53	56,23	267	255	4,5

Table 6. Shows FHAMIPv6/AODV

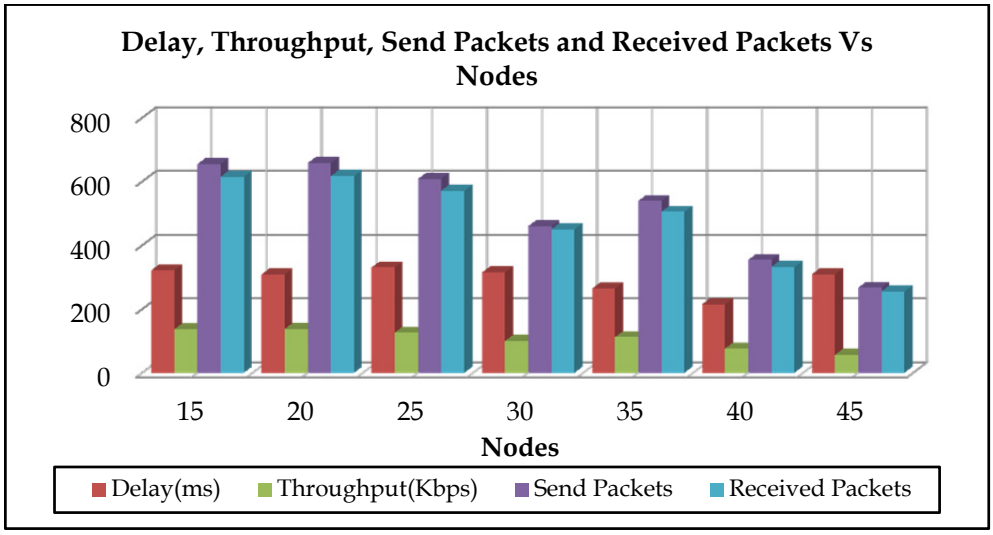


Fig. 26. Delay, Throughput, Send and Received Packets vs. Number nodes

In order to extend the different results obtained in the simulations, the function (figure27) shows the behavior for different scenarios of simulation. With this functions could predict what happens with the metrics (Delay, Throughput, Send and Received Packets) against quantity of number of nodes. In this manner we could predict what happen when the number of nodes and the flow are traffic is increased.

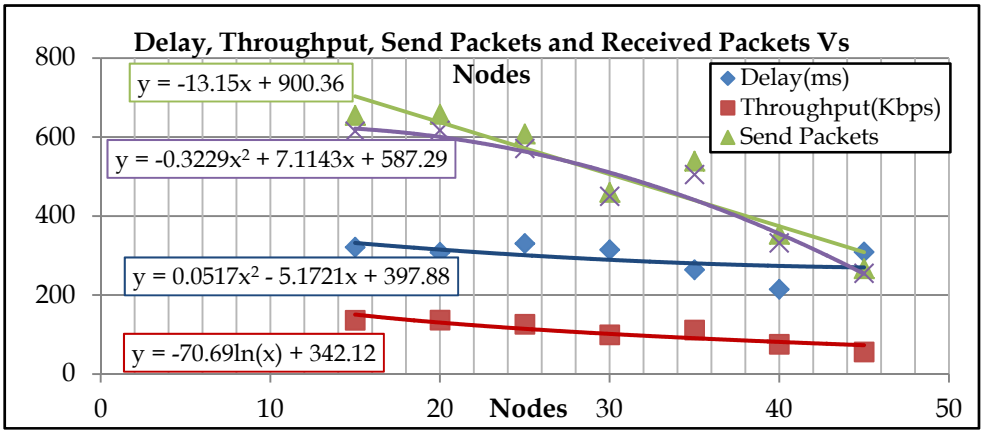


Fig. 27. The figure show the functions Delay, Throughput, Send and Received Packets an Number nodes

The figure 28 shows the results of the following metrics obtained of the table 2. In this manner can visualize the behavior of delay, throughput, send and received packets with different number nodes.

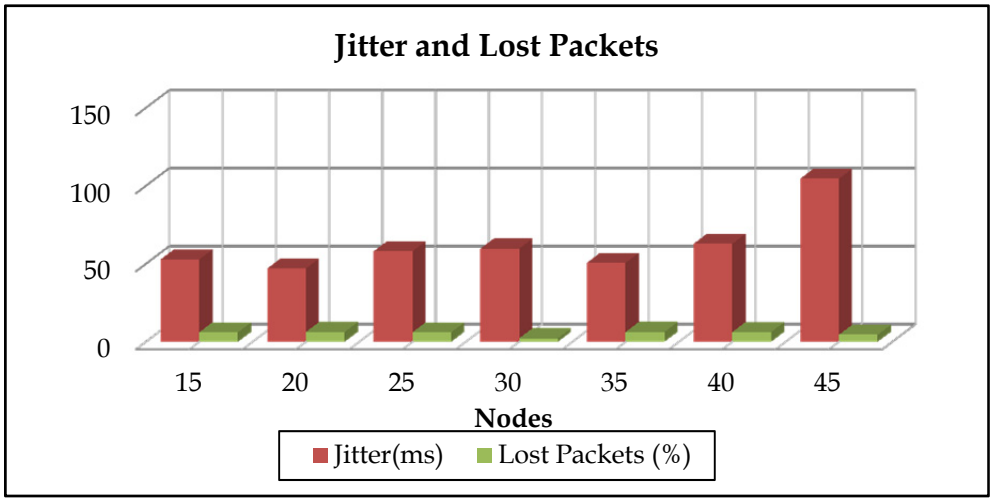


Fig. 28. Jitter and Lost Packets vs Number nodes

In order to extend the different results obtained in the simulations, the function (figure29) shows the behavior of the different simulation scenarios. With this functions could predict what will happen with the metrics (Delay, Throughput, Send and Received Packets) and the number nodes. In this manner, we could predict what happens when the number of nodes and flow of the traffic is increased.

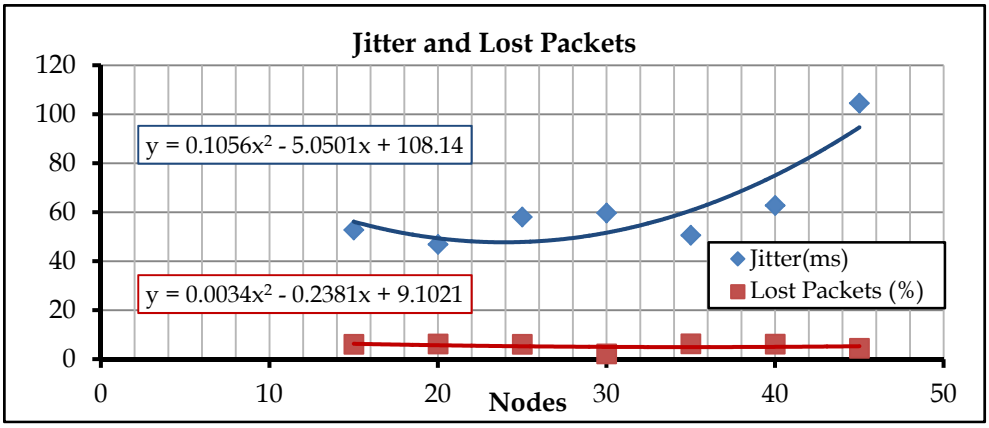


Fig. 29. The figure show the functions Jitter and Lost Packets vs. Number nodes

4.3 Conclusions

This research shows the effects of the FHAMIP/AODV integration over the QoS metrics. The simulation proved that the average delay was approximately 112,27 ms and was penalized by the AODV signalling, so it was necessary to update the status of the routes. On the other hand, the average jitter analysed reached 38 ms.

Regarding the loss of packets, a total of 86 did not reach the destination. Most of them were lost when the AMN moved either towards the APAR or to the ANAR.

The jitter was quite satisfactory given the fact that it exceeds 176 Kbps. In general, the delay and the jitter suffer the strong effects of the AODV routing updates. Some nodes stop sending TCP packets to transmit useful AODV signalling to recalculate routes, increasing the delay in a TCP session significantly. A possible solution (assuming that only a node moves on) would be to modify AODV in order to stop routes updating until the APAR and the ANAR receive a MAP_REG_REQUEST from the AMN. This would indicate that the AMN is in its own area.

5. FHAMIPv6/MPLS integration

FHAMIPv6 protocol was created as an extension to support FHMIPv6 hierarchical addresses in MANET networks, but FHAMIPv6, is not an protocol to provide quality services in such networks. For this reason, it was necessary to integrate MPLS and FHAMIPv6 in order to provide QoS in MANET networks.

To achieve the integration was necessary to modify the source codes of MPLS and FHAMIP. In this section the same way as in the other sections, we used the base scenario proposed by R. Hsieh and then the number of nodes and traffic flow was increased in order to analyze the scalability of the integration. The Tests were realized with: 9, 20 and 30 nodes. The QoS metrics analyzed were: Delay, jitter, throughput, send and received packets and lost packets.

The figure 30 Shows that initially the AMN is in the area of the AHA in communication with the ACN, it can also be observed that the core MPLS is formed by MAP/GW1, LSR2, LSR3, PAR/LER1, NAR/LER2 nodes. Where the MAP/GW1 node performs the functions of default gateway, the nodes LSR3 and LSR2 are used simultaneously as Label Switching Routers and intermediate nodes FHAMIPv6; it can also be observed that the nodes PAR/LER1 and NAR/LER2 have functions of MPLS edge router and access router for FHAMIPv6. Furthermore, the node AN1 only functions as an intermediate FHAMIPv6 node and has no has no MPLS functions, while ACN and AHA nodes correspond to the corresponding node and base agent respectively, lastly the AMN node represents the mobile node. With regards to the characteristics of the wired links, table7 presents details. From the table above, we can highlight the fact that the link AN1 - MAP/GW1 has a superior bandwidth and delay than the rest, because it represents a connection with Internet.

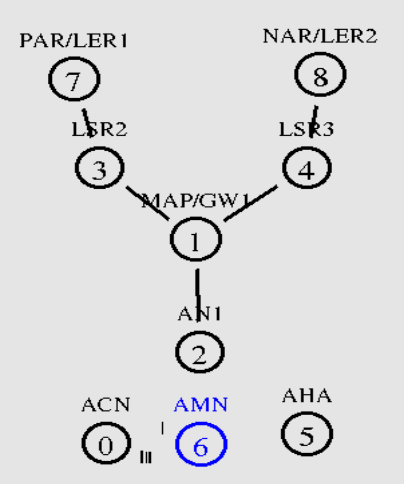


Fig. 30. Scenario of simulation

Link	Bandwith(Mbps)	Delay(ms)
AN1-- MAP/GW1	100	50
MAP/GW1 – LSR2	10	2
MAP/GW1 – LSR3	10	2
LSR2 -- PAR/LER1	1	2
LSR3 -- NAR/LER2	1	2

Table 7. Presents the characteristics of the wired wireless links.

5.1 Description of simulation

This section will describe in detail, and scenario by scenario, all relevant events in each simulation, details on the movement of the AMN will be presented, the moments of reserve of resources through RSVP and some comments on the transfer; not without mentioning that for all scenarios a FTP traffic type was used and the following metrics of QoS were defined:

Delay, jitter, throughput, TCP congestion window and lost packets. The choice of the afore mentioned metrics is because they are the most affected when the amount of network traffic is very high, in addition these are the metrics that affect more significantly the traffics that have high QoS requirements such as video, audio and real-time applications.

It is noted once again that scenarios with different number of nodes were simulated to study the impact of this change in the behavior of the proposed integration compared to the QoS metrics and to evaluate the functionality of the proposed integration to such scenarios.

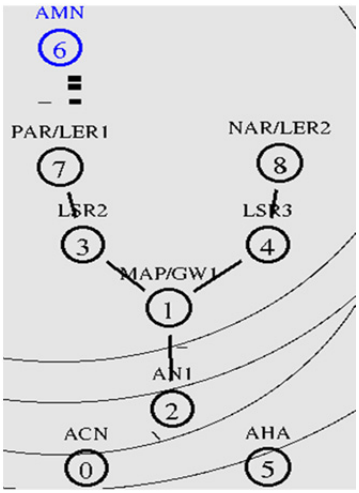


Fig. 31. AMN in PAR/LER zone

At the initial instant the mobile node (AMN) is in the area of its home agent (AHA) as shown in Figure 29, then at t = 1.2s the AMN starts transferring FTP traffic with the ACN, there upon between t = 3.5s and t = 4.5s MPLS / RSVP resource reservation takes place on the path MAP/GW1 - LSR2 - PAR/LER1. Then at time t = 10s the AMN begins its displacement towards the PAR/LER1, at a speed of 100m / s arriving shortly to this area from which it will use the PAR/LER1 - LSR2 - MAP / GW1 - AN1 route to communicate.

A few seconds later between t = 14.5s and t = 15.5s resource reservation along the route MAP/GW1 - LSR3 - NAR/LER2 takes place anticipating the subsequent transfer made by the AMN which moves at 10m / s from PAR/LER1 toward the NAR/LER2 at t = 16s. From there on, the traffic will follow the NAR/LER2 - LSR3 - MAP/GW1 - AN1 route to communicate with the ACN and the AHA. This is illustrated in the figure 32.

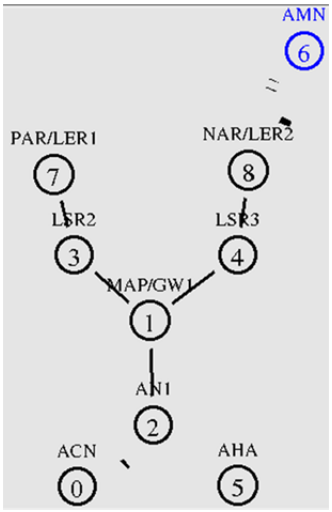


Fig. 32. AMN in NAR/LER zone

5.2 Scalability

In the same way, we simulated with 20 and 30 nodes. See figures 33 and 34.

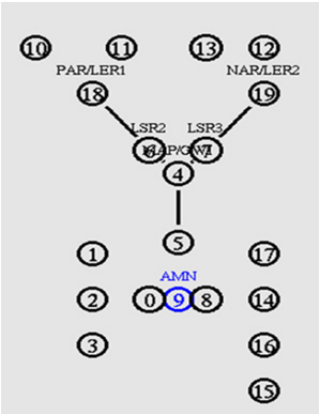


Fig. 33. Scenario with 20 nodes

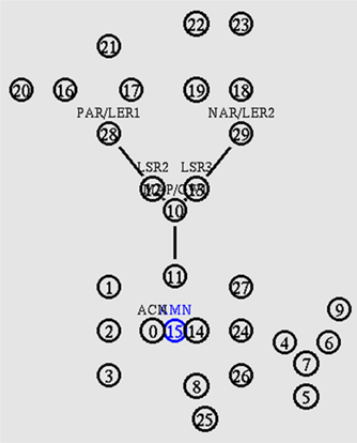


Fig. 34. Scenario with 30 nodes

5.2.1 Analysis of delay

As shown in the (figure 35), in the time that the AMN is in the AHA zone (between t = 1.2s and t = 10s), the traffic experiences a delay below 250ms, this is due mainly to the fact that

the AMN communicates directly with the ACN, that is to say, traffic does not pass through the intermediate nodes. Then we can see a blank space, which corresponds to the time when the AMN moves towards the PAR/LER1 and does not send traffic to the ACN. Beyond the time t = 10s we can observe that the experienced delay increases, at this time the AMN is fully in PAR/LER1 zone. A few seconds later we see a growing tendency of the delay until reaching a blank space, this behavior corresponds to the time when the AMN performs the transfer from the PAR/LER1 to the NAR/LER2 between times t = 16s and t = 20s . Finally the delay adopts a regular behavior close to the 350ms, which is maintained until the end of the simulation. The average delay was 224.521ms.

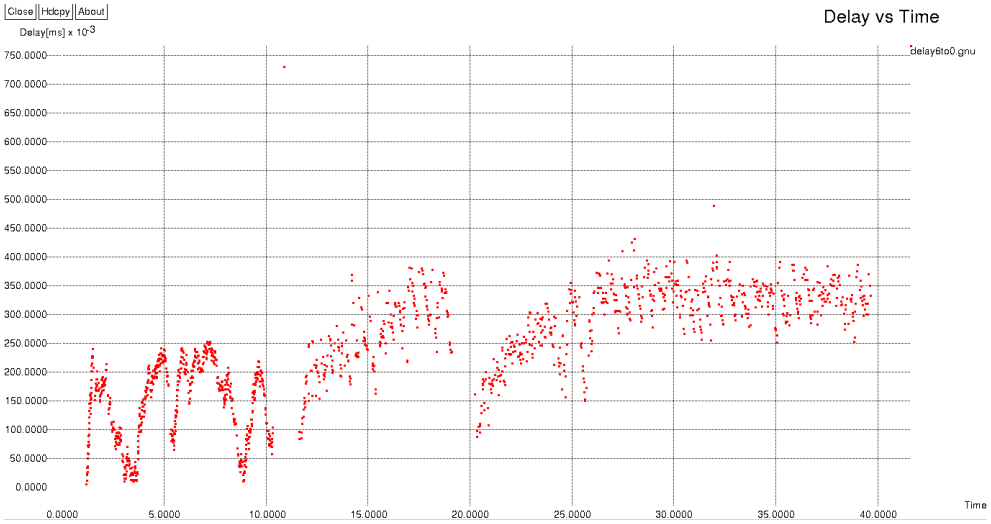


Fig. 35. Illustrates the behaviour of the delay vs. time in the simulation

5.2.2 Analysis of jitter

The (figure36) illustrates the jitter behavior as time in the simulation.

As it can be seen in figure 36, the jitter has a similar behavior to the delay during the first 10s of simulation, in the sense that both present the lowest values throughout the simulation in this range, but after the AMN moves towards the PAR/LER1 a huge peak of about 650ms is registered, this corresponds with the packet that experiences more than 700ms in delay in figure 80. After this, the jitter is stabilized below 50ms when the AMN is in the ANAR/LER2 zone (after t = 20s) and below 100ms when the AMN is in the APAR/LER1 zone (between the 11s and 18s or so). Additionally it is noted that the transfer that takes place near the instant t = 16s has no significant effects on the experienced fluctuation. Finally the average jitter during the simulation was 15.84ms.

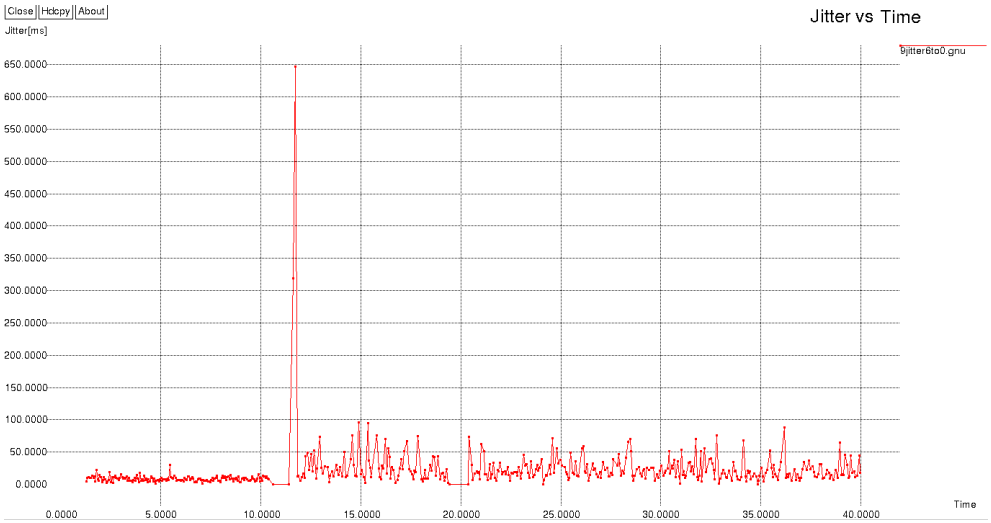


Fig. 36. Jitter vs. time

5.2.3 Analysis of throughput

The figure 37 shows the same trend that is reflected in the previous metrics, related to the fact that while the AMN is located in the AHA zone the metric performs better than in the rest of the simulation. In this occasion the throughput obtains values close to the 800Kbps before the 10s after the start of the simulation. Subsequently when the AMN moves to the PAR/LER1 zone performance drops to 0Kbps which is due to the absence of traffic at that moment and the loss of some packets while the displacement occurs. Then when the AMN reaches the area in question an irregular behavior of the performance is registered, which sometimes comes close to the 800Kbps which is close to the maximum possible limit of 1Mbps due to the LSR2 – PAR/LER1 link, while on the other hand also reaches values of about 50Kbps. Moments later, after the AMN moves towards the ANAR/LER2 the performance drops once again to 0Kbps due to decreasing traffic and the lost of some packets during the transfer. Finally, once the AMN arrives to the above-mentioned area, throughput shows a behavior similar to that reported in the PAR/LER1 zone. The average throughput of the simulation was 343.649Kbps.

5.2.4 TCP windows

This is also supported by the behavior of the TCP congestion window presented in the figure 38 as it can be seen, the instants near t = 11s and t = 20s show the drop in the TCP congestion window which indicates the loss of packets, as was mentioned above.

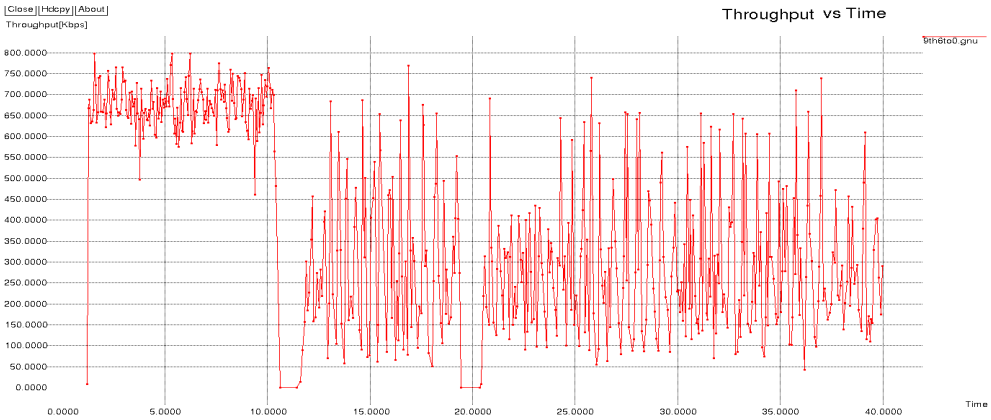


Fig. 37. Illustrate throughput vs. time

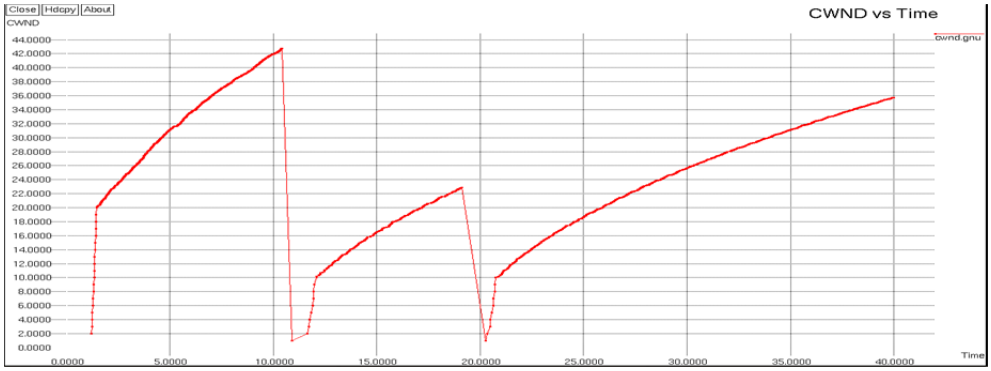


Fig. 38. TCP congestion window vs. time

5.2.5 Analysis of lost packets

During the simulation 1610 packets were sent from the AMN to the ACN, out of which 1586 reached their destination, leading to the loss of 24 packets, which corresponds to 1.49% of the total packets.

5.2.6 Analysis of results

The table 11 presents various facts to highlight: first, both the delay and the fluctuation do not exhibit increasing tendency as the number of nodes increases, this is important because

it shows that the proposed integration is functional in the presence of more than 9 nodes and also that the metrics in question do not deteriorate significantly in scenarios of large volumes of nodes. Another important fact to highlight is that performance observes a relationship of inverse proportion to the number of nodes that make up the simulation scenario that is to say, that with more nodes the performance decreases, however this decrease is not linear but it is less affected with the presence of new nodes. As the tendency is presented, it could be said that it is possible that for scenarios of more nodes performance will be stabilized around a certain value, which means that the decrease has a limit. The last fact to note is that the proportion of lost packets does not increase as the number of nodes increases, but stabilizes after some growth in the network. Therefore we can conclude that the proposed integration is useful for providing QoS in scenarios with large volumes of nodes.

Scenario\Metric	Delay(ms)	Jitter(ms)	Throughput(Kbps)	Send Packets	Lost Packets
9 nodes	224.52	15.84	343.65	1610	24
20 nodes	275.20	27.87	241.60	1168	53
30 nodes	225.52	21.99	206.87	989	34

Table 8. Nodes vs. Different metrics

In this figure 39 we can visualize the following metrics (Delay, Jitter, Throughput, Send and Lost Packets vs. Number nodes). The Delay, Jitter and Throughput have slight variation. The throughput decreases when increasing the number nodes, likewise the send packets decreases when increased the number nodes and traffic flow. This behavior of these metrics is logical, we did not test with more nodes, because we believe that these tests is enough to make an analysis

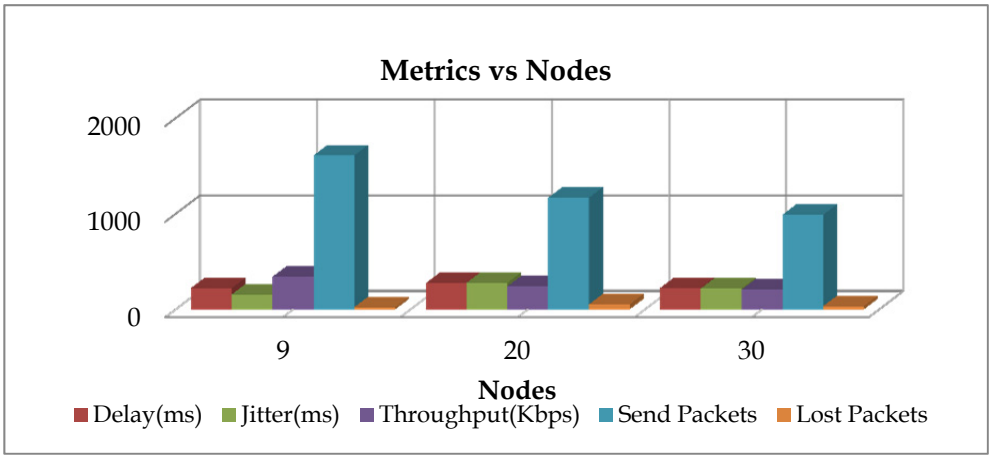


Fig. 39. Metrics vs. number nodes in FHAMIPv6/MPLS integration

5.3 Conclusions

This chapter released the results of the integration FHAMIPv6/MPLS and features to provide QoS. This study is of considerable importance because it is the first to bring mobile capabilities, fast handover and hierarchical IP extensions to MPLS hybrid environments. Thus provides the basis for future research that want to implement prototypes in real environments.

6. General conclusions

This chapter is focused on all IPv6/MPLS scheme for wireless mobile networks. We presented different integrations of mobility protocols (versions6 IP protocol extensions) with quality of service (QoS) protocols (MLPS, RSVP). The initial integrations were performed in infrastructure networks. The results delivered valuable information on how the protocols operated as well as the different coupling options available. This shows that the best coupling option was that where it is necessary to modify the protocols in a way that all could work as one single protocol. Other options were discarded, since protocols operating independently or even synchronised did not deliver satisfactory results. Among the quality of service protocols, we managed to prove that the RSVP was valid as a signalling protocol. This was also confirmed at the IETF when protocol CR-LPD was discarded as a signalling protocol when it was used together with MPLS.

On the other hand, in order to integrate IP protocol extensions (IP mobile, HMIPv6, F-HMIPv6 and FHAMIPv6) and MPLS protocol, it was necessary to modify MPLS nodes to turn them into mobile MPLS nodes. It was proved that IP mobile protocol, when integrated with MPLS, works better in macromobility scenarios. For micromobility scenarios, it is more convenient to use hierarchical IP mobile extensions since the signalling load is higher. The integration MPLS and HMIPv6 protocol extensions formed a good coupling for infrastructure networks in order to provide QoS. On the contrary, in total ad-hoc networks it is almost impossible because MPLS/Diffserv provides end-to-end quality of service, and when integrated with HMIPv6, the signalling load was so high that the network resulted overloaded.

Another problem was the compatibility of the source codes to perform the simulation to migrate from one version to another. The protocols did not work correctly. For this reason, we tested the F-HMIPv6 and MPLS protocols to verify if this was the best option to provide QoS to the next generation of mobile networks. In full ad-hoc mobile networks, FHMIPv6 showed diverse inconveniences, so it had to be modified to assume a new agent. This new agent was in the origin of the FHAMIPv6 protocol and the AHRA routing protocol. In order to solve the problem of the routing protocol AHRA, FHAMIPv6 was integrate with AODV and the result was successful. Similarly, we integrate FHAMIPv6 and MPLS and the result was satisfactory. With this result, we have achieved to propose an alternative to one of the great challenges of ad hoc networks. Because to provide QoS in ad hoc networks is a big challenge.

The quality of service values were obtained when a handover occurred and the results were satisfactory. In general, we can affirm that during a handover, not only metrics such as delay jitter and throughput improved, but also the default quality level was maintained in the integrations performed. The results obtained allowed us to identify which integration

protocols were the most suitable to ensure QoS in all IPv6/MPLS network. A series of architectures for next generation hybrid networks were proposed, including several important applications for universities, industry and the government. In general, the coupling between the quality of service and mobility protocols mentioned before is an excellent option to provide QoS in mobile networks and, especially, in the ad-hoc mobile ones. An interesting topic that we are currently evaluating is the different security issues that are generated in coupling protocols, which can actually degrade the quality of service by the action of malware or malicious users. On the other hand, we can say that, in next generation networks (4G), an all IPv6/MPLS architecture will be critical in next generation wireless mobile networks, compatible with the standards proposed so far (WIMAX, advanced LTE/SAE, LTE/IMT, WiMAX/IMT).

7. References

[1] "Mechanisms of quality of service and mobility in 4G".Jesús Hamilton Ortiz, Juan Carlos López and Carlos Lucero. Global Mobile Congress (GMC), Shanghai, China, 2009.

[2] Integration of HMIPv6 mobility protocol and Diffserv quality of service protocols over M-MPLS to provide QoS on IP mobile networks". Jesús Hamilton Ortiz. 10th IFIP International Wireless Communications Conference PWC´05 IEEE. Colmar, France, August 25-27, 2005.

[3] "Integration of protocols FHMIPv6/MPLS in hybrids networks". Jesús Hamilton Ortiz, Jorge Perea, David Santibáñez, Alejandro Ortiz. Journal JSAT. 2011.

[4] "FHMIPv6/MPLS to provide QoS in 4G". Jesús Hamilton Ortiz, Juan Carlos López.An International Interdisciplinary Journal. 2012.

[5] "Metrics of QoS for HMIPv6/MPLS in a handover". Jesús Hamilton Ortiz, Jorge Perea, Juan Carlos López. An international Interdisciplinary Journal. 2011.

[6] "AHRA: A routing agent supporting the IPv6 hierarchical mobile protocol with fast-handover over mobile ad-hoc network scenarios". Jesus Hamilton Ortiz, Santiago Gonzales, Jorge Perea, Juan Carlos López. Journal, Editorial IJRRCS.

[7] "Integration of protocols FHAMIPv6/AODV in Ad hoc networks" Jesús Hamilton Ortiz, Juan Carlos López, Jorge Perea. "Network Protocols and Algorithms" Macro think Institute.

[8] "Services de mobile IP au dessus de MPLS". DEA, thesis, Lina E. Mekkaoui: Lebanese University, 2005.

[9] "Performance analysis on hierarchical mobile IPv6 with fast-handoff over end-to-end TCP". Robert Hsieh, GLOBECOM, 2002.

[10] "A comparison of mechanisms for improving mobile IP handoff latency for end-to-end TCP". Robert Hsieh, Globes 2003.

2

A QoS Guaranteed Energy-Efficient Scheduling for IEEE 802.16e

Wen-Hwa Liao and Wen-Ming Yen
Department of Information Management, Tatung University, Taipei
Taiwan

1. Introduction

Recently, the IEEE 802.16 standard (IEEE Std 802.16-2004, 2004), a solution to broadband wireless access commonly known as Worldwide Interoperability for Microwave Access (WiMAX), has been considered as a promising standard for next generation broadband wireless access networks. IEEE 802.16e (IEEE Std 802.16e-2005, 2005), also called Mobile WiMAX (Li et al., 2007), provides enhancements to IEEE 802.16 to support the mobility of Mobile Subscriber Stations (MSSs) at vehicular speed. Like other wireless systems, conserving energy is one of the critical issues for MSSs in IEEE 802.16e. Therefore, it is required for the protocol to offer a well-designed energy-efficient algorithm for an MSS.

IEEE 802.16e is expected to support Quality of Service (QoS) for real-time applications such as Voice over IP (VoIP), video streaming, and video conferencing with different QoS requirements (Wongthavarawat & Ganz, 2003; Zhu & Cao, 2004). Such applications are delay and delay variation susceptible. For example, when data packets incur vast delays and delay variations, the quality of the application seriously degrades. In order to avoid such situation, QoS provides the guarantee of transmission. IEEE 802.16e defines five types of service classed: Unsolicited Grant Service (UGS), Real-Time Variable Rate (RT-VR), Non-Real-Time Variable Rate (NRT-VR), Best Effort (BE), and Extended Real-Time Variable Rate (ERT-VR). Among them, the UGS is designed to support Constant Bit Rate (CBR) services, such as T1/E1 emulation, and VoIP without silence suppression. These kinds of services generate fixed-size data packets on a periodic basis. They usually require stringent QoS delay constraints, so determining the length of sleeping duration of an MSS in IEEE 802.16e is not only bounded by the total amount of traffic generated by the connections in the MSS, but is also restricted by the connections' QoS delay constraints. IEEE 802.16e was developed for the targets on mobile devices which are generally powered by energy-limited batteries. Thus, the energy-efficiency is an important issue to extend the lifetime of MSSs (Jang et al., 2006; Mukherjee et al., 2005; Tian et al., 2007). When a connection is established, an MSS may shift the operation status into sleep mode in order to save the power consumption if there are no packets to send or to receive in certain frame durations. Under sleep mode, there are two intervals: sleeping interval and listening interval. During the sleeping interval, an MSS can be powered down by putting its wireless network interface into sleep mode. Aside from this, the MSS would be unable to send or to receive packets during sleeping intervals. After a sleeping interval finishes, the MSS switches to listening interval. The MSS wakes up during the listening interval to check

whether there are packets destined to it. Message packets are checked to determine whether the MSS should be woken up or not. IEEE 802.16e defined three types of Power-Saving Classes (PSCs) for connections with different characteristics, and each PSC is defined for a set of connections with common properties. A PSC is composed of interleaved listening windows and sleep windows. In PSC Type I, the sleep window is exponentially increased from a minimum value to a maximum value. This is typically done when the MSS is doing best-effort and non-real-time traffic. PSC Type II has a fixed-length sleep window and is used for UGS service. PSC Type III allows for a one-time sleep window and is typically used for multicast traffic or management traffic when the MSS knows when the next traffic is expected.

There are many previous researches that have devoted their efforts to adapting the sleeping duration of IEEE 802.11 and IEEE 802.15 (Liao & Wang, 2008; Liu & Liu, 2003; Tseng et al., 2002; Ye et al., 2004; Zheng et al., 2005). However, due to lack of QoS requirements, the results of those searches cannot be applied to IEEE 802.16e directly. Several studies have been proposed to analyze the IEEE 802.16e's power while an MSS operates in the power-saving mode (Han & Choi, 2006; Lei & Tsang, 2006; Seo et al., 2004). Several studies (Fang et al., 2006; Huang et al., 2007; Jang et al., 2006; Tsao & Chen, 2008) investigated the power consumption issues of IEEE 802.16e and suggested algorithms to determine the sleep interval in order to improve energy efficiency. In (Jang et al., 2006), the length of sleeping period is adapted according to the traffic type. This scenario is valid only under one MSS, and the QoS delay constraint is not considered. In (Tsao & Chen, 2008), although the QoS delay constraints are considered, the scenario cannot consider the energy costs of status transition. In (Fang et al., 2006), a scheduling algorithm for multiple MSSs with QoS delay constraints is proposed. To save power, the algorithm grants a primary MSS the right to use the bandwidth in burst mode. Secondary MSSs are only given the necessary bandwidth to meet the requirements of QoS delay constraints. However, its benefit only exhibits when the total traffic loading of all MSSs is low. In (Huang et al., 2007), although the constant bit rate traffic with QoS delay constraint is considered, the scenario cannot consider the jitter constraint.

In this chapter, we propose a QoS guaranteed energy-efficient scheduling for IEEE 802.16e. We consider that delay and jitter types of QoS should be scheduled at the same time and integrate sleep duration in one MSS. The packets would be scheduled successively to reduce the number of status transitions under QoS requirements for delay and jitter. The proposed approaches not only minimize the power consumption of the MSS but also guarantee both delay and jitter QoS of real-time connections.

2. The QoS guaranteed energy-efficient scheduling for IEEE 802.16e

In this section, we first describe the basic idea of our algorithm for QoS guaranteed energy-efficient scheduling. Second, we define the notations of our system model. Finally, we schedule packets in an MSS with our QoS guaranteed energy-efficient scheduling. Additionally, we consider the QoS requirements of jitter constraint to schedule the packets and achieve the guarantees of transmissions.

2.1 Basic idea

The idea behind our proposed algorithm, called successive scheduling scheme (SSS), is to schedule the packet transmission in successively fashion with the minimal interval of listen

periods and a maximum interval of sleep periods without violating the QoS of all connections in an MSS. Additionally, the successive scheduling of time slots would reduce the number of status transitions between sleep periods and listen periods. This improvement greatly contributes to achieve the power-saving. The proposed approach can be adapted to the power-saving class of type III where the length of sleep and listen periods are variable.

2.2 System model

In this chapter, the centrally controlled IEEE 802.16e wireless network with a central BS and an MSS with multiple real-time connections is considered. The uplink and downlink channel is divided into fixed-size frames, and the frames are subdivided into fixed-size time slots. Both the energy consumption and the bandwidth are calculated in time slots. Different QoS parameters have been defined for various type of services in IEEE 802.16e, and all of them can be mapped into the minimum data rate requirements of the MSSs (Andrews et al., 2005). Therefore, we only apply the minimum data rate as the bandwidth requirement of QoS for each type of connection. Additionally, other QoS requirements such as the maximum latency and tolerated jitter would be considered in this chapter. The notations in this chapter are as follows: T_{aw} is the total number of time slots in which an MSS stays in the awake state; T_{st} denotes the total number of status transitions of an MSS from the sleep state to the awake state; Paw stands for the average energy consumption of each time slot by an MSS in the awake state; P_t represents the average energy consumption of each status transition from the sleep state to the awake state in an MSS; n denotes the index of time slot in an MSS; r_n stands for the data rate in which an MSS has been allocated by time slot n; $R_n^{\min}$ stands for the minimum data rate that an MSS should receive in order to guarantee its service quality in time slot n. We assume that there is no energy consumed during the sleep period of an MSS. Thus, the energy consumed of an MSS is determined by the number of the time slots it stays in the awake state and the number of status transitions it has from the sleep state to the awake state. The overall energy consumed by an MSS during period T, denoted as P, can be represented as follows:

$$P = T_{aw} \times P_{aw} + T_{st} \times P_t \tag{1}$$

The goal of the scheduling algorithm is to minimize the average energy consumed by an MSS during period T, while the QoS requirements such as minimum data rate, maximum delay constraint and tolerated jitter of an MSS must be guaranteed. Thus, we can minimize P by allocating the minimum time slots (T_{aw}) to satisfy the minimum data rates ($R_n^{\min}$) and successively schedule the packets to reduce the status transitions (T_{st}). In order to acquire the optimal result, the power-saving scheduling algorithm should consider the properties of the QoS requirements. We discuss the solutions of previous studies and present our QoS guaranteed energy-efficient scheduling to acquire the optimal result in the next section.

2.3 QoS guaranteed energy-efficient scheduling

First, we give the idea of our QoS guaranteed energy-efficient scheduling and perform the algorithm of our successive scheduling in an example. In the second part, we consider the jitter constraint of packet scheduling to provide more precise QoS guarantees.

2.3.1 The successive scheduling scheme (SSS)

To improve the power-saving performance, our algorithm will schedule packets into successive frames in order to reduce the number of status transitions in an MSS. The successive scheduling scheme is performed in two parts. The first part sorts all connections on the scheduling priorities of connections with tight delay requirements. The second part schedules the packets from the first priority connection into the successive frames. An MSS stays idle during sleep periods to save power, and only wakes up to transmit data packets during listen periods. Packets sent to the MSS during sleep periods are buffered at BS and are delivered to the MSS until the listen periods. In other words, the MSS only needs to receive and transmit data in listen periods and stay idle to conserve energy during sleep periods. The next paragraphs describe in detail the steps of our proposed successive scheduling scheme. Also, notations used in this chapter are summarized in Table 1.

Notation	Description
N	The number of connections
D_i	The delay constraint for connection i
I	The interval of packet arrival
$C_{i,j}$	The j^{th} packet for connection i
FIU	The frame-in-used; the frame which had already scheduled the packets without full-filled frame
FFU	The frame fully used; the frame which had already scheduled the packets without any available time slot
F_{next}	The unused frame that is next to the FFU and is more close to the next full-filled frame

Table 1. Notations and their descriptions.

To minimize the energy consumption of an MSS with multiple real-time connections, the successive scheduling scheme schedules the packets into their successive time slots under the radio resource and QoS requirements. Considering an MSS with N real-time connections, D_i is the delay constraint in milliseconds of any two consecutive packets for connection i, and I is the average inter-packet interval time in milliseconds for connection i. In this chapter, these connections could be either downlinked from a BS to an MSS or uplinked from an MSS to a BS. In the scheduling of downlink packets, our proposed scheme should be implemented on a BS. However, the proposed scheme must be realized on both a BS and an MSS if the proposed scheme is to be applied to the uplink packet scheduler. A BS can know the resource requirements of an MSS through negotiations in the requests from the MSS. Thus, a BS can determine the uplink packet schedule according to the proposed algorithm and provide transmission opportunities to an MSS. When a new connection to an MSS is initiated or any existing connection is released, the proposed scheme is activated to schedule or re-schedule resources in the following frames for the MSS. First, the successive scheduling scheme sorts all connections on an MSS according to their delay constraints and schedules these connections with tight delay requirements. The reason for this is that packets of these connections with tight delay requirements need to be sent or received within a small time window. The scheduler must consider these packets first in order not to violate their QoS requirements. Conversely, for packets that could tolerate more delays, the

scheduler can postpone the packets to schedule them successively without violating their delay constraints. After the scheduler decides on the scheduling priorities of connections, the packets from the first priority connection, e.g. connection i, are scheduled. $C_{i,j}$ is represented the j^{th} packet of connection i and the proposed scheme schedules $C_{i,j}$ with following steps: (1) The frames that are within D_i for $C_{i,j}$ and have already scheduled the packets without full-filled frames, called *FIU*. For the various applications, the proposed scheme is based on either the shortest delay or the longest delay. For the shortest delay based, if there are two or more *FIU* for $C_{i,j}$, the *FIU* with shorter delay receives a higher priority for $C_{i,j}$. The shortest delay based is applied to the urgent applications that are very strict with delay requirements and is used to prevent packets loss. Additionally, it is done to reduce the intervals of listen periods and increase the interval of sleep periods. In other words, an MSS cannot sleep in the time slots where there are already schedule packets. Thus, *FIU* is assigned first if the time slots of *FIU* are still available to accommodate $C_{i,j}$. On the other hand, the *FIU* with longer delay receives a higher priority in being scheduled to $C_{i,j}$. The longest delay based applied to scenarios which have loose delay requirements. The BS can decide on which strategy to perform in specific applications. (2) If there is no *FIU* for $C_{i,j}$, the scheduler will then pick a set of frames that are within D_i for $C_{i,j}$ and which already have scheduled packets without any available time slot. These are called *FFU*. The frames in the set are sorted by the D_i for $C_{i,j}$ in ascending order. The *FFU* will be the first frame and last frame in the set with the shortest delay based and longest delay based individuals. In order to reduce the number of status transitions, the scheduler will schedule the packets in successive time slots. In the successive listen periods, the MSS will not enter the sleep periods, and the number of status transitions would be reduced. Additionally, the sleep periods will be longer after their successive listen periods. To schedule the packets successively, the scheduler will find an unused frame that is next to *FFU* and is closer to the next full-filled frame, called F_{next}. The reason for this is that F_{next} is closer to the next full-filled frame has more chances to schedule the listen periods successively. In other words, packets that are scheduled to F_{next} and that is next to *FFU* will become an *FIU*. Obviously, *FIU* gains more opportunities to serve other packets in the following connections. Therefore, *FIU* will become *FFU* after full-filled frame with packets, and this *FFU* will be successive. The listen periods will be continuously without the sleep periods and the number of status transitions would be reduced. (3) If there are no *FIU* and *FFU* within D_i for $C_{i,j}$, the scheduler will schedule the packet into the last unused frame within D_i for $C_{i,j}$ and the unused frame will then become *FIU*. The last unused frame is selected is because once a frame is scheduled to transmit or receive packets, the frame becomes an *FIU*. As we mentioned, an *FIU* has more opportunities to serve other packets in the following connections. After the above steps, the successive scheduling scheme performs packet scheduling and achieves the power-saving for an MSS.

Fig. 1 shows the second step in the second part of the proposed algorithm. Based on the shortest delay, the scheduler chooses the first *FFU* to determine F_{next}. Because the 4th frame is an unused frame and is closer to the next *FFU*, which is the 5th frame, the scheduler determines the 4th frame as F_{next} and schedules the packet into the 4th frame. Once we determine the proper frame to be filled with packets, the time slots for transmission will be more successive for their following connections of scheduling. Thus, the 4th frame becomes *FIU* and has a greater chance to be filled with packets by the proposed algorithm. The status would not be switched from 3rd to 5th frame when the 4th frame is filled up with packets.

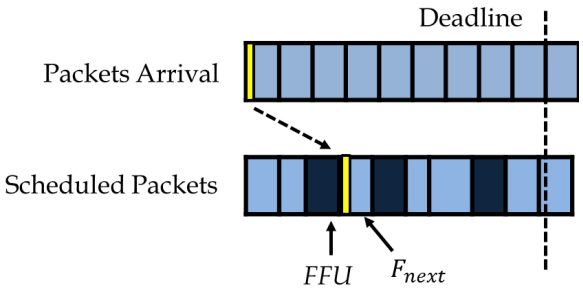


Fig. 1. The shortest delay based scheduling.

Therefore, we can reduce the number of the status transitions by scheduling packets successively and save energy consumption in the status transitions. The longest delay based scheduling is shown in Fig. 2.

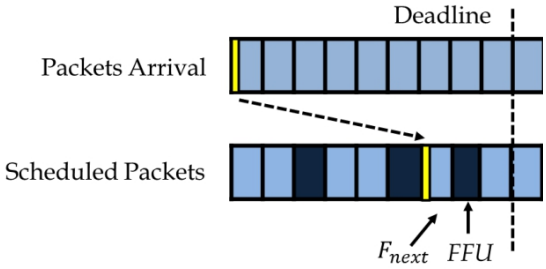


Fig. 2. The longest delay based scheduling.

In Fig. 3, we schedule the packets of connection 1 with the QoS requirement of UGS, and connection 2 with the QoS requirement of RT-VR in an MSS. With the shortest delay based, we schedule the first packet of connection 1 which is $C_{1,1}$. There is no *FIU* or *FFU* in the available frames under this delay constraint. In the third step of the second part in our proposed algorithm, we schedule $C_{1,1}$ into the 5th frame with the maximum delay without violating the constraints, and the 5th frame becomes *FIU*. After that, $C_{1,2}$ is scheduled into *FIU* which is the 5th frame according to the first step in the second part of our algorithm. $C_{1,3}$ and $C_{1,4}$ are also scheduled into *FIU*, which is the 5th frame by the first step in the second part of our algorithm. The 6th packet is scheduled into the 10th frame because there is no *FIU* or *FFU* within D_1 for $C_{1,6}$. The 10th frame becomes *FIU* after $C_{1,6}$ is scheduled inside. The rest packets of connection 1 are scheduled in the same way as are done in previous steps. When we schedule connection 2, the first packet will be scheduled into the 6th frame because there is no *FIU*, while the 5th frame is *FFU*. By the second step in the second part of the algorithm, the F_{next} is the 6th frame. Thus, we schedule $C_{2,1}$ into the 6th frame and $C_{2,2}$ is scheduled into *FIU*, which is the 6th frame. $C_{2,3}$ and $C_{2,4}$ are scheduled into the 9th and 14th frame, respectively. The longest delay based scheduling is shown in Fig. 4. In the result of our examples, our SSS algorithm will schedule the packets into the time slots successively and reduce the number of status transitions in an MSS and minimize the energy consumption of status transitions.

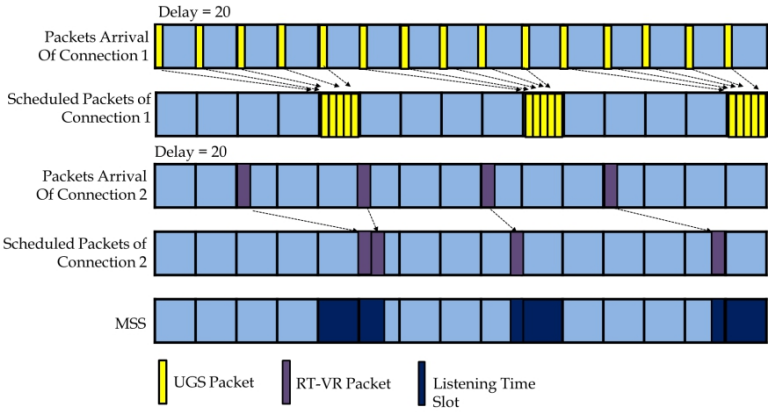


Fig. 3. Our SSS algorithm on the shortest delay based scheduling.

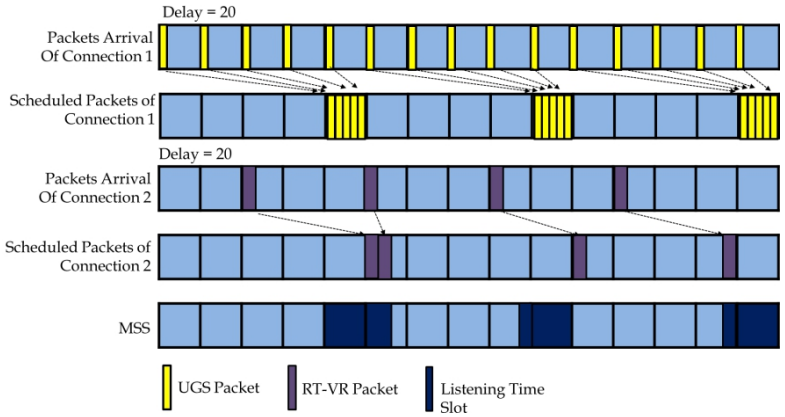


Fig. 4. Our SSS algorithm on the longest delay based scheduling.

2.3.2 The QoS requirements for jitter

In IEEE802.16e broadband wireless access networks, acrucial component of delay is the buffered packet delay between BS and MSS. Due to varying delays in transmission, the delays of scheduling from packet to packet may cause buffered packet delay. This phenomenon is called jitter (Wu & Chen, 2004).

As shown in Fig. 5, we denote $Packet_i$ as the i^{th} packet of certain connections, with the QoS requirement of delay having 7 time slots and 2 jitters. Assume $Packet_{i-1}$ was scheduled in the first time slot, and the delay of $Packet_{i-1}$ is 0. $Packet_i$ may schedule into the time slots of the 2nd time slot to the 8th time slot if we only consider the delay constraint of the QoS requirement. However, it is more realistic to consider the jitter constraint of the QoS requirement. Because the delay of $Packet_{i-1}$ and $Packet_i$ cause jitter, we need to consider the delay of $Packet_i$ to satisfy the jitter constraint. Assume we schedule $Packet_i$ in the 5th time slot, the delay of $Packet_i$ is 3 and the jitter will also 3, and this violates the jitter constraint. Thus, under the jitter

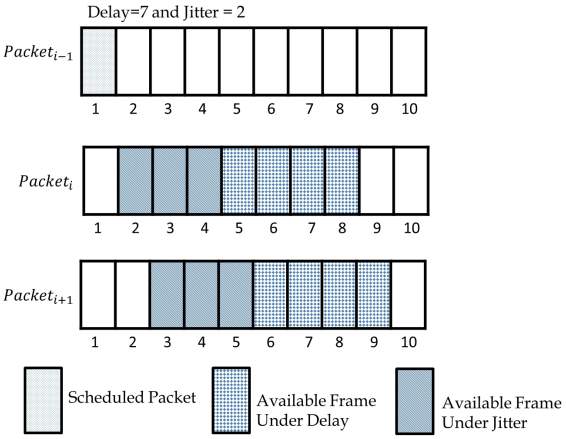


Fig. 5. An example of jitter.

constraint, $Packet_i$ may only schedule into the time slots of the 2nd time slot to the 4th time slot. Assume we schedule $Packet_i$ into the 2nd time slot, $Packet_{i+1}$ may only schedule into the time slots of the 5th time slot to the 7th time slot under the jitter constraint. Thus, the previous approaches to power-saving scheduling with QoS may cause the transmission failure when the jiter constraint is not considered.

QoS requirements include the delay and jitter constraints in scheduling packets. However, previous studies focused on delay constraint without considering the effect of the jitter. Therefore, we take the jitter constraint into account in the scheduling algorithm. In Fig. 6, the first packet was scheduled into the 4th frame which is *FIU* (Tsao & Chen, 2008). Thus, the first packet's delay is 17 and satisfies the delay constraint. The second packet is scheduled into the 4th frame, which is *FIU*. The delay of the second packet is 8 and the jitter between the first and second packet is 9, which satisfies the jitter constraint. The third packet is scheduled into the 9th frame, according to the priorities of the frames. If there is no *FIU*, the first priority will be the frame which has the maximum delay. Therefore, the delay of the third packet would be 20 and the jitter between the second and third packet would be 18, which violates the jitter constraint. Once the scheduling violates the jitter constraint, the QoS is no longer guaranteed.

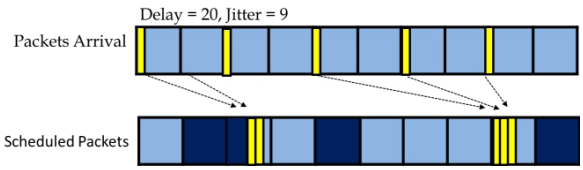


Fig. 6. Example of the scheduling approach (Tsao & Chen, 2008) without considering jitter.

The algorithm of our proposed successive scheduling, which considers jitter constraints, is described in the following two parts. In the first part of our algorithm, the scheduler sorts all connections on an MSS by their delay constraints, and schedules these connections with tight delay requirements. The reason for scheduling connections with tight delay

requirements first is to not violate their QoS requirements, as we mentioned previously. The second part of our algorithm is composed of three steps we described in the Section 2.3.1. In addition to these, the scheduler examines the difference in the delay between the present packet and the previous packet when scheduling each packet in each step. The difference in the delay between the present packet and the previous packet can be viewed as jitter. The scheduler schedules the packets to be earlier or later and into the proper time slots in order to satisfy the jitter constraints.

An example of our algorithm is represented in Fig. 7. The first packet is scheduled into the 4th frame, which is *FIU* within D_1 for $C_{1,1}$. Thus, the delay of $C_{1,1}$ is 17. $C_{1,2}$ is scheduled into the 4th frame, which is *FIU* and with a delay of $C_{1,2}$ being 8. Thus, the jitter between $C_{1,1}$ and $C_{1,2}$ is 9, which satisfies the jitter constraint. $C_{1,3}$ is scheduled into the 5th frame according to our algorithm of successive scheduling scheme and with a delay 4. The jitter between $C_{1,2}$ and $C_{1,3}$ is 4, which is smaller than a jitter constraint of 9. $C_{1,4}$ is scheduled into the 7th frame, with a delay of zero time slots and satisfies the jitter constraint of 9 between $C_{1,3}$ and $C_{1,4}$. $C_{1,5}$ is scheduled into the 9th frame with a delay of 4 and the jitter between $C_{1,4}$ and $C_{1,5}$ being 4. Therefore, in order to provide the QoS guarantees of packets scheduling, we need to satisfy the delay and the jitter constraints.

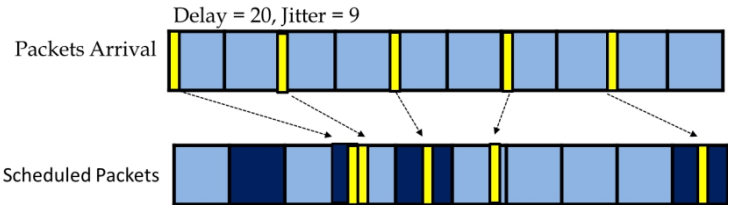


Fig. 7. Example of our SSS algorithm with jitter constraint.

3. Simulation results

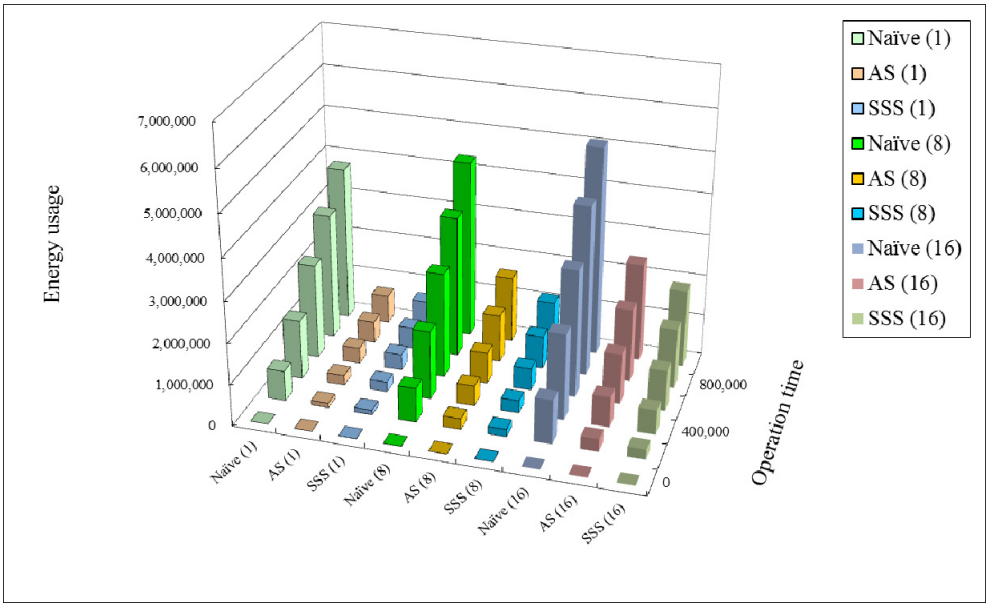
This section evaluated the power consumption of an MSS in terms of the number of listen time slots and status transitions. The QoS requirements of *A*, *B*, *C*, and *D* are listed in Table 2. Both connection types *A* and *B* are VoIP connections. Both connection types *C* and *D* are video streaming connections. The first four connection types on the top half of the list are real-time connections that do not consider the tolerated jitter, and the last four connection types are the same as the first four connection types, but with constrained tolerated jitter. The total energy of an MSS is 1,000,000 units. We compare our proposed SSS algorithm with the Naïve approach without optimizations and the AS approach (Tsao & Chen, 2008). The Naïve approach implies that each connection associates with its preferred type of power-saving class and parameters, and minimizs that packet delay and power consumption for that single connection.

Fig. 8 shows the operation time and energy usage of an MSS by applying three different scheduling schemes in the different connection types with a varied number of connections without the jitter constraints. In the Naïve approach, the energy usage increases faster than the other two approaches. Because the Naïve approach does not consider the optimization of packet scheduling, it results in the enormous energy consumption in status transitions. The energy usage in the AS approach performs the same as our SSS approach when there is

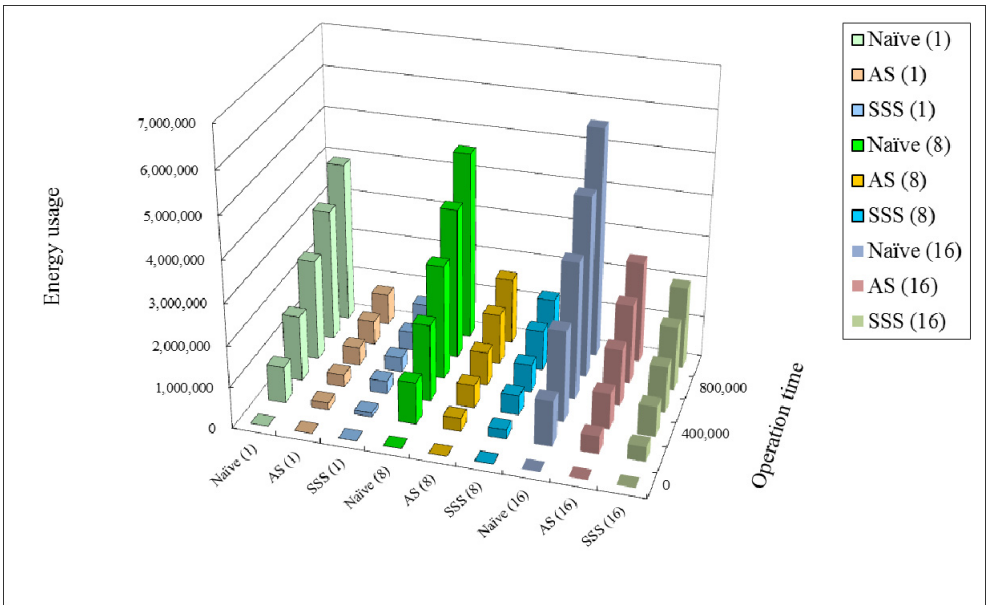
only one connection in an MSS. This is because the two approaches maximize the delay of packets scheduling and schedule the packets into minimal listen periods. However, since we consider status transitions in scheduling the packets, our SSS approach chooses successive frames in scheduling the packets and reducing the number of status transitions. When more connections take into account the scheduling, our SSS approach reduces the number of status transitions by successive scheduling. In other words, while successive time slots are scheduled with packets, they do not place the status transitions in the time slots. Thus, our SSS algorithm saves energy and prolongs the operation time in an MSS.

	Service type of QoS	Packets size (bytes)	Interval of packets arrival (ms)	Delay constraint (ms)	Tolerated jitter (ms)
A	UGS	32	10	50	∞
B	UGS	128	10	50	∞
C	RT-VR	512	30	100	∞
D	RT-VR	1024	30	100	∞
A′	UGS	32	10	50	10
B′	UGS	128	10	50	10
C′	RT-VR	512	30	100	20
D′	RT-VR	1024	30	100	20

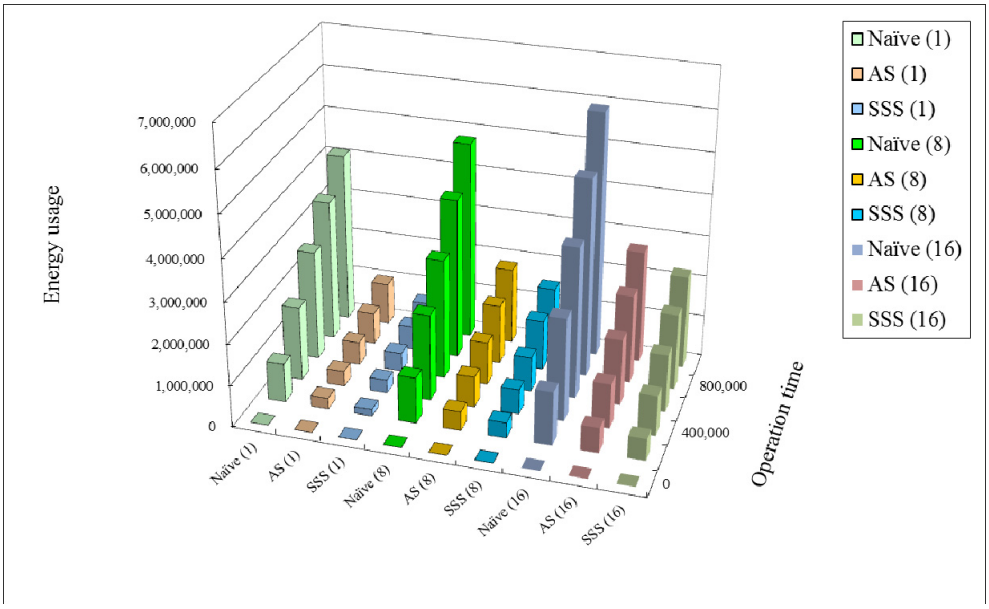
Table 2. QoS parameters of four real-time connections.



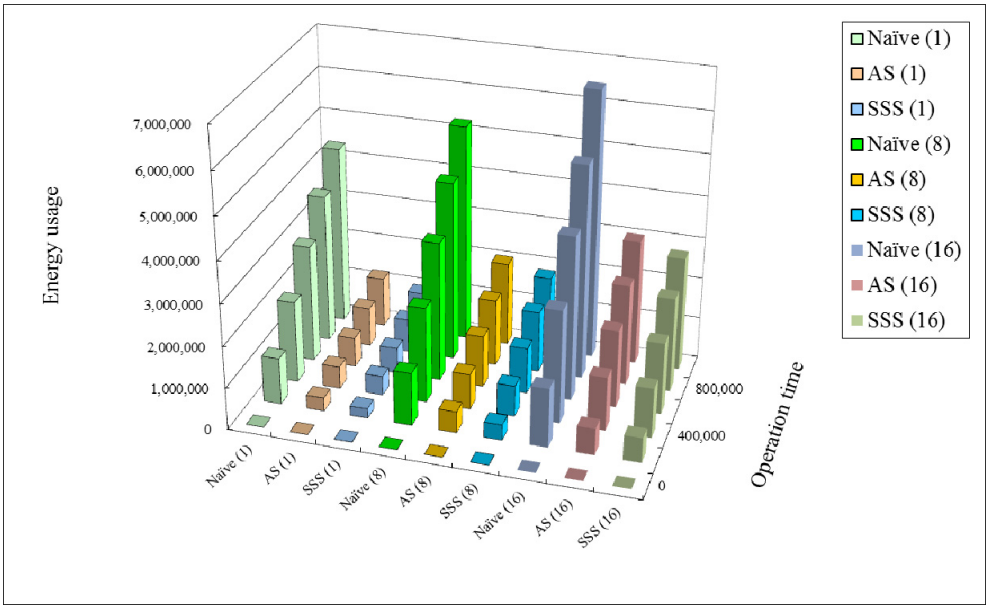
(a)


Energy usage
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
Naïve (1)
AS (1)
SSS (1)
Naïve (8)
AS (8)
SSS (8)
Naïve (16)
AS (16)
SSS (16)
800,000
400,000
0
Operation time

(b)


Energy usage
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
Naïve (1)
AS (1)
SSS (1)
Naïve (8)
AS (8)
SSS (8)
Naïve (16)
AS (16)
SSS (16)
800,000
400,000
0
Operation time

(c)



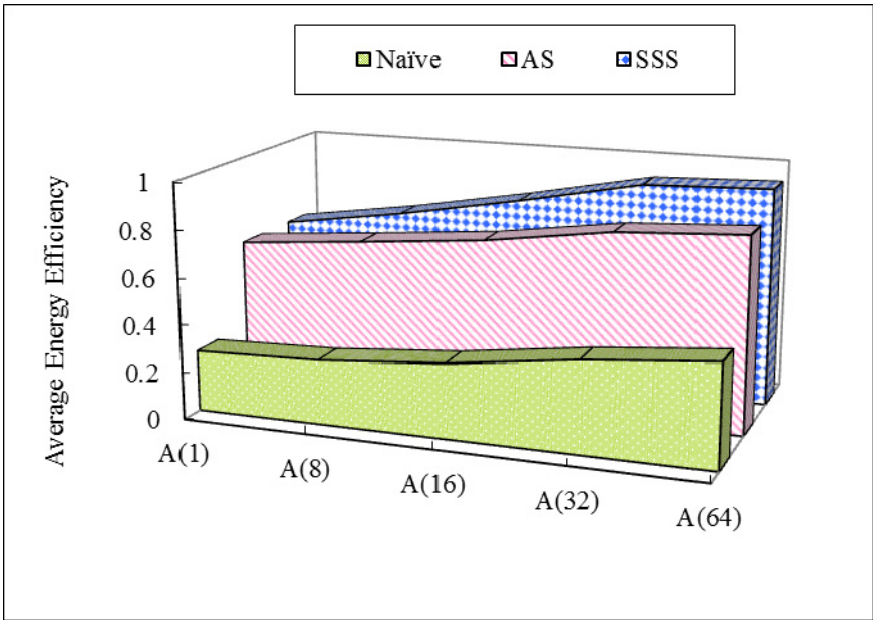
(d)

Fig. 8. The operation time and energy usage of an MSS for three schemes with four connection types with a varied number of connections: (a) connection type *A*, (b) connection type *A*+*B*, (c) connection type *A*+*B*+*C*, and (d) connection type *A*+*B*+*C*+*D*.

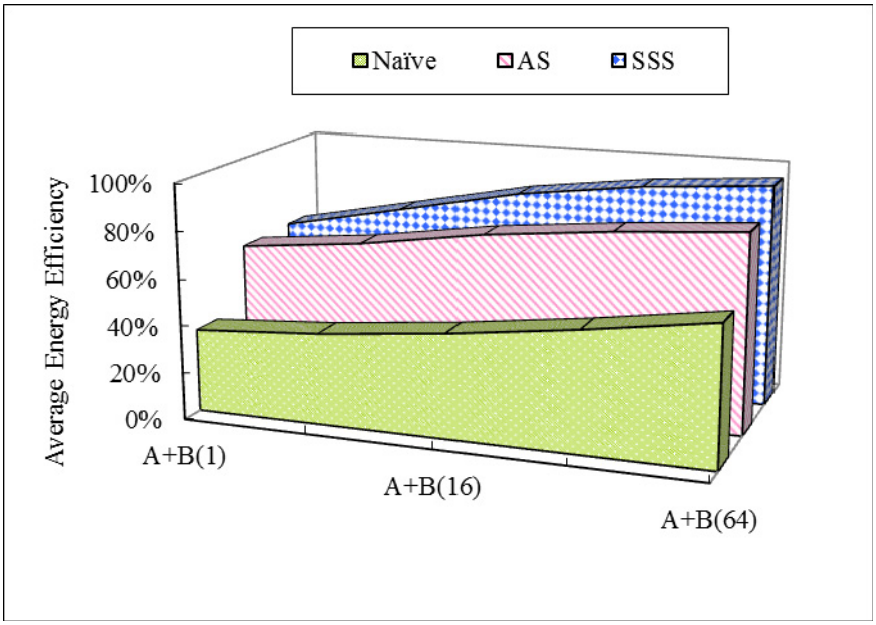
Fig. 9 shows the average energy efficiency of an MSS by applying three different scheduling schemes for different connection types with a varied number of connections without the jitter constraints. We defined E_{trans} as the energy usage for the packet transmission of an MSS during a time period T; E_{total} represents the total energy usage in an MSS during T. The average energy efficiency (*AEE*) for an MSS during T can be represented as follows:

$$AEE = E_{trans} / E_{total} \quad (2)$$

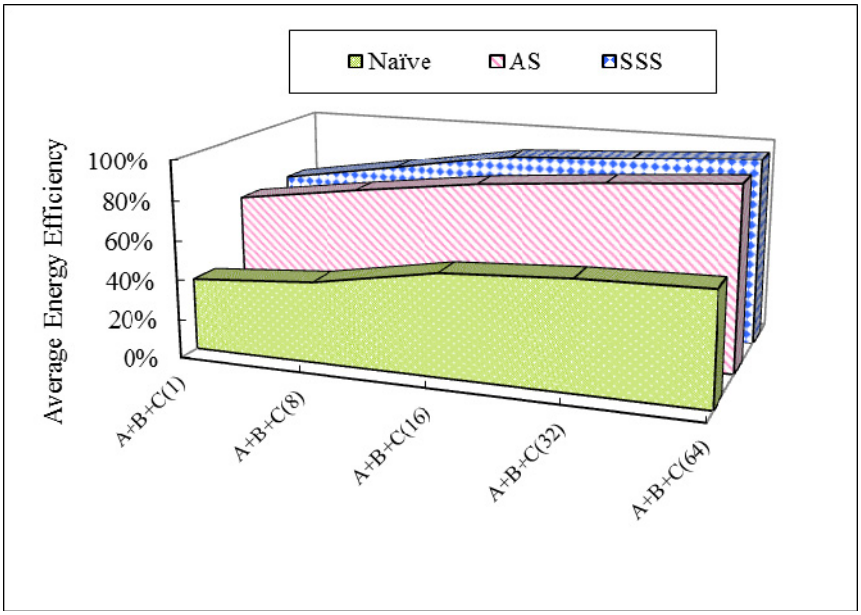
In the Naïve approach, the average energy efficiency is lower than the other two approaches. This is because the Naïve approach processes packets immediately when they arrive, so number of status transitions increase enormously. The energy for status transitions reduce the energy usage for packet transmission from the total energy usage in an MSS. In our SSS algorithm, the average energy efficiency performed the same as the AS approach, where there is only one connection in an MSS. The reason for this is the same as the previous simulation matrix. When there is only one connection in an MSS, the two approaches maximize the delay in packet scheduling and schedules the packets into their minimal listen periods without violating the delay constraints. Thus, the number of status transitions is the same. However, the average energy efficiency in our SSS approach grows up when the number of connections increases. This is because the packets are scheduled more successively when the packets are small, and the number of connections grows large under our proposed algorithm. Fig. 9(c) and (d) reveal that, when the transmission loading encounters a bottleneck, the average energy efficiency stops increasing.


Naïve
AS
SSS
Average Energy Efficiency
1
0.8
0.6
0.4
0.2
0
A(1)
A(8)
A(16)
A(32)
A(64)

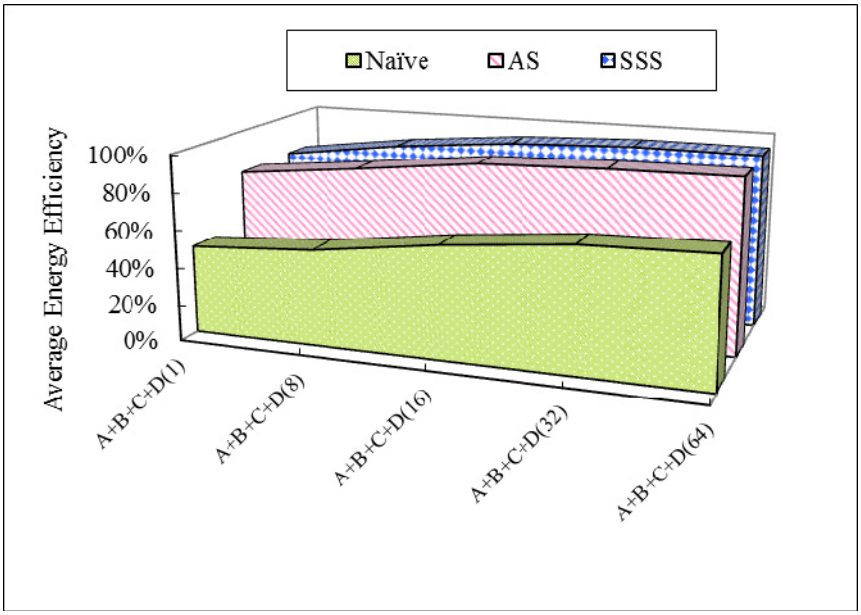
(a)


Naïve
AS
SSS
Average Energy Efficiency
100%
80%
60%
40%
20%
0%
A+B(1)
A+B(16)
A+B(64)

(b)



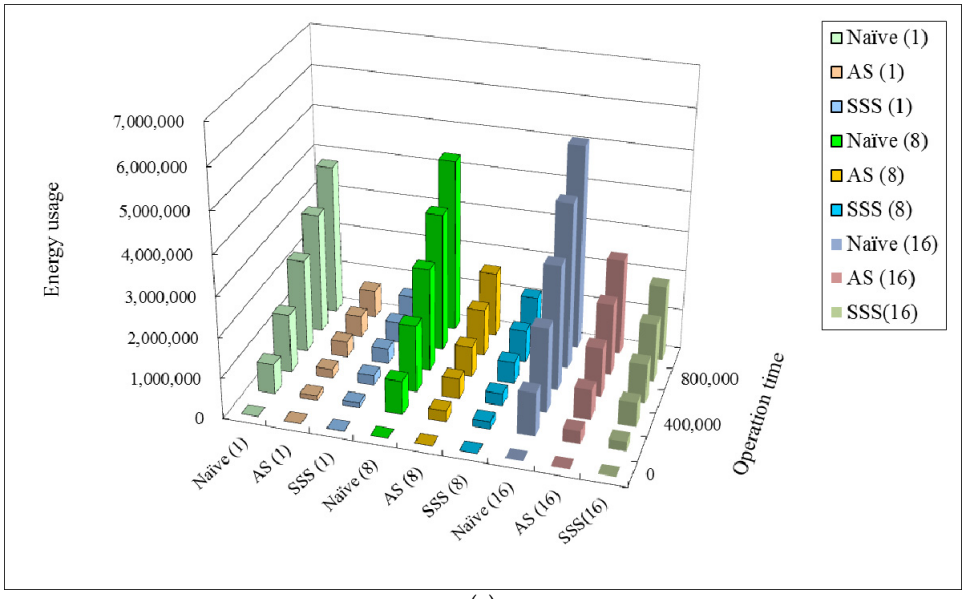
(c)



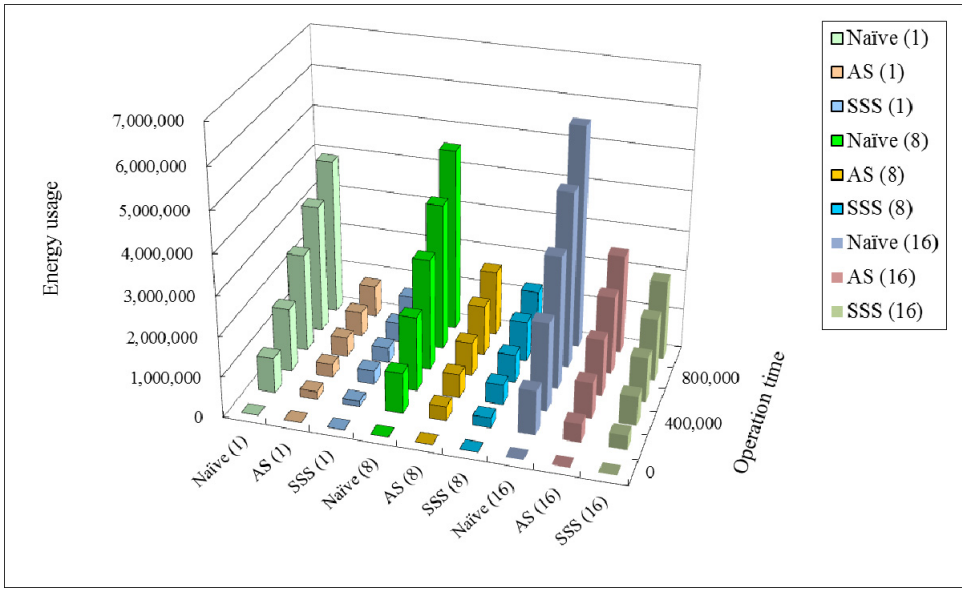
(d)

Fig. 9. The average energy efficiency of an MSS with three schemes and four connection types with a varied number of connections: (a) connection type *A*, (b) connection type *A*+*B*, (c) connection type *A*+*B*+*C*, and (d) connection type *A*+*B*+*C*+*D*.

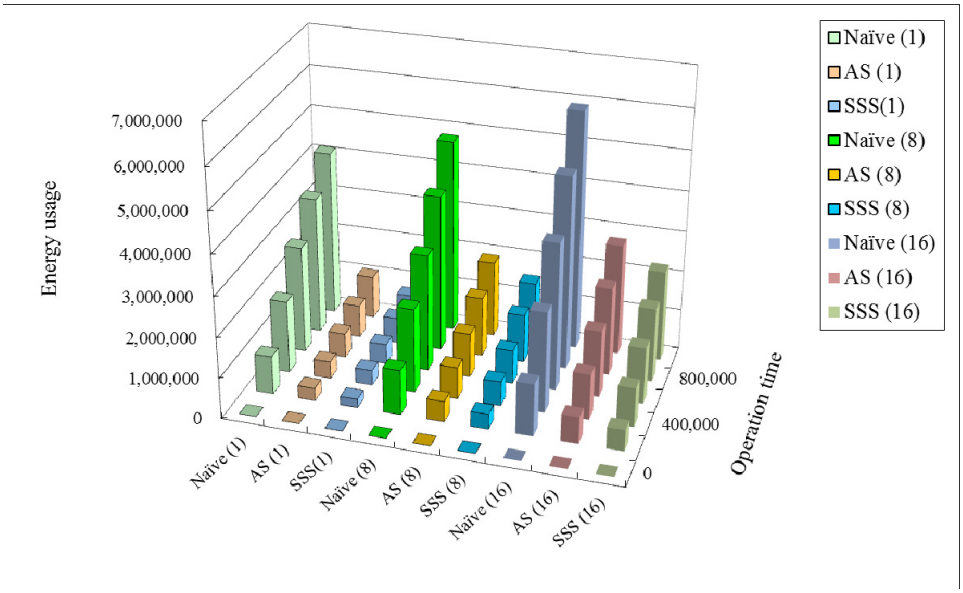
Fig. 10 shows the operation time and energy usage of an MSS by applying three different scheduling schemes for different connection types with a varied numbers of connections with the jitter constraints. The energy usage of different three approaches is higher than the one that does not consider the jitter constraints. The reason for this is that the process is limited



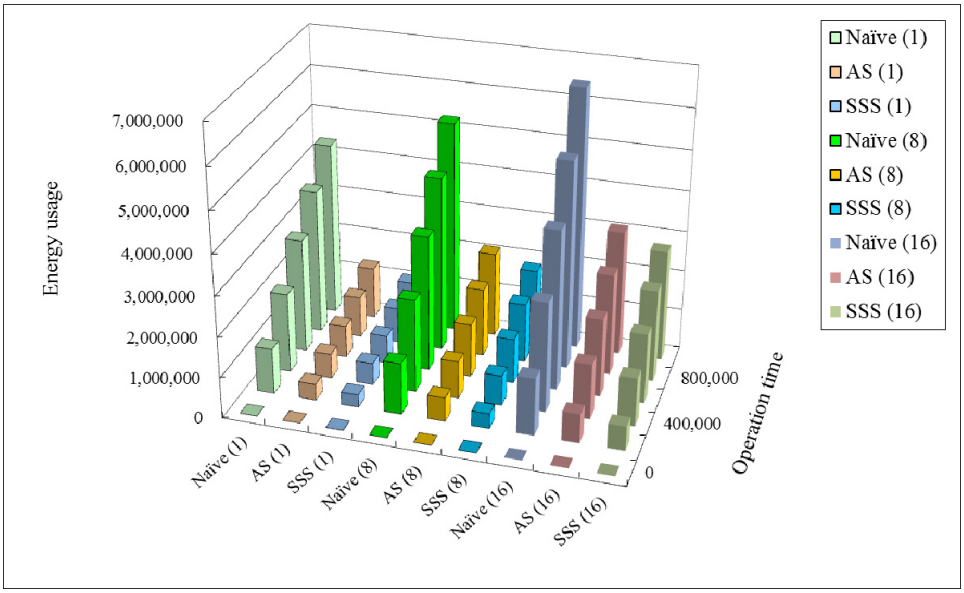
(a)



(b)



(c)

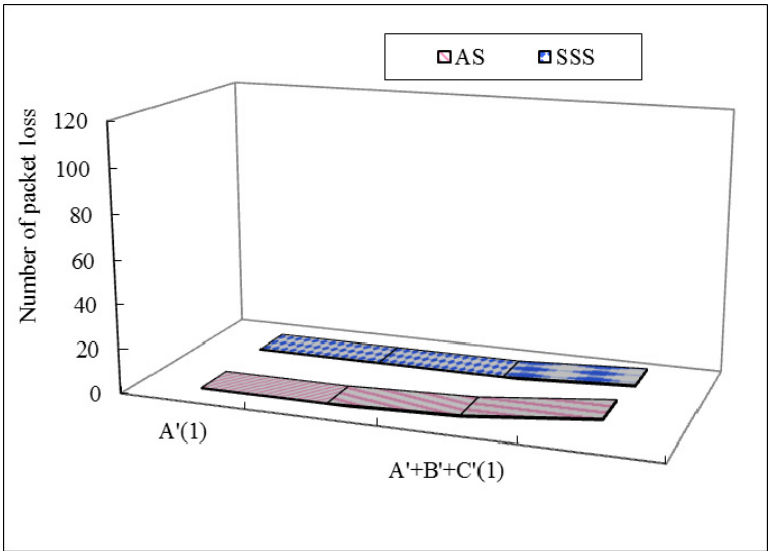


(d)

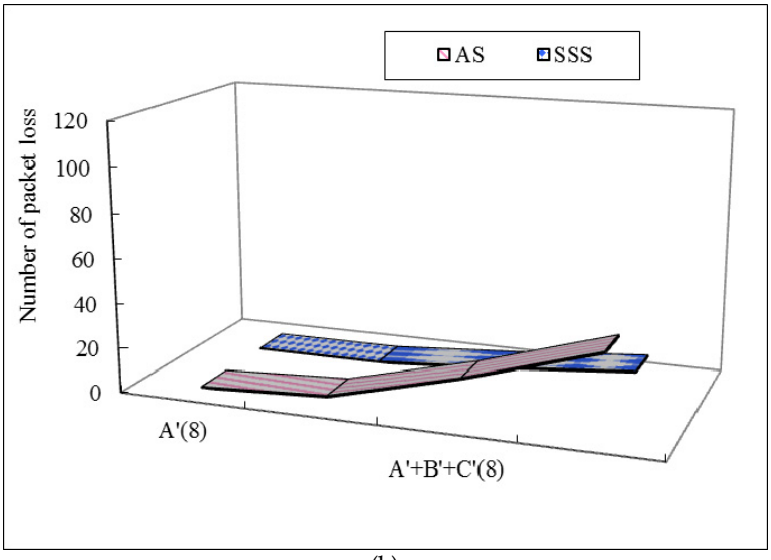
Fig. 10. The operation time of an MSS with three schemes and four connection types with a varied number of connections with jitter constraints: (a) connection type *A'*, (b) connection type *A'*+*B'*, (c) connection type *A'*+*B'*+*C'*, (d) connection type *A'*+*B'*+*C'*+*D'*.

by the jitter constraints, and the limited scheduling increases the number of status transitions. In our SSS approach, the energy usage is lower than the other two approaches under the same connection types. That is because the more connections gain the more chances to be scheduled successively, so the energy consumption of status transitions is reduced.

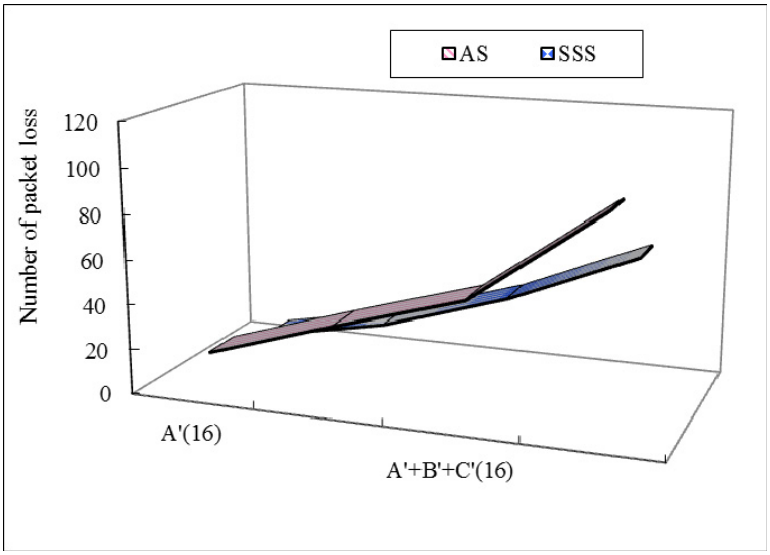
Fig. 11 shows the amount of packet loss of an MSS which applies two different scheduling schemes for different connection types with a varied number of connections with the jitter



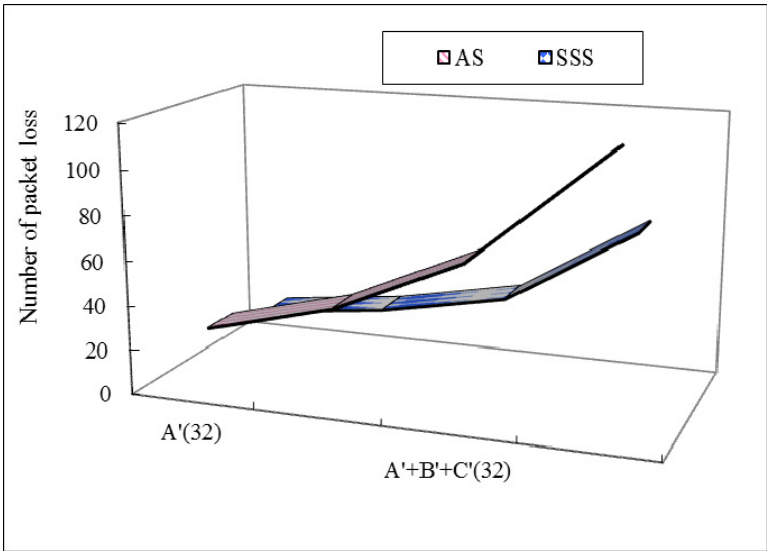
(a)



(b)



(c)



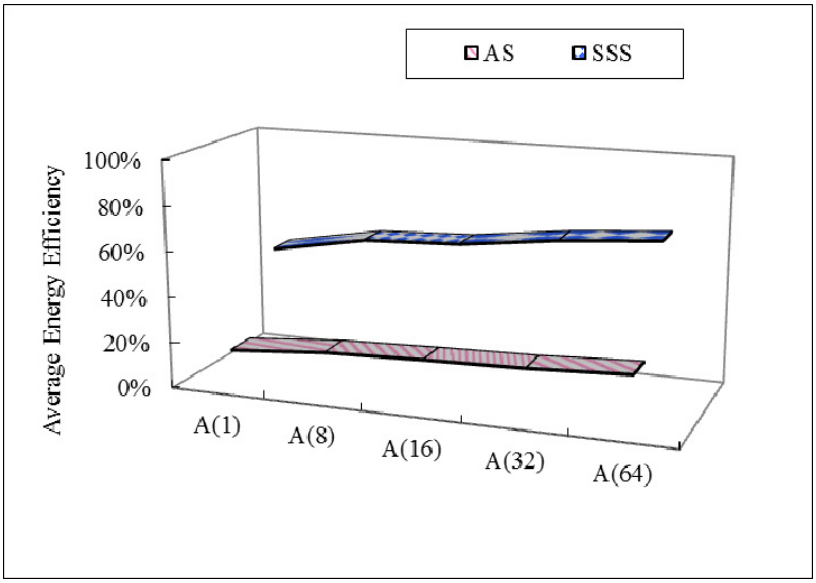
(d)

Fig. 11. The amount of packet loss of an MSS with two schemes and four connection types with a varied number of connections: (a) with 1 connection, (b) with 8 connections, (c) with 16 connections, and (d) with 32 connections.

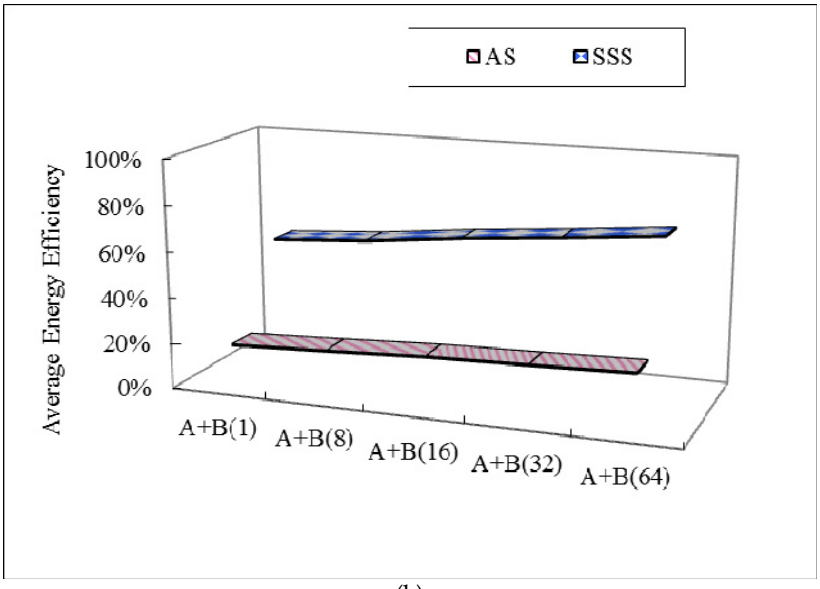
constraints. We only compare the SSS and the AS approaches, which delay the packets, when processing the scheduling. The amount of packet loss is increased when the packet load is raised. In our SSS approach, the number of packet loss is minimized by the algorithm

that considers jitter constraints. The scheduler chooses the proper time slots to schedule the packets in order so as not violate the jitter constraints between each packet.

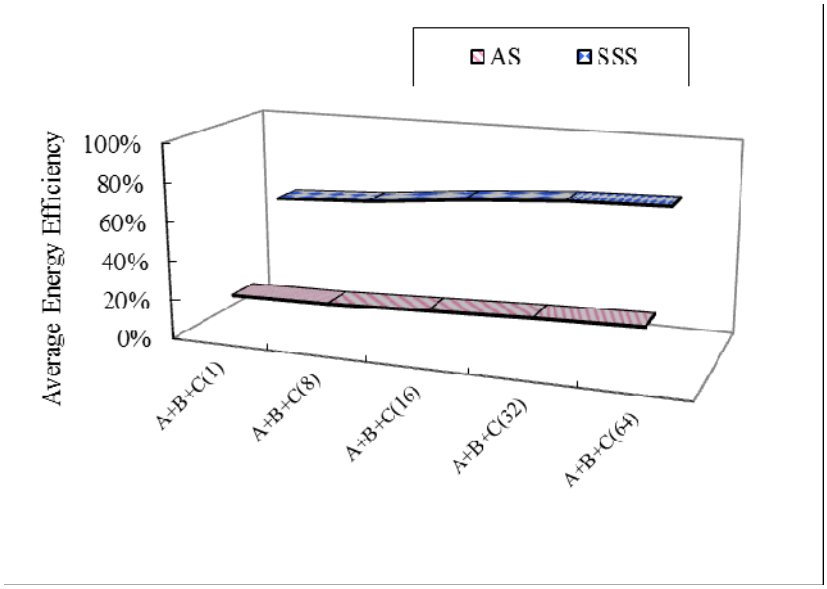
Fig. 12 shows the average energy efficiency of an MSS by applying two different scheduling schemes for different connection types with a varied number of connections with jitter



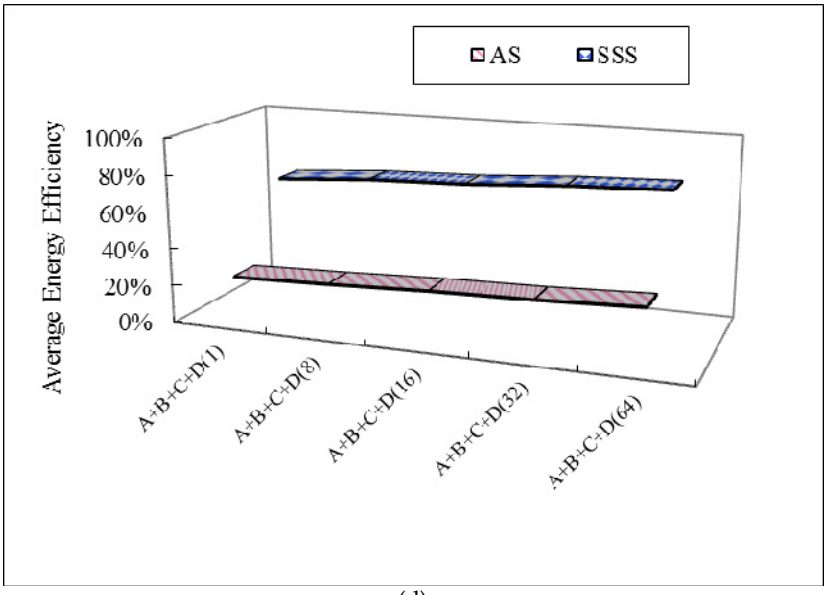
(a)



(b)



(c)



(d)

Fig. 12. The average energy efficiency of an MSS with two schemes and four connection types with a varied number of connections under jitter constraints: (a) connection type *A'*, (b) connection type *A'*+*B'*, (c) connection type *A'*+*B'*+*C'*, and (d) connection type *A'*+*B'*+*C'*+*D'*.

constraints. In this simulation, we only compare the SSS and AS approaches, which delay the packets when processing the scheduling. In our SSS algorithm, the average energy efficient is better than the performance of the AS approach. Due to QoS constraints, the available time slots for scheduling was limited by the delay and jitter constraints. Aside from the energy usage of status transitions, the packets will not be delivered if the scheduling violates the delay and jitter constraints. Meanwhile, the AS does not take the jitter constraints into account when they scheduling the packets. Thus, our SSS approach transmits more packets than the AS, and the average energy efficient in our SSS approach is better than the AS.

4. Conclusion

An energy-efficient scheduling scheme to improve the energy efficiency and guarantee Quality of Service in IEEE 802.16e was proposed. The previous literature only considers the delay constraint of QoS requirement in one MSS. We first consider both the jitter and delay constraints of QoS requirement to schedulethe real-time connections in one MSS. Our proposed algorithmis to schedule the packet transmission in successively fashion with the minimal interval of listen periods and maximal interval of sleep periods without violating the QoS of all connections in an MSS. Additionally, the successive scheduling of time slots would reduce the number of status transitions between the sleep periods and listen periods. The proposed approach can be adapted to the power-saving class of type III where the length of sleep and listen periods arevariable. Simulation results show that, incomparison with the AS and Naïve schemes, the proposed SSS scheduling algorithm can result in a significant overall energy saving and can guarantee the delay and jitter QoS.

5. References

Andrews, M.; Qian, L. & Stolyar, A. (2005). Optimal Utility Based Multi-user Throughput Allocation Subject to Throughput Constraints, *IEEE INFOCOM*, pp. 2415- 2424, 2005.

Fang, G.; Dutkiewicz, E.; Sun, Y.; Zhou, J.; Shi, J. & Li, Z. (2006). Improving Mobile Station Energy Efficiency in IEEE 802.16e WMAN by Burst Scheduling, *IEEE International Conference on Global Telecommunications Conference (GLOBECOM'06)*, pp. 1-5, 2006.

Han, K. & Choi, S. (2006). Performance Analysis of Sleep Mode Operation in IEEE 802.16e Mobile Broadband Wireless Access Systems, *IEEE 63rd Vehicular Technology Conference (VTC 2006-Spring),* pp. 1141-1145, 2006.

Huang, S.-C.; Jan, R.-H. & Chen, C. (2007). Energy Efficient Scheduling with QoS Guarantee for IEEE 802.16e Broadband Wireless Access Networks, *International Conference on Wireless Communications and Mobile Computing (IWCMC'07)*, pp. 547-552, 2007.

IEEE Std 802.16-2004 (2004). IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

IEEE Std 802.16e-2005 (2005). IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2005.

Jang, J.; Han, K. & Choi, S. (2006). Adaptive Power Saving Strategies of IEEE 802.16e Mobile Broadband Wireless Access, *Asia-Pacific Conference on Communications (APCC'06)*, pp. 1-5, 2006.

Lei, K. & Tsang, D.H.K. (2006). Performance Study of Power Saving Classes of Type I and II in IEEE 802.16e, *IEEE Conference on Local Computer Networks*, pp. 20-27, 2006.

Li, B.; Qin, Y.; Low, C.P. & Gwee, C.L. (2007). A Survey on Mobile WiMAX, *IEEE Communications Magazine*, Vol. 45, No. 12, pp. 70-75, Dec. 2007.

Liao, W.-H. & Wang, H.-H. (2008). An Asynchronous MAC Protocol for Wireless Sensor Networks, *Journal of Network and Computer Applications*, Vol. 31, No. 4, pp. 807-820, 2008.

Liu, M. & Liu, M. T. (2003). A Power-Saving Scheduling for IEEE 802.11 Mobile Ad Hoc Network, *IEEE International Conference on Computer Networks and Mobile Computing(ICCNMC 2003)*, pp.238 – 245, 2003.

Mukherjee, S.; Leung, K.K. & Rittenhouse, G. E. (2005). Protocol and Control Mechanisms to Save Terminal Energy in IEEE 802.16 Networks, *IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing (PACRIM)*, pp. 5-8, 2005.

Seo, J.B.; Lee, S.Q.; Park, N.H.; Lee, H.W. & Cho, C.H. (2004). Performance Analysis of Sleep Mode Operation in IEEE 802.16e, *IEEE 60th Vehicular Technology Conference, 2004 (VTC2004-Fall)*, pp. 1169-1173, 2004.

Tian, L.; Yang, Y.; Shi, J.; Dutkiewicz, E. & Fang, G. (2007). Energy Efficient Integrated Scheduling of Unicast and Multicast Traffic in 802.16e WMANs, *IEEE International Conference on Global Telecommunications Conference (GLOBECOM'07)*, pp. 3478-3482, 2007.

Tsao, S.-L. & Chen, Y.-L. (2008). Energy-Efficient Packet Scheduling Algorithms for Real-Time Communications in a Mobile WiMAX System, *Computer Communications*, Vol. 31, No. 10, pp. 2350-2359, 2008.

Tseng, Y.-C.; Hsu, C.-S. & Hsieh, T.-Y. (2002). Power-Saving Protocols for IEEE 802.11-Based Multi-Hop Ad Hoc Networks, *IEEE INFOCOM*, pp. 200-209, 2002.

Wu, E. H.-K. & Chen, M.-Z. (2004). JTCP: Jitter-Based TCP for Heterogeneous Wireless Networks, *IEEE Journal on Selected Areas in Communications*, Vol. 22, No. 4, pp. 757-766, May 2004.

Wongthavarawat, K. & Ganz, A. (2003). Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems, *International Journal of Communication Systems*, Vol. 16, No. 1, pp. 81-96, Feb. 2003.

Ye, W.; Heidemann, J. & Estrin, D. (2004). Medium Access Control with Coordinated Adaptive Sleeping for Wireless Sensor Networks, *IEEE/ACM Transactions on Networking*, Vol. 12, No. 3, pp. 493-506, Jun. 2004.

Zheng, T.; Radhakrishnan, S. & Sarangan, V. (2005). PMAC: An Adaptive Energy-Efficient MAC Protocol for Wireless Sensor Networks, *IEEE International Parallel and Distributed Processing Symposium (IDPDS'05)*, pp. 65-72, 2005.

Zhu, H. & Cao, G. (2004). A Power-Aware and QoS-Aware Service Model on Wireless Networks, *IEEE INFOCOM*, pp. 1393-1403, 2004.

3

A Fast Handover Scheme for WiBro and cdma2000 Networks

Choongyong Shin, Seokhoon Kim and Jinsung Cho
Kyung Hee University
South Korea

1. Introduction

The continuous development of wireless communication technologies yields diverse wireless networks that are widely deployed and successfully serviced according to their communication capabilities. Representative examples are cellular networks (e.g., cdma2000 and UMTS) capable of wide-area coverage and WLAN (Wireless LAN) efficiently used in public hot spots. WLAN and cellular network are complementary technologies. WLAN has several advantages over cellular networks, including higher data rate and lower operating and equipment costs. However, their coverage is typically limited to corporate buildings, residence, and certain public hot spots. On the other hand, cellular networks provide wide-area coverage but at lower speeds and much higher cost. It is indispensably required to integrate WLAN and cellular networks to serve users who need both high-speed wireless access as well as wide-area connectivity (Salkintzis, 2004).

Integrating heterogeneous networks reveals a lot of difficulties due to their different system specifications, standardization, and service scopes. For this reason, most of research work (3GPP, 2003; Ahmavaara et al., 2003; Buddhikot et al., 2003; Luo et al., 2003) are focused on interworking mechanism between network elements rather than integrating whole network architectures. Recently, with the intention of overcoming the limitation of existing wireless networks, new wireless communication service, WiBro is being deployed in Korea. WiBro is based on IEEE 802.16e (IEEE, 2001; Koffman & Roman, 2002) and is similar with Mobile WiMAX (Forum, n.d.). It is expected to provide enough mobility (60km/h) and higher data rate (50Mbps). So the study of the integration between WiBro and cdma2000 will give better effects than the existing works of the integration between WLAN and cdma2000.

In order to provide seamless services across heterogeneous wireless networks, efficient handoff procedure as well as flexible integrated network architecture is essentially required. As for handoff procedure, it is possible to use Mobile IP (Perkins, 2002) which is generally used in for homogeneous network or its extended version, so called low-latency handoff which tends to reduce packet loss and delay during the handoff (Maki, 2004). However, since these handoff procedures exploit L3 (layer 3) signaling messages, they have a problem that packet loss and delay can occur while processing L3 messages. In this paper, we propose an efficient L2 (layer 2) handoff scheme between cdma2000 and WiBro networks. The proposed L2 handoff scheme takes advantages over the existing L3 handoff scheme because it exploits L2 messages instead of L3 messages. We show the efficiency of our proposed L2 handoff through extensive computer simulations.

The paper is organized as follows: Section 2 summarizes existing techniques for handoff scheme as related work. Section 3 presents the proposed L2 handoff scheme and we describe the performance characteristics of the proposed scheme through OPNET simulation in Section 4. Section 5 concludes the paper.

2. Handoff in heterogeneous wireless networks

Since the integration of widely deployed 3G cdma2000 and WLAN can give a lot of benefits to both end users and service providers, there has been a lot of researches about interworking mechanism between cdma2000 and WLAN (3GPP, 2003; Ahmavaara et al., 2003; Buddhikot et al., 2003; Luo et al., 2003; Salkintzis, 2004). The WLAN and 3G cdma2000 integration architecture is characterized by the amount of interdependence it introduces between the two component networks. Two candidate integration architectures, tightly-coupled and loosely-coupled interworking are described in Fig. 1.

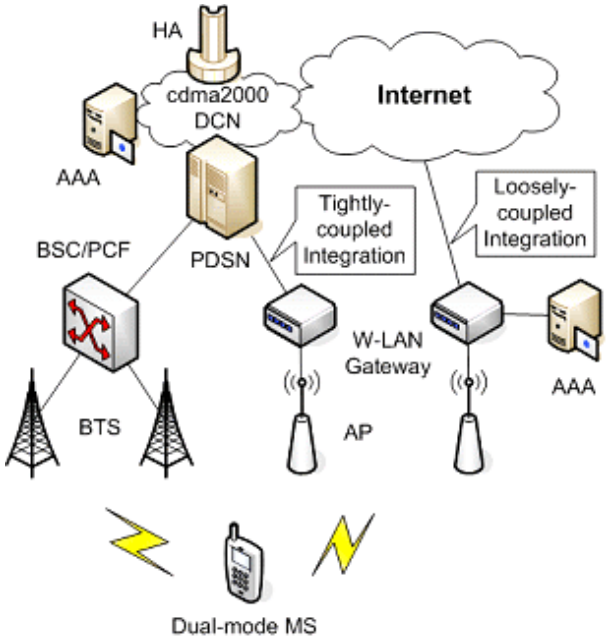
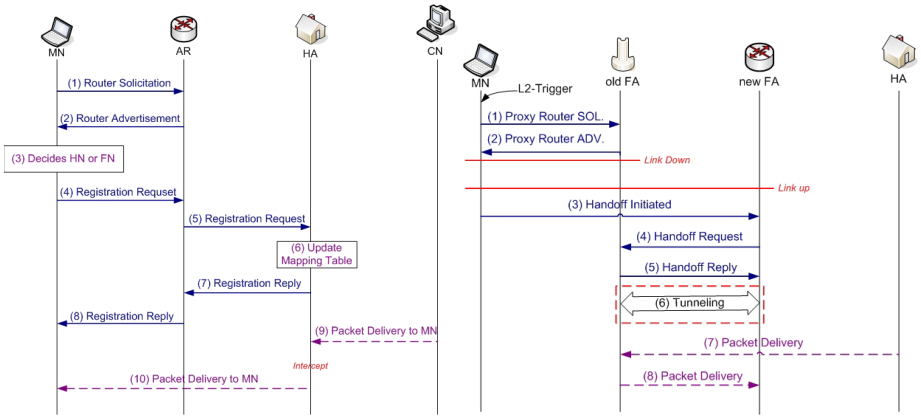


Fig. 1. Interworking architecture of heterogeneous wireless networks

In the tightly-coupled interworking approach, WLAN networks appear to 3G core network as another 3G RAN (Radio Access Network). The WLAN gateway hides the details of the WLAN network to the 3G core network, and implements all the 3G protocols. Even though this approach can share the same authentication, signaling, and billing infrastructures, independent from physical layer interface, it has a disadvantage that the capacity and configuration of each network element is carefully reengineered and can result in high cost. On the contrary, the loosely-coupled interworking approach has several advantages such that 3G networks and WLAN can be independently deployed without extensive capital investments and its implementation is relatively easy. Therefore, it has emerged as a preferred



(a) Mobile IP (b) Low-latency handoff for Mobile IP

Fig. 2. L3 handoff procedure

architecture for the integration of WLAN and 3G networks. As previously mentioned, the L3 handoff procedure is commonly used to provide mobility services in the loosely coupled interworking architecture. Mobile IP is the representative technology of L3 handoff to provide mobility services within homogeneous 3G cdma2000 or between 3G cdma2000 and WLAN. When mobile station moves into new wireless network, Mobile IP performs registration procedure and handoff are completed after registration procedure as shown in Fig. 2(a). The low-latency handoff for Mobile IP (Maki, 2004) is proposed to reduce packet loss occurred in the registration procedure. It reduces packet loss by providing tunneling and buffering between the previous and new networks as depicted in Fig. 2(b). Since the current cdma2000 networks employ Mobile IPv4, we deal with Mobile IPv4 and its extended version in this paper. So the low latency handoff for Mobile IPv4 in Fig. 2(b) will be used in performance evaluation of Section 4. In IPv6 environment, however, fast handoff for Mobile IPv6 (Koodli, 2005) can be considered similarly.

3. Proposed L2 handoff scheme

3.1 Overview of L2 handoff

For wireless communication, mobile terminal performs step-by-step layered procedures. First, when it is initially booted or located in new wireless network area, it scans L1 signal. As soon as it detects L1 signal, it performs L2 connection procedure. After successful L2 connection, mobile terminal can communicate or send/receive packets. The L3 handoff described in Section 2 initiates handoff after L2 connection. Our proposed L2 handoff procedure exploits L2 signaling messages transmitted during L2 connection setup. By acquiring packet flow path between network elements (e.g. PDSN and ACR in Fig. 3) while processing L2 messages, our method reduces packet loss occurred in handoff.

The proposed L2 handoff procedure considers the interworking network architecture as shown in Fig. 3. The cdma2000 and WiBro networks are loosely-coupled integrated through interworking between PDSN of cdma2000 and ACR of Wibro. So in exception of interworking of PDSN and ACR, each network is working and servicing independently. Our L2 handoff

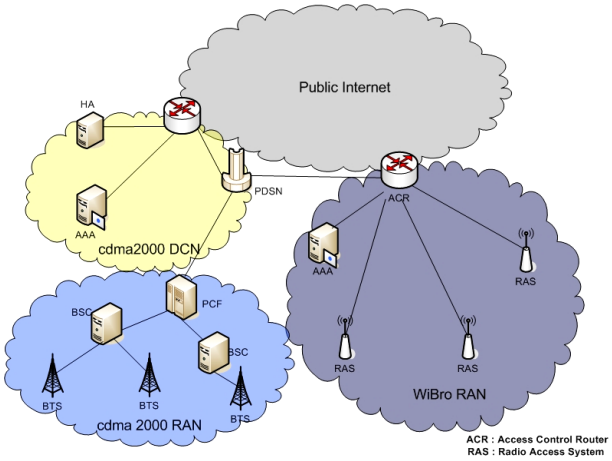


Fig. 3. Interworking Architecture between cdma2000 and WiBro networks

procedure exploits L2 signaling messages commonly existing in cdma2000 and WiBro. It does not require additional signaling messages so that our scheme can be implemented with ease.

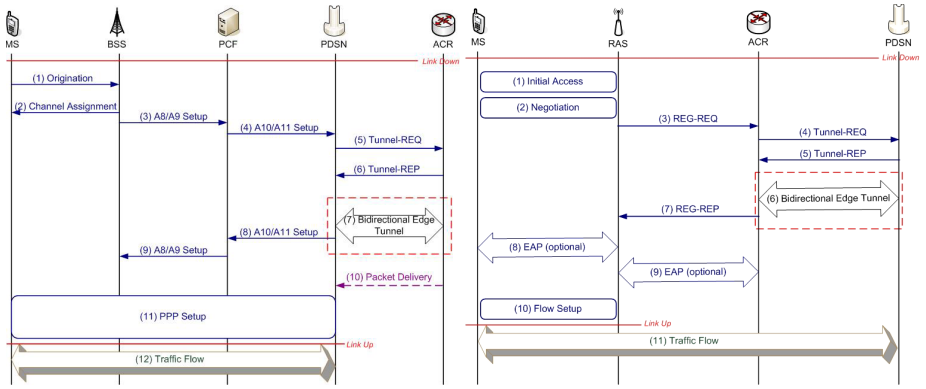
3.2 L2 handoff procedure

In this section, we describe the proposed L2 handoff procedure. We consider two handoff scenarios. The first case is that the mobile station is moving from cdma2000 cellular network into WiBro. The second one is the reverse case. We do not consider the cases that handoff is occurring within its own wireless network, i.e., cdma2000 or WiBro. We also do not consider the initial call setup procedures for its own wireless network, since it is followed as described in standardization of each wireless network.

Fig. 4 describes the proposed L2 handoff procedure. When the mobile station moves into a new network area, it processes L2 connection procedure. At this time, handoff information is transmitted to the network through the Origination message of cdma2000 or L2 REG-REQ message of WiBro as shown in (1) of Fig. 4(a) and (1) of Fig. 4(b), respectively. More specifically, the Origination message includes PANID (Previous Access Network ID). If the cdma2000 system receives the Origination message which contains PANID=ANID of WiBro, the cdma2000 system regards it as a vertical handoff from WiBro network. Similary, the REG-REQ message of WiBro can contain PANID=ANID of cdma2000 which means a vertical handoff from cdma2000 network. With such handoff information, the source PDSN or ACR detects the occurrence of handoff and extracts the target ACR or PDSN address for the handoff. Based on this address information, PDSN and ACR requests each other and generates tunnel for the handoff traffic. After the tunnel is setup, the corresponding PDSN or ACR is buffering packets destined to the mobile station while the requested L2 connection is being made.

3.3 Standardization issues

Consider a mobile station moves from WiBro to cdma2000 networks as depicted in Fig. 4(a). When MS requests a communication channel to BSS, a bearer path MS-BSSPCF-PDSN-ACR



(a) WiBro to cdma2000 L2 handoff (b) cdma2000 to WiBro L2 handoff

Fig. 4. L2 handoff procedure

should be setup, where ACR is the service anchor point in WiBro network. The origination message in (1) of Fig. 4(a) contains PANID field so that PDSN can connect to the appropriate anchor ACR using the field (3GPP2, 2002). To do this, a mapping function from the base station ID in WiBro network to ANID in cdma2000 network is required. A possible solution may be as follows: construct 48-bits base station ID in WiBro with SID (16 bits), NID (16 bits), PZID (8 bits), and base station number in a packet zone (8 bits), where SID, NID, PZID comprise ANID in cdma2000 network.

Next, let us consider a mobile station moves from cdma2000 network to WiBro in Fig. 4(b). Similarly to the aforementioned case, when MS requests a communication channel to RAS, a bearer path MS-RAS-ACR-PDSN should be setup, where PDSN is the service anchor point in cdma2000 network. In order for ACR to connect to the right PDSN, PANID field should be delivered to ACR via RAS. It can be implemented by adding PANID filed in MAC management messages of WiBro standard specifications.

With this slight modification of standard specifications, the proposed L2 handoff scheme can be implemented as explained in this section. In addition, the fast handoff mechanism between PDSN and ACR will provide seamless services on vertical handoff. In the next section, we validate its performance.

4. Experimental evaluation

4.1 Simulation model

The OPNET simulation, as shown in Fig. 5, has been conducted to examine the performance of the proposed scheme. We assume that there are 135 mobile stations used in the simulation and the traffic parameters are set as in Table 1.

$$f_{init}(v) = \begin{cases} k\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(v-m)^2}{2\sigma^2}} \\ 0, v < 0 \end{cases}, v \geq 0 \quad (1)$$

	Conversational	Streaming	Interactive	Background
Max. Bit Rate (Mb/s)	2.4 (EV-DO), < 2 (WiBro)	2.4 (EV-DO), < 2 (WiBro)	2.4 (EV-DO), < 2 (WiBro)	2.4 (EV-DO), < 2 (WiBro)
Max Packet Size (byte)	$\leq$ 1500 or 1502	$\leq$ 1500 or 1502	$\leq$ 1500 or 1502	$\leq$ 1500 or 1502
Packet Error Ratio	10^{-2}, $7*10^{-3}$,10^{-3}, 10^{-4}, 10^{-5}	10^{-1}, 10^{-2}, $7*10^{-3}$,10^{-3}, 10^{-4}, 10^{-5}	10^{-3}, 10^{-4}, 10^{-6}	10^{-3}, 10^{-4}, 10^{-6}

Table 1. Simulation parameters

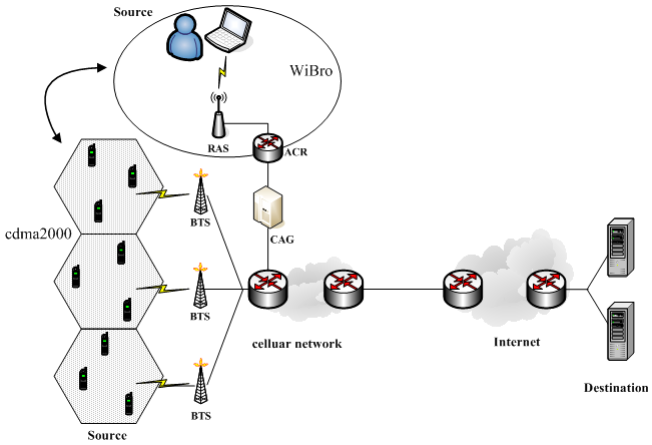


Fig. 5. Simulation model for performance evaluation

Fig. 7 shows only a small part of the entire model for simplicity although there are a lot of cdma2000 and WiBro cells. WiBro network cells and cdma2000 network cells are attached one by another in the simulation. Picocells and microcells are only used to generate frequent handoffs of mobile stations. Since the Markov mobility model used in the simulation, as shown in Fig. 8, is designed for mobile stations at low-speed (20 60km/h), and the following probability density function is used, where m represents the average speed of a mobile station in a cell (Janevski, 2003).

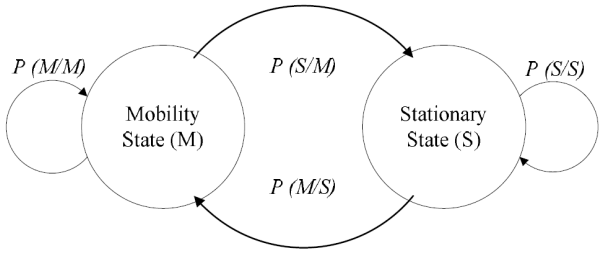


Fig. 6. Mobility model for mobile stations

4.2 Simulation results

Both L3 and L2 handoff schemes are simulated to prove the superiority of the proposed L2 handoff over L3 handoff on most popular services in mobile environment, such as streaming, web browsing, and Email services. These services are categorized into streaming, interactive, and background traffic class, respectively. In addition, conversational class is also added for video conferencing environment. As mentioned earlier, the low latency handoff for Mobile IPv4 in Fig. 3 has been implemented for the L3 handoff scheme.

Fig. 7 - 10 show that the proposed L2 handoff scheme outperforms the L3 handoff on all kinds of service classes. In fact, performances resulted in each scheme should be the same except when handoffs occur. Therefore, performance differences shown in Fig. 9 - 12 are due to handoff processes. Figures also show that handoff occurrence is very frequent at interval times of 6 to 23, 32 to 35, and 55 to 57.

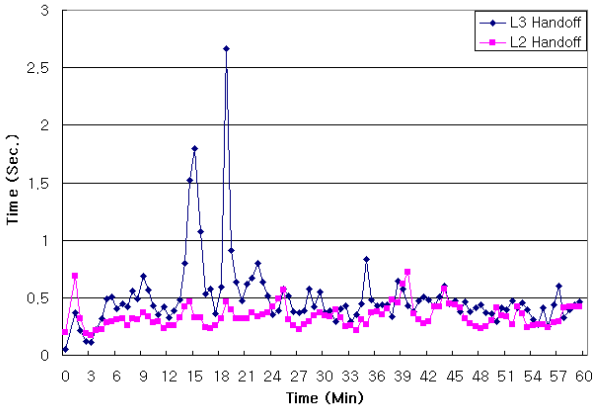


Fig. 7. Packet delay on conversational traffic class

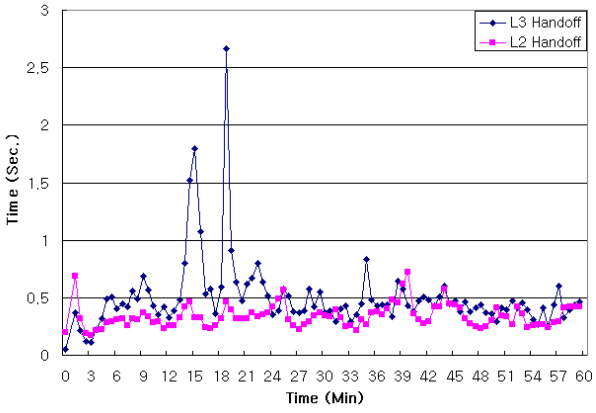


Fig. 8. Packet delay on streaming traffic class

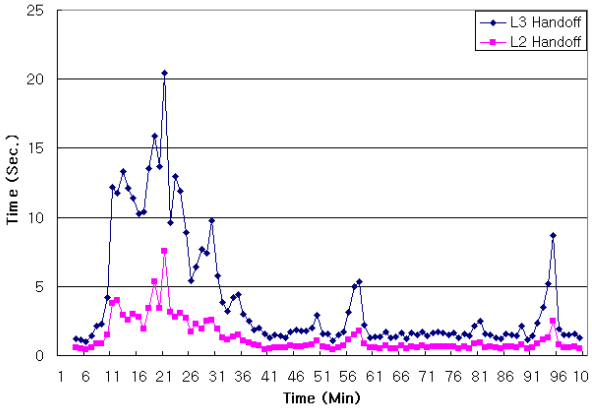


Fig. 9. Packet delay on interactive traffic class

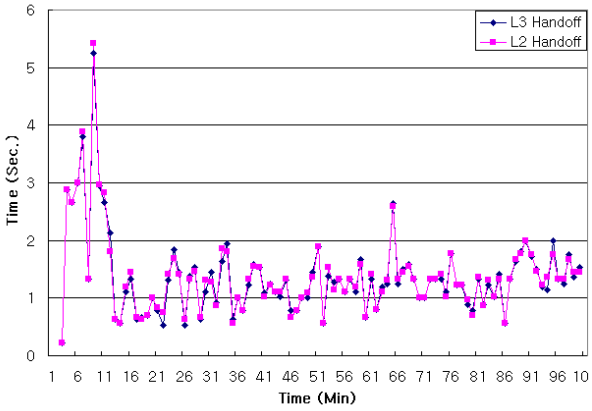


Fig. 10. Packet delay on background traffic class

Fig. 7 and Fig. 8 are the simulation results of conversational and streaming traffic classes, respectively. The graph of the L2 handoff scheme is more stable with small deviation than the L3 handoff because the L2 handoff reduces packet losses and delays.

Fig. 9 shows the case of interactive traffic class using HTTP, where differences in packet delays between L3 and L2 handoff are larger than those of Fig. 7 and Fig. 8 of the burst property of web traffic. We can find from Fig. 10 that background traffic class like Email shows almost no difference in delay performance because background traffic has the lowest priority.

In addition, we performed simulations with different moving speed of mobile stations. Fig. 11 - 14 show the average delay times of each traffic class of MS at different speed. The graph shows the superiority of the L2 handoff scheme over the L3 handoff scheme on all kinds of traffic classes and moving speeds. We can also find that the faster a mobile station moves,

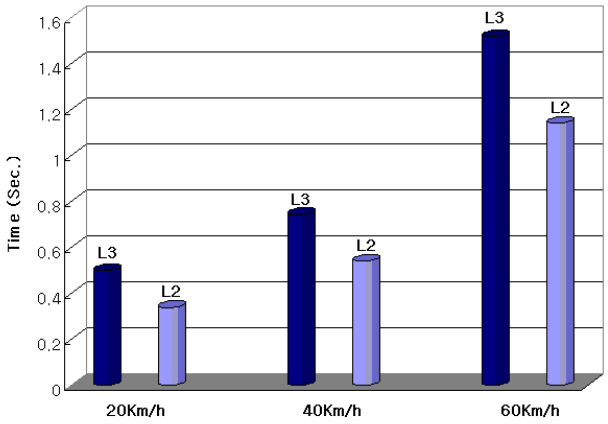


Fig. 11. Average delay according to MS's speed (conversational traffic class)

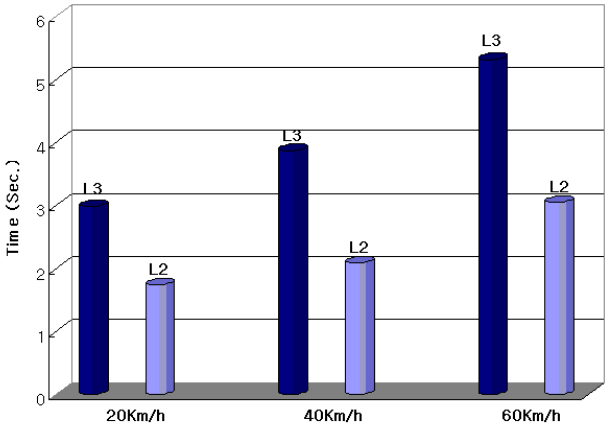


Fig. 12. Average delay according to MS's speed (streaming traffic class)

the larger differences in delay performance arise, because the fast moving causes frequent handoffs. We summarize the average packet delay in Table 2.

Finally, Fig. 15 describes the packet loss ratio in both schemes. Since the L3 handoff scheme employs mobile IP techniques, there may be relatively large number of packet loss.

In summary, the OPNET simulation results in this section indicate that the proposed L2 handoff scheme is an efficient and practical solution because it can be implemented with minimal modification of existing cdma2000 and WiBro networks while providing the reduced packet delay and loss.

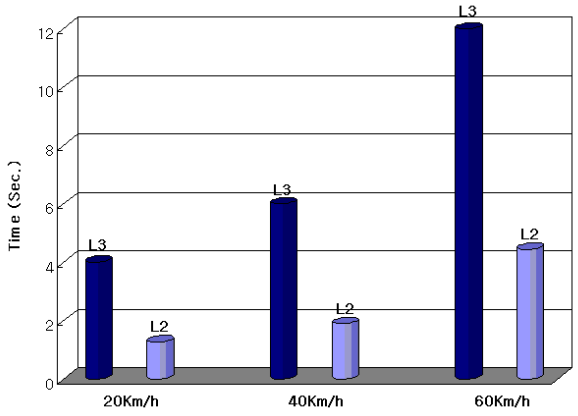


Fig. 13. Average delay according to MS's speed (interactive traffic class)

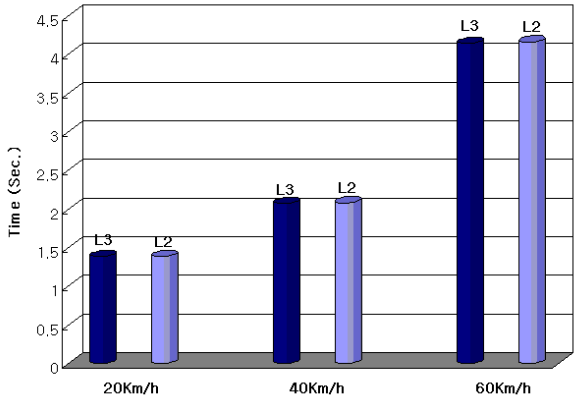


Fig. 14. Average delay according to MS's speed (background traffic class)

Traffic class	Handoff	20Km/h	40Km/h	60Km/h
Conversational	L3	0.495	0.743	1.485
	L2	0.335	0.537	1.117
Streaming	L3	2.953	3.839	5.316
	L2	1.726	2.071	3.020
Interactive	L3	5.829	8.743	17.487
	L2	2.459	3.689	8.608
Background	L3	3.358	4.365	6.549
	L2	3.371	4.382	6.574

Table 2. Summary of average delay according to MS's speed

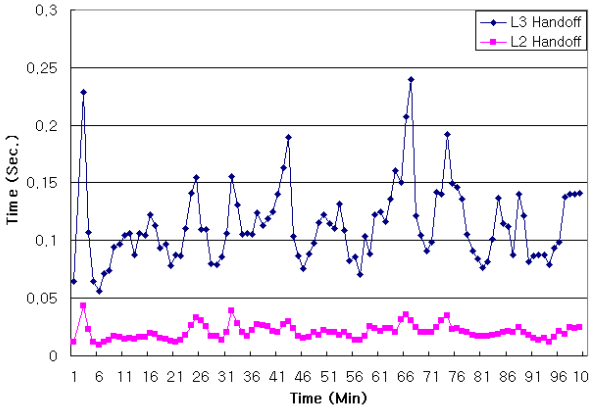


Fig. 15. Packet loss ratio

5. Conclusion

In this paper, we proposed a low-latency L2 handoff procedure between cdma2000 and WiBro which creates a promising next generation wireless network. Even though several efforts are actively in progress to improve mobility services based on Mobile IP, mobility services between different wireless networks, e.g., WiBro and cdma2000 still need more attention. From this viewpoint, we devise an L2 handoff scheme which can provide better performance compared with the L3 handoff. We also define required functionalities of each network element, ACR of WiBro, PDSN of cdma2000, and mobile station.

The proposed L2 handoff procedure does not require additional signaling messages to reduce packet loss which can occur in signaling L3 messages. However, in order to apply our scheme, the necessary functional change of network elements is inevitable. For completion, the detailed protocols above L3, e.g., session control remains to be further studied.

6. References

3GPP (2003). 3gpp tr 22.934, *Feasibility study on 3GPP system to Wireless Local Area Network(WLAN) Interworking (Release 6)* .

3GPP2 (2002). 3gpp2 a.s0011-a, *Interoperability Specification (IOS) for cdma2000 Access Network Interfaces : Part 1 Overview* .

Ahmavaara, K., Haverinen, H. & Pichna, R. (2003). Ieee communications magazine, *Interworking Architecture between 3GPP and WLAN Systems* Vol. 41(No. 11): 74–81.

Buddhikot, M., Chandranmenon, G., Han, S., Lee, Y., Miller, S. & Salgarelli, L. (2003). Ieee communications magazine, *Design and Implementation of a WLAN/cdma2000 Interworking Architecture* Vol. 41(No. 11): 90–100.

Forum, W. (n.d.).
URL: *http://www.wimaxforum.org*

IEEE (2001). Ieee std. 802.16, *IEEE Standard for Local and Metroploitan Area Network Part 16 : Air Interface for Fixed Broadband Wireless Access System* .

Janevski, T. (2003). Artech house, *Traffic analysis and design of wireless IP networks* pp. 186–190.

Koffman, I. & Roman, V. (2002). Ieee communications magazine, *Broadband Wireless Access Solutions based on OFDM access in IEEE 802.16* Vol. 40: 96–103.

Koodli, R. (2005). Ietfrfc 4068, *Fast Handovers for Mobile IPv6* .

Luo, H., Jiang, Z., Kim, B. & Henry, P. (2003). Ieee internet computing magazine, *Integrating Wireless LAN and Cellular Data for the Enterprise* Vol. 7(No. 2): 25–33.

Maki, K. E. (2004). Ietf internet draft, *Low Latency Handoffs in Mobile IPv4* .

Perkins, C. (2002). Rfc3344, *IP Mobility Support for Ipv4* .

Salkintzis, A. K. (2004). Ieee wireless communications magazine, *Interworking Techniques and Architectures for WLAN/3G Integration Toward 4G Mobile Data Networks* Vol. 11(No. 3): 50–61.

4

Design and Analysis of IP-Multimedia Subsystem (IMS)

Wagdy Anis Aziz[1] and Dorgham Sisalem[2]
[1]*Mobinil*
[2]*Tekelec, Berlin*
[1]*Egypt*
[2]*Germany*

1. Introduction

IP Multimedia Subsystem (IMS) has resulted from the work of the Third Generation Partnership Project (3GPP) toward specifying an all-IP communication service infrastructure (24.229 2009). Mainly looking at the needs and requirements of mobile operators, the 3GPP first specified IMS as a service architecture combining the Internet's IP technology and wireless and mobility services of current mobile telephony networks. After that the IMS architecture was extended to include fixed networks as well. By deciding to use session initiation protocol (SIP) as the signaling protocol for session establishment and control in IMS instead of developing its own set of protocols, 3GPP has opened the door toward a tight integration of the mobile, fixed and Internet worlds. Recent reports already indicate that there are more than 200 million subscribers using the IMS technology for telephony services.

In this chapter we provide a theoretical model that can be used by operators and network designers to determine the effects of introducing IMS to their networks in terms of bandwidth usage for example and the effects of losses and delays on the service quality. This model uses as the input various traffic characteristics such as the number of calls per second and mean holding time and network characteristics, such as losses and propagation delays. The output of the model provides details on the bandwidth needed for successfully establishing a session when using SIP over UDP in IMS networks.

Voice traffic in IP Multimedia Subsystem (IMS) will be served using Internet Protocol (IP) which is called Voice over IP (VoIP). This chapter uses the "E-Model", (ITU-T Rec. G.107 2005), as an optimization tool to select network and voice parameters like coding scheme, packet loss limitations, and link utilization level in IMS Network. The goal is to deliver guaranteed Quality of Service for voice while maximizing the number of users served. This optimization can be used to determine the optimal configuration for a Voice over IP in IMS network.

2. Bandwidth calculation for IMS session establishment

2.1 Introduction

By deciding to use session initiation protocol (SIP) as the signaling protocol for session establishment and control in IMS instead of developing its own set of protocols, 3GPP has

opened the door toward a tight integration of the mobile, fixed and Internet worlds. SIP can be used over various transport protocols such as UDP, TCP or SCTP. To enable the reliable transmission of SIP messages even when used over UDP, SIP supports application level retransmission mechanisms. That is in case no response was received for a sent request then after a timeout the request is retransmitted. Thereby, losses due to overloaded servers or lossy links would cause delays in the session establishment and hence reduce the perceived service quality.

In this part of the chapter we provide a theoretical model that can be used by operators and network designers to determine the effects of introducing IMS to their networks in terms of bandwidth usage for example and the effects of losses and delays on the service quality. This model uses as the input various traffic characteristics such as the number of calls per second and mean holding time and network characteristics, such as losses and propagation delays. The output of the model provides details on the bandwidth needed for successfully establishing a session when using SIP over UDP in IMS networks. In Sec. 2.2 we provide the related work to this chapter and present a brief overview of the literature concerning modeling of SIP. In Sec. 2.3 the IMS and SIP in IMS are presented. In Sec. 2.4 the IMS session establishment phases is presented. The SIP model for IMS session establishment is presented in Sec. 2.5.

2.2 Related work

With the success of SIP, there have already been a number of studies addressing aspects of performance evaluation and modeling of SIP. Chebbo et al. describe in (Chebbo et al. 2003) a modeling tool with which it is possible to estimate the number of required SIP entities for supporting certain traffic. Gurbani et al. present in (Gurbani et al. 2005) a theoretical model of a SIP server using queuing theory. This model is then used to evaluate the performance of a SIP server in terms of response time and number of served requests. Wu et al. analyze in (Wu et al.2003) the usage of SIP for carrying telephony information in terms of queuing delay and delay variations.

In general, these studies aim at investigating the performance of SIP servers in terms of the number of SIP sessions that can be supported by a SIP server or the processing delays at such servers. In contrast, in our work we do not aim at modeling the performance of a SIP server but to investigate the performance of SIP in terms of the number of messages and amount of time needed by SIP for establishing a session in lossy environments.

Fathi et al. (Fathi et al.2006) present a model of SIP in VoIP networks and investigate the effects of mobility on the performance of session establishment using SIP. The used model is however rather simplified and is only applicable to stateless SIP proxies which have no notion of transactions. Alam et al. (Alam et al.2005) discuss different performance model for SIP deployment scenarios in mobile networks. This involves providing models for evaluating the performance of push-to-talk applications or the effects of different mobility concepts. The work does not however provide for a model of how SIP itself deals with losses. Sisalem et al. (Sisalem et al. 2008) provided a theoretical model of the effects of losses and delays on the performance of SIP. While that work is providing the basis for our work here, it is rather limited to simple SIP networks as are discussed in IETF. The work in this chapter takes the multi-hop nature of IMS into account as well as the SIP specifications.

2.3 Background

In this section, a description of the IP Multimedia Sub system (IMS) architecture including the function of the key components and SIP function in IMS is also presented.

2.3.1 IMS architecture

3GPP has standardized the IP Multimedia Subsystem specifications (24.229 2009).IETF also collaborates with them in developing protocols that fulfill their requirements. Figure 1 shows the common nodes included in the IMS .These nodes are:

- CSCF (Call/Session Control Function): CSCF is a SIP server which processes SIP signaling in the IMS. There are three types of CSCFs depending on the functionality they provide,
- Proxy Call Session Control Function (P-CSCF)
- Interrogating Call Session Control Function (I-CSCF)
- Serving Call Session Control Function (S-CSCF).
- P-CSCF (Proxy-CSCF): The P-CSCF is the first point of contact between the IMS terminal and the IMS network. All the requests initiated by the IMS terminal or destined to the IMS terminal traverse the P-CSCF.
- I-CSCF (Interrogating-CSCF): It has an interface to the SLF (Subscriber Location Function) and HSS (Home Subscriber Server). This interface is based on the Diameter protocol (Calhoun et al. 2003).
- The I-CSCF retrieves user location information and routes the SIP request to the appropriate destination, typically an S-CSCF.
- S-CSCF (Serving-CSCF): It maintains a binding between the user location and the user's SIP address of record (also known as Public User Identity). Like the I-CSCF, the S-CSCF also implements a Diameter interface to the HSS.
- SIP AS (Application Server): The AS is a SIP entity that hosts and executes IP Multimedia Services based on SIP.
- ENUM (E.164 NUmber Mapping): The ENUM allows telephone numbers to be resolved into SIP URLs using the Domain Name System (DNS) (Faltstrom 2000).
- MRF (Media Resource Function): The MRF provides a source of media in the home network. It is further divided into a signaling plane node called the MRFC (Media Resource Function Controller) and a media plane node called the MRFP (Media Resource Function Processor). The MRFC acts as a SIP User Agent and contains a SIP interface towards the S-CSCF. The MRFC controls the resources in the MRFP via an H.248 interface (ITU-T H.248.1 2005).
- BGCF (Breakout Gateway Control Functions): BGCF a SIP server that includes routing functionality based on telephone numbers.
- SGW (Signaling Gateway): SGW performs lower layer protocol conversion.
- MGCF (Media Gateway Control Function): MGCF implements a state machine that does protocol conversion and maps SIP to either ISUP (ISDN User part) over IP or BICC (Bearer Independent Call Control) over IP. The protocol used between the MGCF and the MGW is H.248 (ITU-T H.248.1 2005)
- MGW (Media Gateway): The MGW interfaces the media plane of the PSTN. On one side the MGW is able to send and receive IMS media over the Real-Time Protocol (RTP) (Schulzrinne et al. 2003)

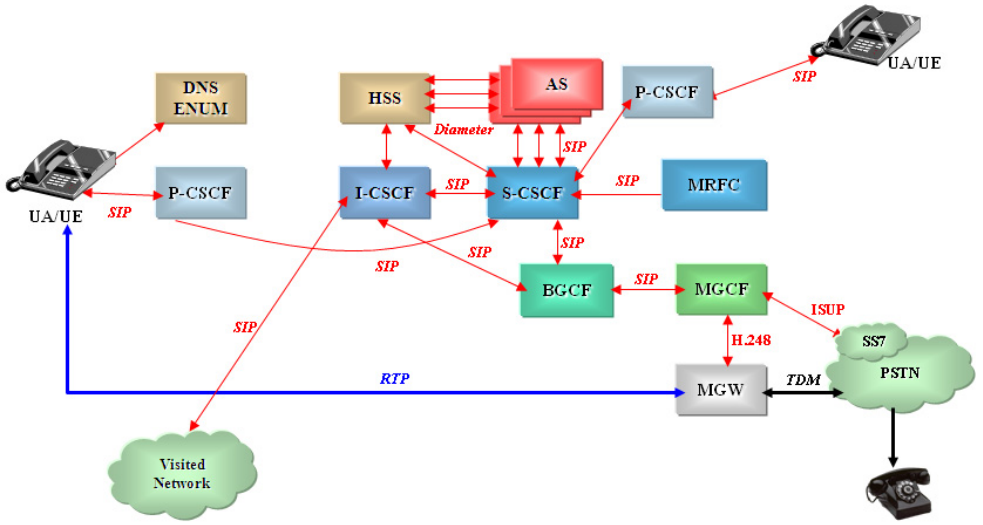


Fig. 1. IMS Functional Elements

- The Home Subscriber Server (HSS) contains all the user related subscription data required to handle multimedia sessions.

2.3.2 SIP in IMS

All IP voice and multimedia call signaling in IMS will be performed by SIP providing a basis for rapid new service introductions and integration with fixed network IP services. With regard to the SIP messages we distinguish between requests and responses. A request indicates the user's wishing to start a session (INVITE request) or terminate a session (BYE request). We further distinguish between session initiating requests and in-dialog requests. The INVITE request used to establish a session between two users is a session initiating request. The BYE sent for terminating this session would be an in-dialog request. Responses can either be final or provisional. Final responses can indicate that a request was successfully received and processed by the destination. Alternatively, a final response can indicate that the request could not be processed by the destination or by some proxy in between or that the session could not be established for some reason. Provisional responses indicate that the session establishment is in progress, e.g., the destination phone is ringing but the user did not pick up the phone yet. A SIP proxy acts in either stateful or stateless mode. In the stateful mode, the proxy forwards an incoming request to its destination and keeps state information about the forwarded request until either a response is received for this request or a timer expires. When used over an unreliable transport protocol such as UDP, if the proxy did not receive a response after some time, it will resend the request. In the stateless mode, the proxy would forward the request without maintaining any state information. In this case the user agent would be responsible for retransmitting the request if no responses were received. SIP uses an exponential retransmission behavior. So if a sender of a SIP message does not receive a response after some time, it will resend the request after some waiting time. In case no response was received for the retransmission, the

sender increases the waiting time and tries again and so up to a certain number of retransmissions, (Rosenberg et al.2002) (29.328 2008).In general one can distinguish between two retransmission modes in SIP:

2.3.2.1 INVITE-retransmissions

This behavior applies for INVITE requests as well as some other messages exchanged during session establishment. If this mode is used then the sender retransmits a message if no confirmation was received after T1 seconds. The retransmission timer is then increased exponentially up to a maximum retransmission timer (Timer B). Once this timer is reached the sender drops the message and stops the retransmission.

2.3.2.2 Non-INVITE-retransmissions

This behavior applies to all requests other than INVITE. In this mode the sender retransmits a message if no confirmation was received after T1 seconds. The retransmission timer is then increased exponentially up to a maximum retransmission timer called T2. Once this timer is reached the sender continues retransmitting the request every T2 up to a maximum timer (TimerF). Once this timer is reached the sender drops the message and stops the retransmission.

2.4 IMS session establishment phases

Figure 2 illustrates a basic session establishment in the IMS with the caller and callee roaming to foreign networks. The example is based on the scenario provided in (Sisalem et al. 2009) .The session establishment in IMS is triggered by the sending of an INVITE request. In general, the session establishment can be considered as consisting of five phases, namely:

- Phase 1: Initiation (INVITE / 100 Trying): This phase is initiated with the sending of an INVITE request and is terminated when the client receives a provisional or final response. (INVITE-retransmission mode).
- Phase 2: Negotiation (Session Progress 183 / PRACK / 200 OK) : In this phase the caller and callee negotiate the audio and video codes to be used as well as the QoS criteria. The 183 provisional responses is sent reliably. Session Progress 183 (INVITE-retransmission mode) , PRACK and 200 OK (Non-INVITE-retransmission mode)
- Phase 3: Confirmation (UPDATE / 200 OK): In this phase the caller and callee complete the code and QoS negotiations. UPDATE and 200 OK (Non-INVITE-retransmission mode)
- Phase 4: Ringing: (Ringing 180 / PRACK / 200 OK): In this phase the callee informs the caller that the user is being alerted about the call. Ringing 180 (INVITE-retransmission mode) , PRACK and 200 OK (Non-INVITE-retransmission mode)
- Phase 5: Final Response (200 OK / ACK): In this phase the callee informs the caller that the call was accepted. 200 OK / ACK (Non-INVITE-retransmission mode)

2.5 Modeling IMS session establishment

In this section we will provide a theoretical model for SIP retransmission techniques in lossy network, bandwidth calculation for IMS session set up and estimation of IMS session set up delay. Figure.2. shows that the calls traverse five SIP proxies. Each link of the depicted network has a loss rate of (l) and has a propagation delay of (D) seconds.

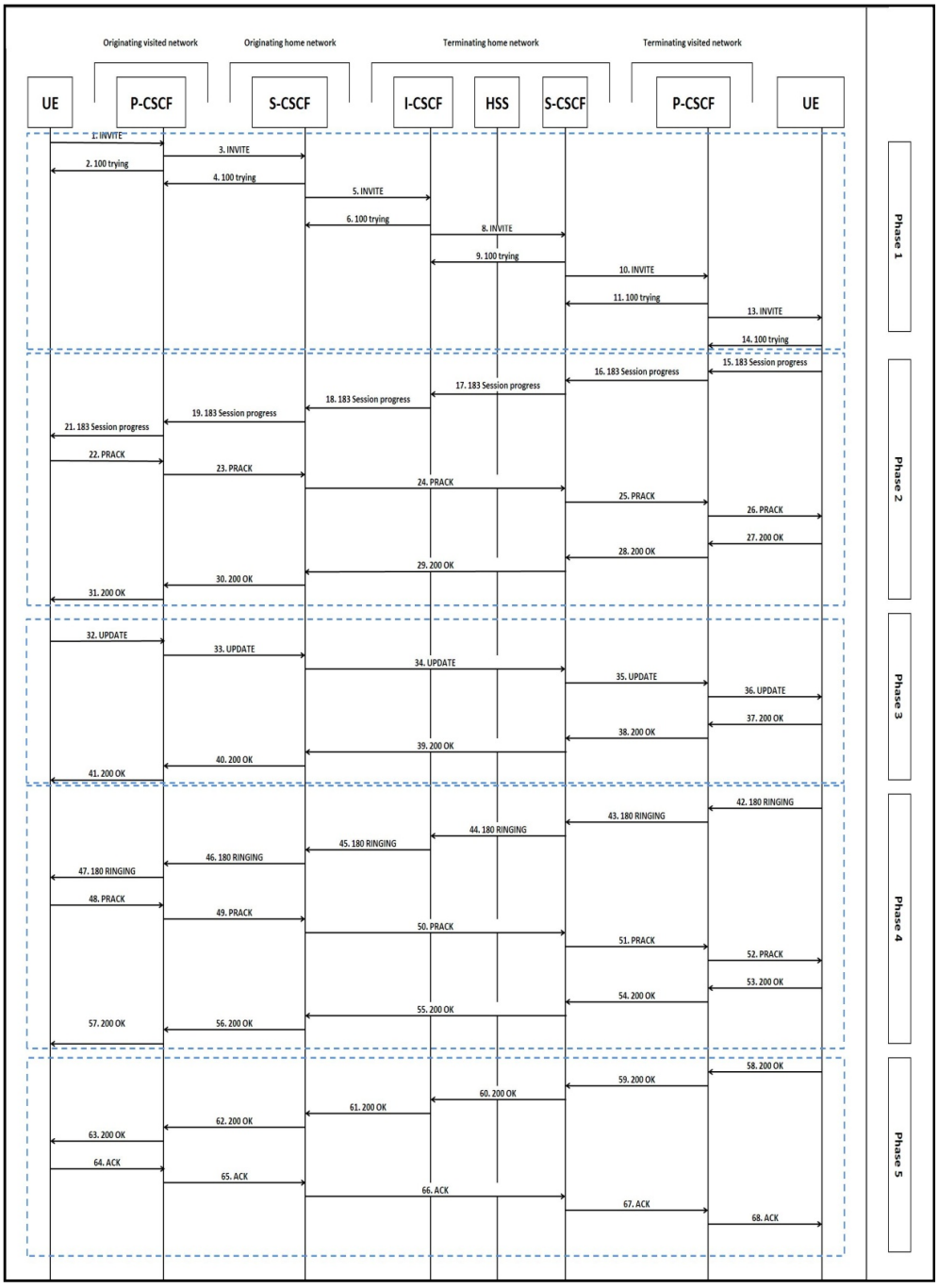


Fig. 2. Basic session establishment scenario (Sisalem et al. 2009)

2.5.1 Modeling the retransmission of the INVITE request

For the case of INVITE requests, the exponential retransmission behavior is used up to a timer called TimerB. That is a request is retransmitted at time points T1, 3T1, 7T1, 15T1 and up to TimerB. This can be represented as a series in the form of:

$$(2^1 - 1)\mathrm{T1}, (2^2 - 1)\mathrm{T1}, (2^3 - 1)\mathrm{T1} \ldots\ldots (2^{\mathrm{N_i}} - 1)\mathrm{T1} \quad (1)$$

With $(2^{\mathrm{N_i}} - 1)\mathrm{T1} = \mathrm{TimerB}$. Thereby the maximum number of retransmitted INVITE requests ($\mathrm{N_i}$) is

$$\mathrm{N_i} = \left\lfloor \log_2\left(\frac{\mathrm{TimerB}}{\mathrm{T1}} + 1\right) \right\rfloor \quad (2)$$

With a loss rate of l, out of r issued INVITE requests per T1 seconds (r x l) packets would be lost on average. These would be retransmitted T1 seconds later. The retransmitted packets would also suffer from a loss and will have to be retransmitted later. Hence, the call generation rate (Ri) can be depicted is shown in Table 1.

Time	R_i	Lost
0	r	lr
1 T1	$r + lr$	$lr + l^2r$
2 T1	$r + lr$	$lr + l^2r$
3 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
4 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
5 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
….	….	….
7 T1	$r + lr + l^2r + l^3r$	$lr + l^2r + l^3r + l^4r$
8 T1	$r + lr + l^2r + l^3r$	$lr + l^2r + l^3r + l^4r$
….	….	….

Table 1. Retransmission Behavior of Invite Requests Due To Network Losses in IMS Network (Sisalem et al. 2008)

At time point 0, r requests are sent per T1 seconds. After T1 seconds the senders will continue generating r new INVITE requests per T1 second and will retransmit the lost (l x r) requests, e.g, (r + (l x r)) will be sent. Out of those (l x (r + (l x r)) will be lost. These would be retransmitted at time 3T1.At time 2T1 r new requests will be sent plus the requests that were lost T1 seconds ago, e.g., (l x r) requests. Out of the sent request (l x (r + (l x r)) will be lost. These would be retransmitted at time 4T1 and so on.The number of INVITE requests (Ri) sent by the sender at any time point (n) can, hence, be determined as:

$$\mathrm{R_i}(\mathrm{l}, \mathrm{n}) = \mathrm{r} \times \left(1 + \sum_{\mathrm{m}=1}^{\mathrm{m}=\mathrm{k}} \mathrm{l}^{\mathrm{m}}\right) = \mathrm{r} \times \sum_{\mathrm{m}=0}^{\mathrm{m}=\mathrm{k}} \mathrm{l}^{\mathrm{m}} \quad (3)$$

with ($k = \lfloor \log_2(n+1) \rfloor$). n can be maximally ($\frac{\text{TimerB1}}{\text{T1}}$).

At this stage the number of new losses, e.g. losses of newly generated requests, would become equal to the number of retransmissions that will be terminated as the maximum number of attempts was already tried. Hence, at this stage the system reaches a steady state.

2.5.2 Losses during INVITE phase

As already described, an INVITE is sent reliably on a hop-by-hop basis. Hence if the INVITE sent by the Caller UE or the 100 response sent by the first proxy in originating visited network (P-CSCF) request were lost then the Caller UE would retransmit the INVITE. The same applies between the each two proxies and between the last proxy in terminating home network (P-CSCF) and the Callee UE. Hence, we can consider the six hops as independent from each other. For each hop a request is considered to be successfully sent if the request and its response arrive at their destinations successfully. Thereby, one needs to consider the losses in both directions, e.g., an end-to-end loss (l) of

$$l_e = 1 - (1 - l)^2 \tag{4}$$

2.5.3 Modeling the retransmission of the Non-INVITE request

The Non-INVITE requests use the exponential retransmission behavior up to a timer called T2 and then every T2 seconds up to the so called TimerF. That is a request is retransmitted at time points T1, 3T1, 7T1, 15T1 and up to T2. Then at 2T2 , 3T2, and up to TimerF. This can be represented as a series in the form of:

$$(2^1 - 1)\text{T1}, (2^2 - 1)\text{T1}, (2^3 - 1)\text{T} \ldots . \left(2^{N_n^e} - 1\right)\text{T1}, 2\left(2^{N_n^e} - 1\right)\text{T1} \ldots \text{TimerF} \tag{5}$$

with

$$\left(2^{N_n^e} - 1\right)\text{T1} = \text{T2}$$

The number of retransmissions (N_n^e) conducted in the exponential manner up to the T2 Timer is determined as:

$$N_n^e = \left\lfloor \log_2\left(\frac{\text{T2}}{\text{T1}} + 1\right) \right\rfloor \tag{6}$$

After T2 seconds, the retransmission timeout is kept constant to T2. The maximum number of retransmissions of a Non-INVITE request (Nn) can then be determined as the sum of N_n^e in the exponential part and N_n^l of the linear part:

$$N_n = N_n^e + N_n^l = \left\lfloor \log_2\left(\frac{\text{T2}}{\text{T1}} + 1\right) \right\rfloor + \left\lfloor \frac{\text{TimerF} - \text{T2}}{\text{T2}} \right\rfloor$$

$$= \left\lfloor \log_2\left(\frac{\text{T2}}{\text{T1}} + 1\right) \right\rfloor + \left\lfloor \frac{\text{TimerF} - \left(2^{N_n^e} - 1\right) \times \text{T1}}{\text{T2}} \right\rfloor \tag{7}$$

with $\left(\left(2^{N_n^e}-1\right)\times T1\right)$ indicating the time point at which the sender goes from exponential backup to a constant timeout value. For Non-INVITE requests, the transmission behavior is slightly different, see Table 2.

With the value of T2 set to 4 seconds and T1 set to 0.5 seconds. In this case, the number of Non-INVITE requests (R_o) sent by the sender at any time point (n) can be determined as:

$$R_o(l,n)=\begin{cases} r\times\left(1+\sum_{m=1}^{m=k} l^m\right) & n\le\frac{T2}{T1} \\ r\times\left(1+\sum_{m=1}^{m=k} l^m+\sum_{m=k+1}^{m=q} l^m\right) & \text{otherwise}\end{cases} \tag{8}$$

with $\left(k=\left\lfloor\frac{\ln(n+1)}{\ln(2)}\right\rfloor\right)$ while $\left(n\le\frac{T2}{T1}\right)$, e.g., k would be maximally equal to N_o^e.

$$q=\left\lfloor\frac{(n-2^{N_n^e}-1)\times T1}{T2}\right\rfloor \tag{9}$$

which ensures that q is incremented every T2 seconds.Note that the maximum value of n here is $\left(\frac{TimerF}{T1}\right)$ at which stage the steady state is reached, e.g., number of new retransmissions equals the number of terminated retransmissions.

Time	R_o	Lost
0	r	lr
1 T1	$r + lr$	$lr + l^2r$
2 T1	$r + lr$	$lr + l^2r$
3 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
4 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
5 T1	$r + lr + l^2r$	$lr + l^2r + l^3r$
….	….	….
7 T1	$r + lr + l^2r + l^3r$	$lr + l^2r + l^3r + l^4r$
8 T1	$r + lr + l^2r + l^3r$	$lr + l^2r + l^3r + l^4r$
….	….	….
15 T1	$r + lr + l^2r + l^3r + l^4r$	$lr + l^2r + l^3r + l^4r + l^5r$
16 T1	$r + lr + l^2r + l^3r + l^4r$	$lr + l^2r + l^3r + l^4r + l^5r$
….	….	….
23 T1	$r + lr + l^2r + \ldots + l^5r$	$lr + l^2r + \ldots + l^6r$
24 T1	$r + lr + l^2r + \ldots + l^5r$	$lr + l^2r + \ldots + l^6r$
….	….	….
31 T1	$r + lr + l^2r + \ldots + l^6r$	$lr + l^2r + \ldots + l^7r$
….	….	….

Table 2. Retransmission Behavior of Non-Invite Requests Due To Losses in IMS Network (Sisalem et al. 2008)

2.5.4 Losses during Non-INVITE phase

After sending a final response, the UA server expects to receive a reply, e.g., an ACK, before T1 seconds. Hence, the relation between the final response and the ACK is similar to that between an INVITE and a provisional response. Unlike the INVITE requests, the relation between the final response and the ACK is an end-to-end one, e.g., the proxies in between would not retransmit the lost messages. Hence for determining the number of final response (P_f) and ACK requests (P_a) needed on the average for setting up a session, one can use the same equations as previously but by taking into account the end-to-end delay. Assuming that all requests follow the same path, e.g., the proxies record-route themselves in the SIP requests and with a loss rate of l on each link, the one way end to end loss (L) of a request traversing the η links would be

$$L = 1 - (1 - l)^{\eta} \tag{10}$$

The end-to-end loss for a request plus response would in this case be

$$L_e = 1 - (1 - L)^2 = 1 - (1 - l)^{2\eta} \tag{11}$$

2.6 Bandwidth consumption of IMS session establishment in a lossy network

Before we start calculating the Bandwidth Consumption of SIP Signaling of IMS call establishment in a Lossy Networks, we have to distinguish between the eight phases of call establishment in terms of INVITE or Non-INVITE retransmission type.

2.6.1 For INVITE/100 trying phase (INVITE)

$$R_i(l_e) = r \times \sum_{m=0}^{m=N_i} {l_e}^m = \; r \times \sum_{m=0}^{m=N_i} (1 - (1 - l)^2)^m \tag{12}$$

(R_i INVITE Retransmission with two ways Losses)

$$R_{100}(l) = r \times \sum_{m=0}^{m=N_i} l^m \tag{13}$$

(R_{100} INVITE Retransmission with one way Losses)

$$N_i = \left\lfloor \log_2\left(\frac{\text{TimerB}}{T1} + 1\right) \right\rfloor$$

2.6.2 For session progress 183/PRACK/200 OK phase

$$R_{183} = \sum_{m=0}^{m=N_{183}} L^m = \sum_{m=0}^{m=N_{183}} (1 - (1 - l)^{\eta})^m \tag{14}$$

(R_{183} INVITE Retransmission with one way End to End Losses)

$$N_{183} = \left\lfloor \log_2\left(\frac{T_{183} + T_{PRACK}}{T1} + 1\right)\right\rfloor \tag{15}$$

$$T_{183} = T1 \times (2^{N_i} - 1) \tag{16}$$

$$N_i = \left\lfloor \log_2\left(\frac{TimerB}{T1} + 1\right)\right\rfloor \tag{17}$$

$$T_{PRACK} = T1 \times \left(2^{N_o^e} - 1\right) + \text{ Max } (0, T2 \times N_n^l) \tag{18}$$

$$T_{PRACK} = T2 + (\text{ TimerF} - T2) = T2 + 64T1 - 8T1 = T2 + 56T1 = 64T1$$

$$N_n = N_n^e + N_n^l = \left\lfloor \log_2\left(\frac{T2}{T1} + 1\right)\right\rfloor + \left\lfloor \frac{TimerF - T2}{T2}\right\rfloor \tag{19}$$

$$R_{PRACK} = \sum_{m=0}^{m=N_n} {L_e}^m = \sum_{m=0}^{m=N_n} (1 - (1 - l)^{2\eta})^m \tag{20}$$

(R_{PRACK} Non-INVITE Retransmission with two ways End to End Losses)

$$R_{200} = \sum_{m=0}^{m=N_n} L^m = \sum_{m=0}^{m=N_n} (1 - (1 - l)^{\eta})^m \tag{21}$$

(R_{200} Non-INVITE Retransmission with one way End to End Losses)

$$N_n = N_n^e + N_n^l = \left\lfloor \log_2\left(\frac{T2}{T1} + 1\right)\right\rfloor + \left\lfloor \frac{TimerF - T2}{T2}\right\rfloor \tag{22}$$

$$L = 1 - (1 - l)^{\eta} \tag{23}$$

$$L_e = 1 - (1 - L)^2 = 1 - (1 - l)^{2\eta} \tag{24}$$

2.6.3 For UPDATE/200 OK phase

$$R_{UPDATE} = \sum_{m=0}^{m=N_o} {L_e}^m = \sum_{m=0}^{m=N_o} (1 - (1 - l)^{2\eta})^m \tag{25}$$

(R_{UPDATE} Non-INVITE Retransmission with two ways End to End Losses)

$$R_{200} = \sum_{m=0}^{m=N_o} L^m = \sum_{m=0}^{m=N_o} (1-(1-l)^{\eta})^m \tag{26}$$

(R_{200} Non-INVITE Retransmission with one way End to End Losses)

$$N_o = N_o^e + N_o^l = \left\lfloor \log_2\left(\frac{T2}{T1}+1\right)\right\rfloor + \left\lfloor \frac{TimerF - T2}{T2}\right\rfloor \tag{27}$$

$$L = 1-(1-l)^{\eta} \tag{28}$$

$$L_e = 1-(1-L)^2 = 1-(1-l)^{2\eta} \tag{29}$$

2.6.4 For ringing 180/PRACK/200 OK phase

$$R_{180} = \sum_{m=0}^{m=N_{180}} L^m = \sum_{m=0}^{m=N_{180}} (1-(1-l)^{\eta})^m \tag{30}$$

(R_{180} INVITE Retransmission with one way End to End Losses)

$$N_{180} = \left\lfloor \log_2\left(\frac{T_{180} + T_{PRACK}}{T1}+1\right)\right\rfloor \tag{31}$$

$$T_{180} = T1 \times (2^{N_i} - 1) \tag{32}$$

$$T_{PRACK} = T1 \times \left(2^{N_n^e} - 1\right) + \text{Max}\ (0, T2 \times N_o^l) \tag{33}$$

$$R_{PRACK} = \sum_{m=0}^{m=N_n} {L_e}^m = \sum_{m=0}^{m=N_n} (1-(1-l)^{2\eta})^m \tag{34}$$

(R_{PRACK} Non-INVITE Retransmission with two ways End to End Losses)

$$R_{200} = \sum_{m=0}^{m=N_n} L^m \sum_{m=0}^{m=N_n} (1-(1-l)^{\eta})^m \tag{35}$$

(R_{200} Non-INVITE Retransmission with one way End to End Losses)

$$N_n = N_n^e + N_n^l = \left\lfloor \log_2\left(\frac{T2}{T1}+1\right)\right\rfloor + \left\lfloor \frac{TimerF - T2}{T2}\right\rfloor \tag{36}$$

$$L = 1-(1-l)^{\eta} \tag{37}$$

$$L_e = 1 - (1 - L)^2 = 1 - (1 - l)^{2\eta} \tag{38}$$

2.6.5 Final response 200 OK/ACK phase (Non-INVITE)

$$R_{200} = \sum_{m=0}^{m=N_n} {L_e}^m = \sum_{m=0}^{m=N_n} (1 - (1 - l)^{2\eta})^m \tag{39}$$

R_{200} Non-INVITE Retransmission with two ways End to End Losses)

$$R_{ACK} = \sum_{m=0}^{m=N_n} L^m = \sum_{m=0}^{m=N_n} (1 - (1 - l)^{\eta})^m \tag{40}$$

(R_{ACK} Non-INVITE Retransmission with one way End to End Losses)

$$N_n = N_n^e + N_n^l = \left\lfloor \log_2\left(\frac{T2}{T1} + 1\right) \right\rfloor + \left\lfloor \frac{TimerF - T2}{T2} \right\rfloor \tag{41}$$

2.7 Total bandwidth usage

The total amount of bandwidth (B) that would be caused by the SIP signaling for IMS call establishment rate of r , η hops and an SIP packet size of S as shown in Table 3. The message sizes for session sequences are provided in Table 3. These values are consistent with (Kueh et al.2003) and (Fathi et al.2006) these references evaluated SIP-based session performance under UDP in different types of networks for instance UMTS etc.).

SIP Message	Size (Bytes)
SIP INVITE	810
SIP REGISTER	225
183 SESSION PROGRESS	260
SIP 180 RINGING	260
SIP PRACK	260
SIP 100 TRYING	260
SIP UPDATE	260
SIP 200 OK	100
SIP SUBSCRIBE	100
SIP ACK	60

Table 3. Message Size for SIP over UDP (Kueh et al.2003)

$$B = \eta \times (R_i \times S_i + R_{100} \times S_{100}) + (R_{183} \times S_{183} + R_{prack} \times S_{prack} + R_{200} \times S_{200}) + (R_{update} \times S_{update} + R_{200} \times S_{200}) + \mu \times (R_{180} \times S_{180} + R_{prack} \times S_{prack} + R_{200} \times S_{200}) + (R_{200} \times S_{200} + R_{ack} \times S_{ack}) \quad (42)$$

l : is the probability of losses between two hops (Assume l is the constant).
L : is the one way End to End Losses.
L_e: is the two ways End to End Losses.
S: is the SIP Message Size.
η : is the number of hops.
μ: is the number of ringing.
r : is the number of Calls or Sessions per Second.
B: The Bandwidth needed for IMS Sessions Establishment.

3. VoIP quality optimization in IMS

The IP Multimedia subsystem (IMS) is an overlay system that is serving the convergence of mobile, wireless and fixed broadband data networks into a common network architecture where all types of data communications are hosted in all IP environments using the session initiation protocol (SIP) protocols infrastructure (23.228 2009). IMS is logically divided into two main communication domains, one for data traffic, i.e., real time protocol packets consisting of audio, video and data and the second one is for SIP signaling traffic. This chapter focuses on the VoIP Quality of Service over IMS using SIP as a signaling protocol. Quality is a subjective factor, which makes it difficult to measure. Taking an end to end perspective of the network further complicates the QoS measurements. The reasons for low quality voice transmission are due to degrading parameters like delay, packet delay variation, codec related impairments like speech compression, echo and most importantly packet loss. Large research efforts have been made to solve the vital quality of service issues. There are some models were developed to measure the VoIP end to end QoS. The output of these models is generally a single quality rating correlated to the subjective Mean Opinion Score (MOS score) which represents the QoS for Voice calls. Many of the developed models for measuring VoIP quality of service are inappropriate for smaller, private networks. They may take too much process resource, are intrusive on the regular traffic or contain very complicated test algorithms. One of the best models used for measuring VoIP quality of service is the E-model, which is a parameter-based model.

The E-Model, (ITU-T Rec. G.107 2005), is a model that allows users to relate Network impairments to voice quality. This model allows impairments to be introduced and voice quality to be assessed. Three cases are considered to demonstrate the effectiveness of optimizing the VoIP over IMS network using E-Model. New equations were also provided to enhance E-Model that can be used to relate packet loss to the level of Equipment Impairment (Ie) with different codecs. The objective function for all cases is to maximize the number of calls that can be active on a link while maintaining a minimum level of voice quality.

The cases considered are:

1. Find voice coder given link bandwidth, packet loss level, and link utilization level.
2. Find voice coder and packet loss level given link bandwidth and background link utilization.

3. Find voice coder and background link utilization level given link bandwidth and packet loss level

OPNET and MATLAB are the optimization tools that are used in this chapter.

3.1 Assumptions for E-Model

The E-Model, (ITU-T Rec. G.107 2005) is extremely complex with 18 inputs that feed interrelated components. These components feed each other and recombine to form an output (R). The recommendation (ITU-T Rec. G.108 1999) gives a thorough description on how to carry out an E-model QoS calculation within VoIP networks.

Due to the complexity of the E-Model, the approach used here is to try to identify which E-Model parameters are fixed and which parameters are not. In the context of this research the only parameters of the E-Model that are not fixed are:

- T and Ta – Delay variables
- Ie – Equipment Impairment Factor
- Id – Delay Impairment Factor

Where (T) is the mean one way delay of the echo path, (Ta) is the absolute delay in echo free conditions. In addition, parameters that affect delay Id and Ie are introduced:

- PL - Packet Loss %
- ρ- Link Utilization
- Coder Type

Next, the relationship between these parameters is identified. Since, we are making the assumption that the echo cancellers on the end are very good, we can say that T = Ta and Ie is directly related to a particular coding scheme and the packet loss ratio.

According to the above assumption, R-Factor equation can be reduced to the following expression (ITU-T Rec. G.107 2005):

$$R = 93.2 - Id\,(Ta) - Ie\,(codec, packet\ loss) \tag{43}$$

3.1.1 Calculation of the delay impairment Id

The factor Id is the delay impairment factor that can be calculated as follow (ITU-T Rec. G.107 2005):

$$Id = 25\left\{\left(1+X^6\right)^{1/6} - \left(1+\left[\frac{X}{3}\right]^6\right)^{1/6} + 2\right\} \tag{44}$$

With:

$$X = \frac{\log\left(\frac{Ta}{100}\right)}{\log 2}$$

3.1.2 Calculation of the equipment impairment Ie

The loss impairment Ie captures the distortion of the original voice signal due to low-rate codec, and packet losses in both the network and the play out buffer. Currently, the E-Model (ITU-T Rec. G.107 2005) can only cope with speech distortion introduced by several codecs i.e. G.729 (ITU-T Rec. G.729 2007) or G.723 (ITU-T Rec. G.723 2006). Specific impairment factor values for codec operation under random packet-loss have formerly been treated using tabulated, packet-loss dependent Ie values. Now, the Packet-loss Robustness Factor Bpl is defined as codec specific value. The packet-loss dependent Effective Equipment Impairment Factor Ie-eff is derived using the codec specific value for the Equipment Impairment Factor at zero packet-loss. Ie and the Packet-loss Robustness Factor Bpl are listed in Table I for several codecs. With the Packet-loss Probability Ppl, (ITU-T Rec. G.113 2007).Ie-eff is calculated using formula (44)

$$\mathrm{Ie-eff} = \mathrm{Ie} + (95 - \mathrm{Ie})\frac{\mathrm{Ppl}}{\mathrm{Ppl} + \mathrm{Bpl}} \tag{45}$$

- Ie is the equipment impairment factor.
- Bpl is called the packet-loss robustness factor, which depends on the used codec.
- Ie,eff represents the packet loss dependent effective equipment Impairment factor, derived from the value of Ie depending on codec and at zero packet loss.
- Ppl is the packet loss probability

As can be seen from this formula (44), the Effective Equipment Impairment Factor in case of Ppl = 0 (no packet-loss) is equal to the Ie value defined in Table 4.Ie represents the effect of degradation introduced by codecs, Packet Loss. (ITU-T Rec. G.113 2007) provides parameters for use in calculating Ie from codec type and Packet Loss rate.

Codec	Rate (Kbps)	Packet Size (msec)	Ie (no packet loss)	Bpl
G 711	64	10	0	25.1
G.729A + VAD	8	20	11	19.0
G.723.1+VAD	6.3	30	15	16.1

Table 4. Provisional planning values for the equipment impairment factor Ie and for packet-loss robustness factor Bpl (ITU-T Rec. G.113 2007)

Coming to the VoIP traffic Characterization, Human speech is traditionally modeled as sequence of alternate talk and silence periods whose durations are exponentially distributed and referred as to ON-OFF model. On the other hand all of the presently available codecs with VAD (Voice Activity Detection) have the ability to improve the speech quality by reproducing Speakers back ground by generating special frame type called SID (Silence Insert Descriptor). SID frames are generated during Voice Inactivity Period.

3.1.3 The R factor

The main output from the E-model is the single R value, produced by an equation combining all relevant impairments. The R factor ranges from 0-100 but is basically

unacceptable below 50. It is recommended for most networks to arrive at a score above 70 when measuring. A user should therefore always reach for an R value as high as possible for all possible connections. Table 5 describes the various quality and satisfaction categories related to the E-model factor R (ITU-T Rec. G.107 2005).

Range of E-Model Rating R	Speech Transmission Quality Category	User Satisfaction
90 < R < 100	Best	Very Satisfied
80 < R < 90	High	Satisfied
70 < R < 80	Medium	Some Users Dissatisfied
60 < R < 70	Low	Many Users Dissatisfied
50 < R < 60	Poor	Nearly All Users Dissatisfied

NOTE - Connection with E-Model Rating R Below 50 are not Recommended

Table 5. E-model related quality and satisfaction categories (ITU-T Rec. G.107 2005).

3.2 Project setup

The simulation is done using OPNET simulation tool IT Guru Academic Edition 9.1 for VoIP in IMS network using SIP Protocol. The network consists of IP-Telephones (VoIP or IMS Clients) connected to the Internet by routers which act as IP gateway, the network is managed by the SIP proxy server (act as P-CSCF) which uses the SIP protocol to establish the voice calls (VoIP) on the IMS network as shown in figure 3. The links between the routers and the Internet are T1 with link speed 1.544 Mbps and the links between the dialer, dialed, Proxy Server and the routers are 1000 Base-x. The idea is to configure the network with a certain parameters and run the simulation then getting from the tool the result values which used in E-Model equations to measure the Quality of service Factor R. The objective function for all cases is to maximize the number of calls that can be active on a link while maintaining a minimum level of voice quality (R). The cases considered are:

1. Find the optimal voice coder given link bandwidth, packet loss level, and background link utilization level.

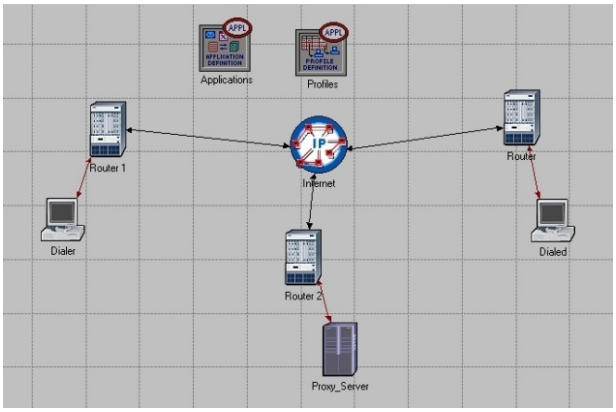


Fig. 3. Network Topology

2. Find the optimal voice coder and the optimal packet loss level given link bandwidth and background link utilization.
3. Find the optimal voice coder and the optimal background link utilization level given link bandwidth and packet loss level.

Table 6 shows standard parameters for each codec used in the analysis

Codec Information	Bandwidth Calculations					
Codec & Bit Rate (Kbps)	Voice Payload Size (Bytes)	Voice Payload Size (ms)	Number of Voice Frames per Packet	Packets Per Second (PPS)	Bandwidth with RTP/UDP/IP/ PPP Header	MTU Size (Bytes)
G.711 (64 Kbps)	160 Bytes	20 ms	1	50	82.8 Kbps	207 Bytes
G.729 (8 Kbps)	20 Bytes	20 ms	1	50	26.8 Kbps	67 Bytes
G.723.1 (6.3 Kbps)	24 Bytes	30 ms	1	34	18.9 Kbps	71 Bytes

Table 6. Codec Parameters

3.3 Results

The results are divided into three general cases. For all cases, the aim is to maximize the number of calls that can be carried on a link while maintaining a minimum voice quality level (R > 70).If two combinations produce the same number of calls, the highest R value will be considered the best selection.

3.3.1 Case 1 – Optimizing for coder selection

The goal of this case is to find the optimal voice coder given link bandwidth, packet loss level, and background link utilization level. Table 4 and Table 5 are containing the coding parameters used in Case1 1 of the simulation. OPNET is configured by these parameters which are according to the ITU-T G.107 (ITU-T Rec. G.107 2005).

Table 7 shows the main differences between the different codecs G.711,G.729 and G.723.1 with respect to the coding type, coder bit rate, frame length , number of voice frames per packet and finally the Ie for each coder in case of no packet loss.

Table 8 also shows other differences between voice codes G.711, G.729 and G.723.1 with respect to the bandwidth calculations like voice payload size, number of packets per second and the bandwidth required after adding the headers of other protocols. For Case 1 with a link speed of 1.544 Mbps, The simulation was run for 2 hours and 4 hours and in all cases G.723.1 gave the max. Number of calls with R value more than 70, so G.723.1 was selected as the optimum Coder. G.711 gave the max. Quality of service (Highest R value) but the lowest number of calls, G.729 gave middle number of calls between G.711 and G.723 and also

Standard	Type	Codec Bit Rate (kbps)	Voice Frame Length (ms)	Look ahead (ms)	Frame length (ms) Packet Length	Number of Voice Frames per Packet	I_E No PL
G.711	PCM	64	0.125	0	20	1	0
G.729	CS-ACELP	8	10	5	20	2	11
G.723.1	MP-MLQ	6.3	30	7	30	1	15

Table 7. Codec Parameters for case1-1 (ITU-T Rec. G.107 2005) & (ITU-T Rec. G.113 2007)

Codec Infor-mation	Bandwidth Calculations					
Codec & Bit Rate (Kbps)	Voice Payload Size (Bytes)	Voice Payload Size (ms)	Number of Voice Frames per Packet	Packets Per Second (PPS)	Bandwidth with RTP/UDP/IP/ PPP Header	MTU Size (Bytes)
G.711 (64 Kbps)	160 Bytes	20 ms	1	50	82.8 Kbps	207 Bytes
G.729 (8 Kbps)	20 Bytes	20 ms	2	50	34.8 Kbps	87 Bytes
G.723.1 (6.3 Kbps)	24 Bytes	30 ms	1	34	18.9 Kbps	71 Bytes

Table 8. Codec Parameters for case1-2 (ITU-T Rec. G.107 2005) & (ITU-T Rec. G.113 2007)

middle R value. As shown in figure 5. Figure 4 shows the average packet end to end delay for different codecs and figure 6 shows the number of connected calls for different coders.

Table 9 contains data collected from OPNET in this case of 4 hours observation and shows that G.723.1 provides the maximum number of calls with accepted voice quality (R=78.2 >70)

PL Ratio	CODEC	Delay (T_A) (ms)	I_D	I_E	R	MOS	Calls
0 %	G.711	115	2.76	0	90.44	4.33	28
	G729a	89	0	11	82.776	4.02	30
	G723.1	97	0	15	78.2	3.86	**39**

Table 9. OPNET Results for case (1)

The results of this case are shown in Figure 6 not surprising, as G.723.1 is a more efficient but lower quality of voice.

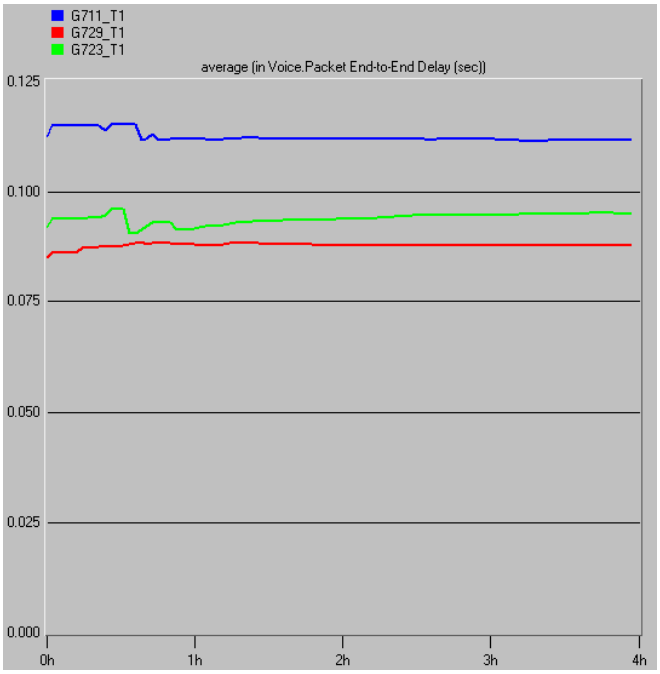


Fig. 4. Average Packet End to End Delay

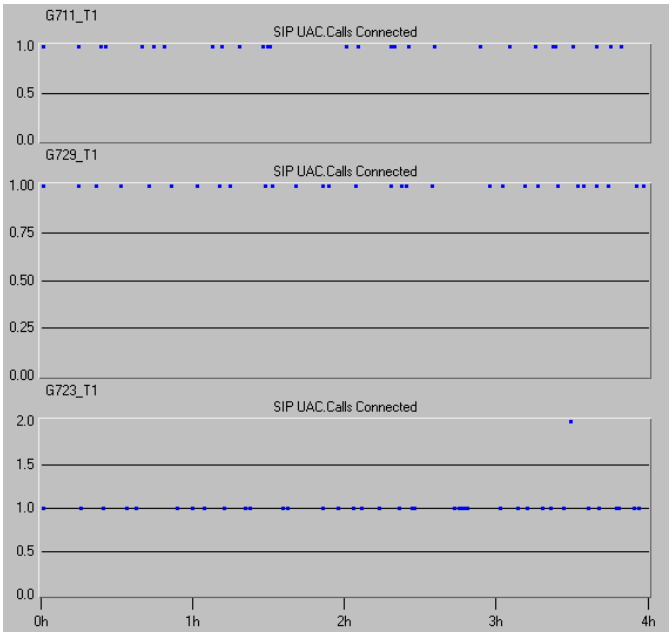


Fig. 5. Number of Connected Calls for different codecs

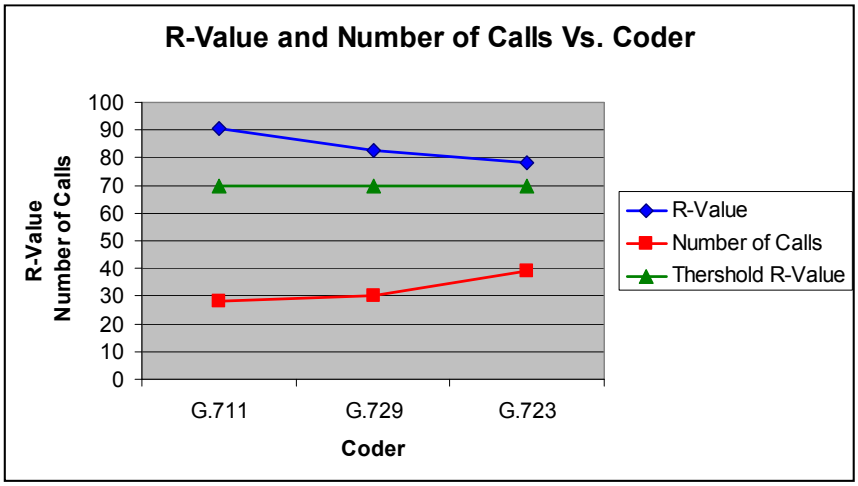


Fig. 6. R Value, Number of Calls vs. Coder – case (1)

3.3.2 Case 2 – Optimizing for Coder and Packet Loss Level Selection

The goal of this case is to find the optimal voice coder and the optimal packet loss level given link bandwidth and background link utilization. Table 10 contains the parameters used for case 2 of the simulation which is according to the ITU-T recommendation G.113

Codec	Rate (Kbps)	Packet Size (ms)	Ie With no PL	Bpl
G.711	64	10	0	25.1
G.729A+VAD	8	20	11	19.0
G.723.1+VAD	6.3	30	15	16.1

Table 10. Codec Parameters for case 2 (ITU-T Rec. G.113 2007)

The OPNET simulation is configured by the above parameters like the codec bit rate and the packet size and the number of voice frames per packet but other values like Ie and Bpl are coming from ITU-T G.107 and G.113 for the mentioned codecs. The simulation was run for 1 hour, 2 hours and 4 hours and for the 3 coders G.711, G.729 and G.723 with different values of packet loss ratio. For the 3 coders G.711, G.729 and G.723 with different values of packet loss ratio (0.5 %, 1 %, 1.5 %, 2 % and 5 %) knowing that the maximum allowable ratio is 2% but the simulation was run for PL% equal 5% to observe the network behavior in case of big crisis as shown in Figure 7.The test was run with a link speed of 1.544 Mbps. The maximum number of calls was 29 calls. G.723.1 with packet loss of 0.5% was the combination chosen and the same combination was chosen till packet loss of 1.5 %. When packet loss ratio reached 2 %, G.723.1 became not feasible as its R value is less than 70 and G.729 with packet loss 2% was the combination chosen. For packet loss more than 2 % G.723.1 and G.729 became not feasible and the only feasible coder is G.711.G.711 with packet loss more than 2 % was the combination chosen.

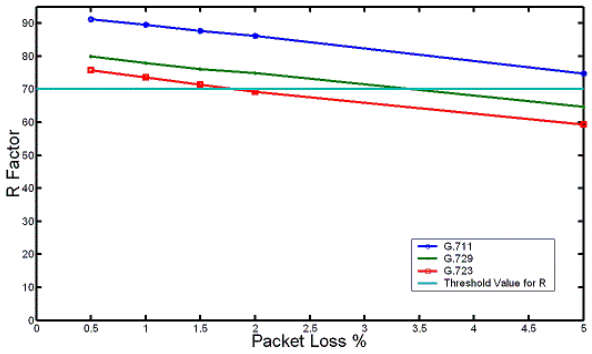


Fig. 7. R Value and PL % vs. Coder – case (2)

The optimization Results for case (2) is listed in Table 11 from the results in table 7we can notice that:

1. With Packet Loss 2% G.723 is not feasible as its R-value < 70.
2. With Packet Loss more than 3.5% G.723 and G.729 are not feasible as their R-value < 70 and you have no choices because G.711 is the only feasible coder with R >70.

Case #	Packet Loss %	Optimum Coder
1	0.5%	G.723.1
2	1%	G.723.1
3	1.5%	G.723.1
4	2%	G.729
5	>3.5%	G.711

Table 11. Results of E-Model Optimization Case (2)

3.3.3 Case 3 – Optimizing for coder and background link utilization

The goal of this case is to find the optimal voice coder and the optimal background link utilization level given link bandwidth and packet loss level. The simulation was run for link speed T1 1.544 Mbps with variable background link utilization (90%, 92%, 94% and 95%) and different coders; the packet loss ratio was considered 0 % in all cases. G.711 with Back ground link utilization 90 % was the combination chosen as the all coders gave the same number of calls and G.711 gave the highest R-value which means the highest quality. With Back ground link utilization 92%, 94% and 95%, G.723.1 became the chosen coder. With back ground link utilization 95% G.711 became not feasible as its R value was below 70 and the feasible coders were G.729 and G.723.1 as shown in figure 9. Looking at Figure 8 we can see that all three coders were in a feasible range until background link utilization reached approximately 94%. In Figure 9 R remains constant for all coders until a point where R declines rapidly. This is important because it suggest that there is optimal link utilization where the system can be operated prior to the R value decline. The sudden

decrease in R is due to the fact that as utilization values approach 100%, the delay becomes unbounded, which negatively affects the R value. It is noticed that at 90% background Link utilization that all coders give the same number of calls, so the selection in this case is based on the R-Value which is the highest for G.711. It is also noticed that the most affected parameter in this case is the Id which is expected as the Background Link Utilization is affecting the Delay (Id) parameters as shown in Figure 10.The optimization Result for case (3) is listed in Table 12

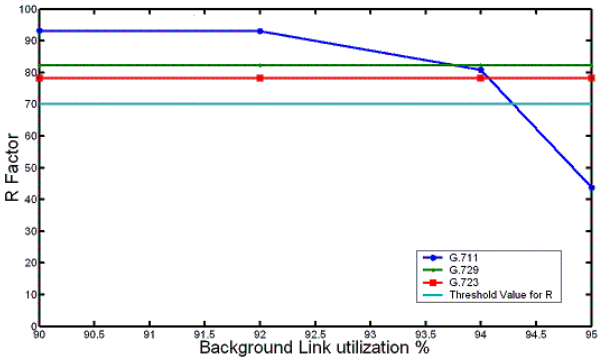


Fig. 8. R Value, Background link Utilization % vs. Coder – case (3)

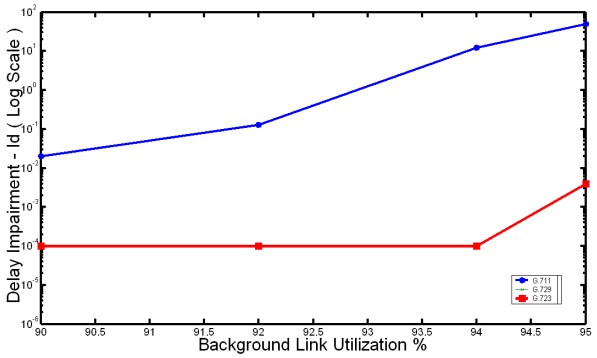


Fig. 9. Background Link Utilization % and Id vs. Coder – case (3)

Case #	Background Link Utilization	Optimum Coder
1	90%	G.711
2	92%	G.723
3	94%	G.723
4	95%	G.723

Table 12. Results of E-Model Optimization Case (3)

Table 12 shows that G.711 with Back ground link utilization 90 % was the combination chosen as the all coders gave the same number of calls and G.711 gave the highest R-value which means the highest quality. With Back ground link utilization 92%, 94% and 95%, G.723.1 became the chosen coder. With back ground link utilization 95% G.711 became not feasible as its R value was below 70 and the feasible coders were G.729 and G.723.1.

3.4 Discussion of E-Model optimization results

This chapter utilized the E-Model to assist with the selection of parameters important to assure the QoS of VoIP in IMS Networks. These parameters include the voice coder, allowable packet loss and the allowable background link utilization. It was based on the concept that maximization of the link usage with respect to the number of calls which is important to the user. It was shown that an optimization of the E-Model is possible and useful. Table 13 reviews the total results of the optimization problem.

Case #	Variables	Optimum Solution
1	Coder	G.723.1
2	Coder, Packet Loss %	G.723.1 with 0.5% PL
3	Coder, Packet Loss %	G.723.1 with 1% PL
4	Coder, Packet Loss %	G.723.1 with 1.5% PL
5	Coder, Packet Loss %	G.729 with 2% PL
6	Coder, Packet Loss %	G.711 with 5% PL
7	Coder, Background Link utilization	G.711With Link Utilization 90%
8	Coder, Background Link utilization	G.723 With Link Utilization 92%
9	Coder, Background Link utilization	G.723 With Link Utilization 94%
10	Coder, Background Link utilization	G.723 With Link Utilization 95%

Table 13. The total results of the optimization problem.

All of the three cases found that G.723.1 is optimal depending on the Circumstances. G.723.1 looks more favorable due to the fact that G.723.1 uses less bandwidth per audio stream. In case 2 , G.723.1 with 0.5%, 1% and 1.5% packet loss was optimal but with packet loss 2 % it was not feasible and G.729 was the optimum coder. In case3 , G.711 coder was selected in case of background link utilization of 90% but in all other cases till 95% G.723.1 was the optimal coder giving the maximum number of calls with R Value more than 70%. The ability to analyze various coders, delay, packets loss and the effect of background link utilization is vital to the QoS of VoIP in IMS network. The optimization of the E-Model provides a tool that is useful for this purpose.

3.5 Proposed enhancement in E-Model

In this section we propose some new equations developed by us using MATLAB could be added to E-Model equations to enhance E-Model performance. It is noticed that the most affected parameter in case (2) when trying to find the best coder in different packet losses condition is the Ie which is expected as the PL is affecting the Ie parameters as shown in Figure10.

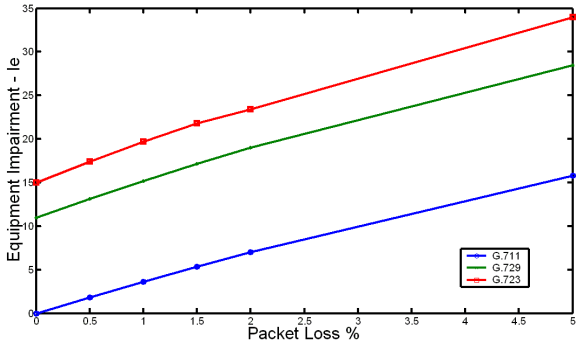


Fig. 10. PL % and Ie vs. Coder – case (2)

Case 2 required additional analysis because not all of the packet loss percentages have been tested and recorded. A polynomial fit was completed for each of the three coders. For G.711, the following polynomial was generated, where x represents the level of packet loss and y represents the level of impairment (Ie).Figure 11, shows a graph of the observed results versus the curve fit.

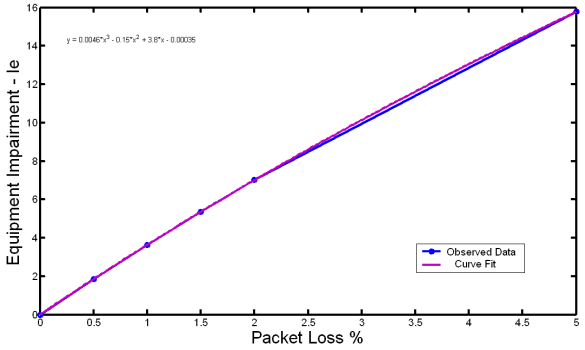


Fig. 11. G.711 Polynomial Fit

The following equation was driven and could be added to enhance E-Model for some codecs.

For G.711, the following polynomial was generated, where x represents the level of packet loss and y represents the level of impairment (Ie).

$$y = 0.0046\,x3 - 0.156x2 + 3.8x - 0.00035 \tag{46}$$

For G.729, the following polynomial was generated, where x represents the level of packet loss and y represents the level of impairment (Ie). Figure 12, shows a graph of the observed results versus the curve fit.

For G.729A, the following polynomial was generated where x represents the level of packet loss and y represents the level of impairment (Ie):

$$y = 0.0081x3 - 0.22x2 + 4.4x + 11 \tag{47}$$

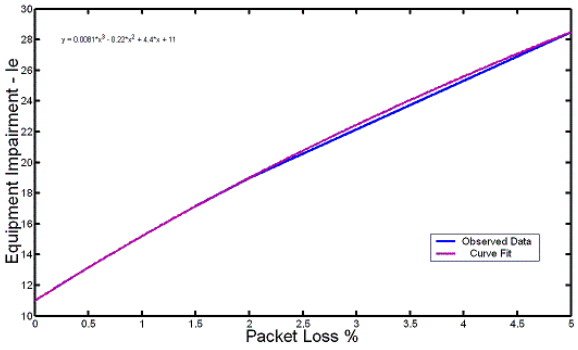


Fig. 12. G.729A Polynomial Fit

For G.723.1, the following polynomial was generated, where x represents the level of packet loss and y represents the level of impairment (Ie). Figure 14, shows a graph of the observed results versus the curve fit.

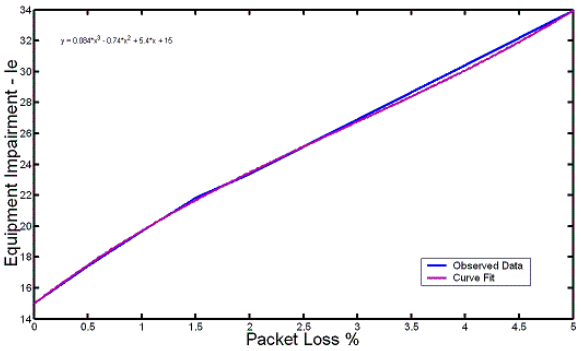


Fig. 13. G.723.1 Polynomial Fit

For G.723.1, the following polynomial was generated where x represents the level of packet loss and y represents the level of impairment (Ie):

$$y = 0.084x3 - 0.74x2 + 5.2348x + 15 \quad (48)$$

3.6 Conclusions

IP Multimedia Subsystem (IMS) is very important due to the critical role it plays in the Next Generation Network (NGN) of the Fixed and Mobile Networks.

In this chapter we provide a theoretical model that can be used by operators and network designers to determine the effects of introducing IMS to their networks in term of bandwidth usage needed to establish IMS session. The inputs of this model are the required number of Calls or Sessions per Second, Network losses, SIP Messages size, Number of Network hops and number of ringing times. The output of this model is the bandwidth needed to insert IMS in the network.

Voice traffic in IMS will be served using Internet protocol (IP) which is called Voice over IP (VoIP). This chapter uses the "E-Model" developed by ITU-T as design tool to select network and voice parameters like coding scheme, packet loss limitations, and link utilization level in IMS Network.

The objective function for all cases is to maximize the number of calls that can be active on a link while maintaining a minimum level of voice quality (R> 70).The cases considered are:

1. Find voice coder given link bandwidth, packet loss level, and link utilization level.
2. Find voice coder and packet loss level given link bandwidth and background link utilization.
3. Find voice coder and background link utilization level given link bandwidth and packet loss level

OPNET and MATLAB are the optimization tool that is used in this chapter.

In case 1, we found that G.723.1 is the optimized coder as it gives the maximum number of calls keeping its R factor more than 70. The quality of speech is generally higher with G.729A and G.711. But G.729A and G.711 uses more bandwidth than G.723.1. In Case 2, both G.729A and G.723.1 were sensitive to changes in packet loss, but G.711 was not as sensitive. In Case 3, voice quality was not sensitive to changes in the link load until the link load grew above approximately 94%.

The chapter also provides new equestrians can be added to enhance E-Model to relate packet loss to the level of Equipment Impairment (Ie) with different codecs.

4. References

23.228 T 2009 IP Multimedia Subsystem (IMS) – Stage 2 (Release 9), Technical specification group core network and terminals, 3rd Generation Partnership Project.

24.229 T 2009 IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (Release 9). Technical specification group core network and terminals, 3rd Generation Partnership Project.

29.208 T 2007 End-to-end Quality of Service (QoS) signalling flows. Technical specification groupcore network and terminals, 3rd Generation Partnership Project.

29.328 T 2008 IP Multimedia Subsystem (IMS) Sh interface; Signalling flows and message contents. Technical specification group core network and terminals, 3rd Generation Partnership Project.

Calhoun, P.; Loughney, j.; Guttman,E.; Zorn,G.& J. Arkko, J.(2003).Diameter Base Protocol. RFC 3588, Internet Engineering Task Force, (September 2003).

Chebbo, H.(2003).Traffic and Load Modelling of an IP Mobile Network. 4th International Conference on 3G Mobile Communication Technologies, London, UK, June 2003.

Fathi, H.; Chakraborty, S. & Prasad, R. (2006). Optimization of SIP Session Setup Delay for VoIP in Wireless Networks, IEEE Transactions on Mobile Computing, vol. 5, no 9, pp. 1121–1132, (September 2006)

Fathi, H.; Chakraborty, S. & Prasad, R. (2006). On SIP session setup delay for VoIP services over correlated fading channels, IEEE Transactions on Vehicular Technology, vol. 55, no 1, pp. 286–295, January 2006.

Faltstrom, P. (2000) E.164 number and DNS, RFC 2916, Internet Engineering Task Force, (September 2000).

Gurbani, V.; Jagadeesan,L. & Mendiratta, V.(2005). Characterizing session initiation protocol (SIP) network performance and reliability. *In ISAS, ser. Lecture Notes in Computer Science,* M. Malek, E. Nett, and N. Suri, Eds. Springer, pp. 196–211.

ITU-T (2005) H.248.1 ITR 2005 Gateway control protocol. Technical report, ITU-T.

ITU-T (2005) Rec. G.107. The E-model, a Computational Model for use in Transmission Planning.Technical report, ITU-T.

ITU-T (1999) Rec. G.108. Application of the E-model: A planning guide. Technical report, ITU-T.

ITU-T (2007) Rec. G.729. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)", Technical report.ITU-T.

ITU-T (2006) Rec.G.723. Dual rate speech coder for multimedia communications transmitting at 5.3 and6.3 kbit/s. Technical report, ITU-T.

ITU-T (2007) Rec. G.113.Transmission impairments due to speech processing. Technical report, ITU-T.

Kueh, V.; Tafazolli, R.& Evans, B.(2003). Performance Evaluation of SIP-based Session Establishment Over Satellite-UMTS. Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual, Vol: 2, 22-25 Apr 2003, pp: 1381 – 1385.

Muhammad T. Alam (2005). An Optimal Method for SIP-Based Session Establishment over IMS. International Symposium on Performance Evaluation of Computer And Telecommunication Systems (SCS 2005), July 24-28, Hilton Cherry Hill/Philadelphia, Philadelphia, Pennsylvania, Sim Series.Vol 37, No. 3, pp: 692-698.

Rosenberg, J.; H. Schulzrinne , H.; Camarillo,G.; Johnston, A.; J. Peterson,J.; R. Sparks, R.;M. Handley,M.& Schooler,E.(2002) SIP: Session Initiation Protocol. Internet Engineering Task Force, RFC 3261,(June 2002).

Schulzrinne, H.; Casner, S.; Frederick, S.; & Jacobson, R.(2003) RTP: a Transport Protocol for Real-Time Applications, STD 64, RFC 3550, Internet Engineering Task Force, (July 2003).

Sisalem, D.; Floroiu, J.; Kuthan, J.; Abend, U. & Schulzrinne, H.(2009) . SIP Security. ISBN 978-0-4-470.51636.2 (cloth)

Sisalem, D.; Liisberg, M. & Rebahi, Y.(2008).A Theoretical Model of the Effects of Losses and Delays on the Performance of SIP. Globecom 2008, New Orleans, (December 2008)

Wu, J.; & Wang, P.(2003).The Performance Analysais of SIP-T Signaling System in Carrier Class VoIP Network, Proceedings of the 17th International Conference on Advanced Information Networking and Applications. Washington, DC, USA: IEEE Computer Society, 2003, p. 39.

5

Dynamic Spectrum Access in Cognitive Radio: An MDP Approach

Juan J. Alcaraz, Mario Torrecillas-Rodríguez, Luis Pastor-González and Javier Vales-Alonso
Universidad Politécnica de Cartagena
Spain

1. Introduction

Cognitive radio refers to a set of technologies aiming to increase the efficiency in the use of the radio frequency (RF) spectrum. Wireless communication systems are offering increasing bandwidth to their users, therefore the spectrum demand is becoming higher. However, RF spectrum is scarce and operators gain access to it by a licensing scheme by which public administrations assign a frequency band to each operator. Currently, this allocation is static and inflexible in the sense that a licensed band can only be accessed by one operator and their clients (licensed users). However, it is a known fact that while some RF bands are heavily used at some locations and at particular times, many other bands remain largely underused FCC (2002). This is, in fact, a classical property of tele-traffic systems, *i.e.* traffic intensity is highly variable during a day. The consequence is a paradoxical situation: while the spectrum scarcity problem hinders the development of new wireless applications, there are large portions of unoccupied spectrum (*spectrum holes* or spectrum opportunities).

Cognitive radio provides the mechanisms allowing unlicensed (or secondary) users to access licensed RF bands by exploiting spectrum opportunities. Cognitive radio is based on software-defined radio, which refers to a wireless communication system that can dynamically adjust transmission parameters such as operating frequency, modulation scheme, protocol and so on. It is crucial that this opportunistic access is performed with the least possible impact on the service provided to licensed users. Therefore, cognitive users should implement algorithms to detect the spectrum use (*spectrum sensing*), identify the spectrum holes (*spectrum analysis*) and decide the best action based on this analysis (*decision making*). Once the decision is made, the cognitive user performs the *spectrum access* according to a medium access control (MAC) protocol facilitating the communication among unlicensed users with minimum collision with other licensed and unlicensed users.

Dynamic spectrum access (DSA) refers to the mechanism that manages the spectrum use in response to system changes (e.g. available channels, unlicensed user requests) according to certain objectives (e.g. maximize spectrum usage) and subject to some constraints (e.g. minimum blocking probability for licensed users). DSA can be implemented in a centralized or distributed fashion. In the former one, a central controller collects all the information required about current spectrum usage and the transmission requirements of secondary users in order to make the spectrum access decision, which is generally derived from the solution of some optimization problem. In distributed DSA unlicensed users make their own decisions autonomously, according to their local information. Compared to centralized DSA, this

scheme requires greater computational resources at the user terminal and generally does not achieve globally optimal solutions. On the other side, distributed schemes imply a smaller communication overhead.

MAC protocols for DSA can also include spectrum trading features. In situations of low spectrum usage, the licensed operator may decide to sell spectrum opportunities to unlicensed users. In order to do this in real-time, a protocol is required to support negotiations on access price, channel holding time, *etc*, between the spectrum owner and secondary users. There are several models for spectrum trading. In this work, we consider the bid-auction model, in which secondary users bid for the spectrum of a single spectrum owner.

This chapter addresses the design of DSA MAC protocols for centralized dynamic spectrum access. We explore the possibilities of a formal design based on a Markov decision process (MDP) formulation. We survey previous works on this issue and propose a design framework to balance the grade-of-service (e.g. blocking probability) of different user categories and the expected economic revenue. When two or more contrary objectives are balanced on an optimization problem, there is not an optimal solution, in the strict sense, but a Pareto front, defined as the set of values, for each individual objective, such that any objective can not be improved without worsening the others. In this work we study the Pareto front solutions for two possible access models. The first one consists of simply providing priority to the licensed users, and the second one is an auction-based model, where unlicensed users offer a bidding price for the spectrum opportunities. In the priority-based access, the centralized policy should balance the blocking probability of each class of users. In the auction-based access, the trade-off appears between the blocking probability of primary users and the expected revenue.

The content structure of the rest of this chapter is the following. Section 2 provides a brief introduction to Markov Decision Processes. Section 3 reviews previous works using the MDP approach in cognitive radio systems. Section 4 explains the system model and MDP formulation for both DSA procedures considered. Section 5 contains the performance analysis of each model based on numerical evaluations of practical examples. Section 6 summarizes the conclusions of this work.

2. Introduction to Markov Decision Processes

Markov Decision Processes (MDPs) are an application of a more general optimization technique known as dynamic programming (DP). The goal of DP is to find the optimal values of a variable when these values (decisions or actions) must be chosen in consecutive stages. The algorithms to solve DP problems rely on the principle of optimality, which states that in an optimal sequence of decisions, every subsequence must also be optimal. DP is generally applied in the framework of dynamical systems. Several basic concepts must be introduced to understand this framework:

- *State*: Is determined by the values of the variables that characterize the system.
- *Stage*: In a discrete-time dynamical system, a stage is a single step in the temporal advance of the process followed by the system. At each stage the system performs a transition from on state to an adjacent one. A process may consist of a finite or infinite number of stages.
- *Action*: At each state, there may be one or several variables whose value can be chosen in order to influence the transition performed at the present stage. The values selected constitute the action at this stage.

- *Cost*: Each pair state-action is associated to a return or outcome, which we will generally refer to as cost. Sometimes the outcome has a positive meaning and is considered a benefit. Additionally, we can compute the total outcome obtained in the whole process. Depending on how it is computed, this overall cost is referred to as total discounted cost or average cost, among others.
- *Policy*: A policy is a function that relates the states with the actions taken at each stage, for the whole duration of the process considered. An optimal policy is the one that attains the best overall cost for a given objective.

As can be anticipated from previous definitions, the goal of DP is to find the optimal policy for a given process. DP is, in fact, a decomposition strategy for complex optimization problems. In this case, the decomposition exploits the discrete-time structure of the policy.

Markov Decision Processes are the application of DP to systems described by controlled discrete-time Markov chains, that is, Markov chains whose transition probabilities are determined by a decision variable.

Let the integer k denote the k-th stage of an MDP. At a given stage, let i and u denote the state of the system and the action taken, respectively. The set of possible values of the state, the *state space*, is denoted by S, therefore $i \in S$. The control space U, is defined similarly. In general, at each state i only a subset of actions $U(i) \subseteq U$ is allowed. We restrict our attention to processes where both S, $U(i)$ and U are independent of k. In this case, the transition probability from state i to state j is denoted as $p_{ij}(u)$. A policy takes the form: $u = \mu(i)$, and because it does not depend on k it is said to be a *stationary* policy. It is said that a policy is admissible if $\mu(i) \in U(i)$ for $i \in S$. At each state i, the policy provides the probability distribution of next state as $p_{ij}(\mu(i))$, for $j \in S$.

The cost of each pair action-state is denoted by $g(i,u)$. Sometimes the costs are associated to transitions instead of states. Let $\tilde{g}(i,u,j)$ denote the transition cost from state i to state j. In this case, we use the *expected cost* per stage defined as:

$$g(i,u) = \textstyle\sum_{j \in S} \tilde{g}(i,u,j) p_{i,j}(u) \tag{1}$$

The objective of the MDP is to find the optimal stationary policy μ such that the total cost is minimized. The total cost may be defined in several ways. We will focus our attention on average cost problems. In this case, the cost to be optimized is given by the following equation

$$\lambda = \lim_{N \to \infty} \frac{1}{N} E \left\{ \sum_{k=1}^{N-1} g\left(x_k, \mu(x_k)\right) \right\} \tag{2}$$

where x_k represents the system's state at the k-th stage. Note that in the definition of the average cost λ we are implicitly assuming that its value is independent of the initial state of the system. This is generally not always true. However there are certain conditions under which this assumption holds. For example, in our scenario, the value of the per-stage cost is always bounded and both S and U are finite sets. Moreover, there is at least one state, n that is *recurrent* in every stationary policy. Given previous conditions, the limit in the right side of (2) exists and the average cost does not depend on the initial state.

Sometimes the system is modeled as a continuous-time Markov chain. In this case, as we shall see, the definition of the average cost is slightly different. In order to solve it by means of the known equations for average cost MDP problems, we have to construct an auxiliary discrete-time problem whose average cost equals the one of the continuous-time problem.

Given the conditions for the limit in 2 to exist, the *optimal* average cost can be obtained by solving the following Bellman's equation

$$h(i) = \min_{u \in U} \left[g(i,u) - \lambda + \sum_{j=1}^{N} p_{ij}(u) h(j) \right] \quad i \in S \tag{3}$$

with the condition $h(n) = 0$. It is known (see Bertsekas (2007)) that previous equations have a unique solution and the stationary policy μ providing the minimum at the right side of (3) is an optimal policy. $h(i)$ is known as relative or differential cost for each state i. It represents the minimum, over all policies, of the difference between the expected cost to reach n from i for the first time and the cost that would be incurred if the cost per stage were equal to the average λ at all states.

There are several computational methods for solving Bellman equation: the value iteration algorithm, the policy iteration algorithm and the linear programming method provide exact solutions to the problem (see Bertsekas (2007) and Puterman (2005)). However, when the dimension of the sets S and U is relatively large, the problem becomes so complex that solving it exactly may be computationally intractable. This is known as the *curse of dimensionality* in dynamic programming. In some situations, we are not able to compute all the transition probabilities $p_{ij}(u)$ of the model, therefore obtaining an exact solution is impossible. For these cases multiple approximate methods have been developed within the framework of approximate dynamic programming (see Powell (2005)) or reinforcement learning.

There are several variations for MDP problems. One of the most important ones refers to the time horizon over which the process is assumed to operate. It may be finite, when the optimization is done over a finite number of stages, or infinite, when the number of stages is assumed to be infinite. The latter type of problems present some theoretical difficulties, and some technical conditions must hold to be solvable. However, when these conditions are present, infinite-horizon problems require less computational effort than finite-horizon problems with similar dimension. Sometimes, more than one performance objective must be attained. In these cases, it is usual to set bounds in all the objectives except one, which should be optimized assuring that the other objectives remain within their bounds, *i.e.* the rest of objectives constitute constraints on the MDP problem. This strategy is known as constrained MDP (CMDP). To solve these problems, the most usual approaches are to re-formulate the problem as a linear-programming one or to use Lagrangian relaxation on the constraints. Finally, in some problems, the control decision at each state must be taken without complete knowledge of the state. Instead of directly observing the state, the controller observes an additional variable related with the state, so that the probability of each state can be inferred. These problems are known as Partially Observable MDP (POMDP) and are tractable, in general, only for small dimensional problems. The more complex versions of MDPs are, in fact, generalizations of the problem. As we will see, some problems must be formulated as Constrained POMDP, for which very few results are available so far and are generally addressed by heuristic methods.

3. MDP applications in cognitive radio

MDP has been frequently applied in the design of MAC protocols in cognitive radio. They can be classified into two classes: decentralized and centralized access protocols. In the decentralized case, each unlicensed user is responsible of performing spectrum sensing and spectrum access, in general with limited, and sometimes unreliable, information about the

spectrum usage. In consequence, it is usual to find partially observed MDP (POMDP) formulations of the problem, which easily become intractable when the dimension of the problem increases. The access of secondary users to the spectrum should have the less possible impact on licensed users. When including these restrictions on the formulation the resulting problem is a constrained POMDP. In the centralized case, a central device, generally referred to as spectrum broker, performs spectrum management, controlling the access of secondary users to idle spectrum channels. It is usually assumed that the spectrum broker has perfect information about the spectrum usage, therefore the problem is formulated as an MDP, or as a CMDP if constraints are included.

3.1 Decentralized access

In Zhao et. al. (2007), the activity of a licensed user is modeled as an on-off model represented by a two-state Markov chain. The problem of channel sensing and access in a spectrum overlay system was formulated as a POMDP. The actions consists on sensing and accessing a channel, and the channel sensing result is considered an observation. The reward is defined as the number of transmitted bits. The objective is to maximize the expected total number of transmitted bits in a certain number of time slots under the constraint that the collision probability with a licensed user should be maintained below a target level.

Geirhofer et. al. (2008) propose a cognitive radio that can coexist with multiple parallel WLAN channels, operating below a given interference constraint. The coexistence between conventional and cognitive radios is based on the prediction of WLAN's behavior by means of a continuous-time Markov chain model. The cognitive MAC is derived from this model by recasting the problem as a constrained Markov decision process (CMDP).

The goal in Chen et. al (2008) is to maximize the throughput of a secondary user while limiting the probability of colliding with primary users. The access mechanism comprises the following three basic components: a spectrum sensor that identifies spectrum opportunities, a sensing strategy that determines which channels to sense and an access strategy that decides whether to access based on potentially erroneous sensing outcomes. This joint design was formulated as a constrained partially observable Markov decision process (POMDP).

The approach in Li et. al. (2011) is to maximize the throughput of the secondary user subject to collision constraints imposed by the primary users. The formulation follows a constrained partially observable Markov decision process.

3.2 Centralized access

In Yu et. al. (2007) the spectrum broker controls the access of secondary users based on a threshold rule computed by means of an MDP formulation with the objective of minimizing the blocking probability of secondary users. In order to cope with the non-stationarity of traffic conditions, the authors propose a finite horizon MDP instead of an infinite horizon one. The drawback is that the policy cannot be computed off-line, imposing a high computational overhead on the system.

Tang et. al. (2009) study several admission control schemes at a centralized spectrum manager. The objective is to meet the traffic demands of secondary users, increasing spectrum utilization efficiency while assuring a grade of service in terms of blocking probability to primary users. Among the schemes analyzed, the best performing one is based on a constrained Markov decision process (CMDP).

Centralized access has received less attention than decentralized access in cognitive radio research in general and in the application of MDP in particular. On the one hand, decentralized access constitutes a harder research challenge because each agent only has partial and sometimes unreliable information about the wireless network and the spectrum bands. This leads to the harder POMDP problems. On the other hand, although centralized access relies on a spectrum broker which generally has full information about the system state, the dimension of the problem increases proportionally to the total number of managed channels. Therefore, although the MDP or CMDP problem may be solvable, its dimension imposes a serious computational overhead. This drawback may be overcome with an off-line computation of the policies. However, when traffic conditions are non-stationary this approach is not applicable and approximate solutions based on reinforcement learning strategies should be explored. In this work we focus on the application of MDP to centralized access and how it can be exploited to balance GoS of each class of user.

3.3 Other applications

Other applications of MDP have been found within the framework of cognitive radio. In Hoang et. al. (2010), authors propose an algorithm based on finite-horizon MDP to schedule the duration of spectrum sensing periods and data transmission periods at the cognitive users aiming to improve their throughput. Berthold et. al. (2008) formulate the spectral resource detection problem as an MDP allowing the cognitive users to select the frequency bands with the most available resources. Galindo-Serrano and Giupponi (2010) deals with the problem of aggregated interference generated by multiple cognitive radios at the receivers of primary (licensed) users. The problem is formulated as a POMDP and it is solved heuristically by means of an approximated dynamic programming method known as distributed Q-learning.

In this paper we highlight another application of MDP: dynamic trading of spectrum bands. While this issue has been typically addressed with a game-theoretic approach, we explore the use of MDP and CMDP formulations to balance benefit and grade of service for primary users in a centralized spectrum access framework.

4. System model

In this section we consider two models for coordinated spectrum access. In the first one, secondary users are accepted or rejected according to an admission policy that only considers the impact on the blocking probability for primary users. In this first model there is a trade-off between the blocking probability of licensed and unlicensed users. The second model includes a spectrum bidding procedure, in which secondary users offer a price, within a finite countable set of prices for mathematical tractability, for the use of a channel. In the second model the trade-off appears between the the blocking probability of licensed users and the expected benefit obtained from spectrum rental.

4.1 Priority-based access

This access is only based on priority, not in bidding price, *i.e.* licensed users are given higher priority than secondary users. Therefore the objective is to minimize the blocking probability of licensed users but also that of unlicensed users. The general rule is that primary users are always accepted if there are available channels but, depending on the available channels, the controller can deny access to secondary users. Once a secondary user occupies a channel, it is this user who decides when to release this channel and it can not be removed by the controller.

There are several approaches to address this type of problems. One of them is to formulate an MDP where the expected cost is obtained as a linear combination (more precisely a convex combination) of the blocking probability of each class of users. By adjusting the weighting factors we can compute a Pareto front for both blocking probabilities. A Pareto front is defined as the set of values corresponding to several coupled objective functions such that, for every point of the set, one objective cannot be improved without worsening the rest of objective values. In this type of access, the Pareto front allows to fix a blocking probability value for the licensed users and know the best possible performance for unlicensed users.

Incoming traffic is characterized by a classic Poisson model. Licensed users arrive with a rate of λ_L arrivals per unit of time. The arrival rate for unlicensed users is denoted by λ_U. The licensed spectrum managed by the central controller is assumed to be divided into channels (or bands) with equal bandwidth. Each user occupies a single channel. The average holding times for licensed and unlicensed users are given by $1/\mu_L$ and $1/\mu_U$ respectively, where μ_L and μ_U denote the departure rate for each class. Because a Poisson traffic model is considered, both the inter-arrival time and the channel holding times are exponentially distributed random variables for both user classes. The model can be easily extended including more user classes, the probability that a user occupies two or more channels, and so on. Essentially the procedure is the same, but the Markov chain would comprise more states as more features are considered in the model. In this model, the state of the Markov chain is determined by the number of channels k occupied by licensed users (LU), and the number of channels s occupied by secondary users (SU). Because spectrum is a limited resource, there is a finite number N of channels. Figure 1 depicts a diagram of the model and its parameters. Note that we can map all the possible combinations of (k, s) for $0 \leq k \leq N$, $0 \leq s \leq N$ and $k + s \leq N$ to a single integer i such that

$$0 \leq i \leq \frac{N(N+1)}{2} + N + 1. \tag{4}$$

The number in the right hand side of 4 is the total number of states. Let N_T denote this number.

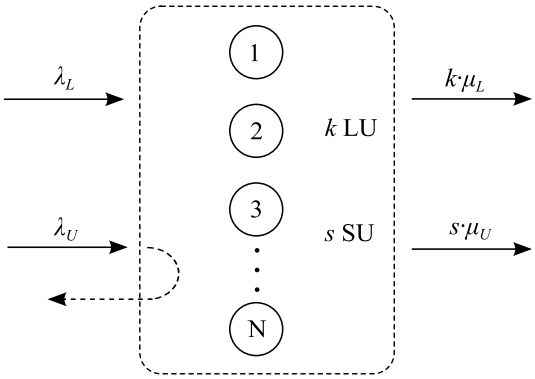


Fig. 1. Diagram of the priority based access model. The system has N channels that can be occupied by k licensed users (LU) and s secondary users (SU) such that $k + s \leq N$. The total departure rates for each type of users depend on k and s.

The model described above consists of a continuous-time Markov chain. In the framework of MDPs we have to define the actions and the costs of these actions. Let $g(i, u)$ denote the

instantaneous cost of taking action u at state i. In the system considered action u is simply defined as

$$u = \begin{cases} 1 & \text{, if incoming user is not accepted} \\ 0 & \text{, otherwise} \end{cases} \tag{5}$$

The above formula refers only to unlicensed users. It is assumed that licensed users are always accepted unless all channels are occupied. The function $g(i,u)$ is given by the convex combination of two per-stage cost functions, *i.e.* $g(i,u) = \alpha g_L(i,u) + (1-\alpha) g_U(i,u)$, where

$$g_L(i,u) = \begin{cases} 1 & \text{, if } i \equiv (k,s) \text{ and } k+s = N \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

where the symbol "$\equiv$" denotes equivalence, i.e. i maps a state (k,s) such that $k+s=N$. Similarly,

$$g_U(i,u) = \begin{cases} 1 & \text{, if } i \equiv (k,s) \text{ and } k+s = N \\ u & \text{, otherwise} \end{cases} \tag{7}$$

These functions determine the blocking probability per unit of time for each class of users. Note that the blocking probability is defined as the probability that the system does not provide a channel to an incoming user. The objective is to find a policy such that, for a relative importance given to each cost (determined by α), the expected average value of the combined cost is minimized. The function to minimize is then given by

$$\lim_{K\to\infty} \frac{1}{E\{t_K\}} E\left\{ \int_0^{t_K} g\left(x(t), u(t)\right) \right\} \tag{8}$$

where t_K is the completion time of the K-th transition. The problem can be solved by formulating its auxiliary discrete-time average cost problem. Let γ be a scalar greater than the transition rate at any state of the chain, *i.e.* $\gamma > v_i(u)$. We can compute the transitions probabilities $\tilde{p}_{i,j}(u)$ for the auxiliary discrete-time problem from the probabilities $p_{i,j}(u)$ of the original problem as

$$\tilde{p}_{i,j}(u) = \begin{cases} \frac{v_i(u)}{\gamma} p_{i,j}(u) & \text{, if } i \neq j \\ 1 - \frac{v_i(u)}{\gamma} & \text{, if } i = j \end{cases} \tag{9}$$

It is known (see Bertsekas (2007)) that if the scalar λ and the vector $\tilde{h}$ satisfy

$$\tilde{h}(i) = \min_{u\in\{0,1\}} \left[g(i,u) - \lambda + \sum_{j=1}^{N_T} \tilde{p}_{ij}(u) \tilde{h}(j) \right] \quad i = 1, \ldots, n \tag{10}$$

then λ and the vector h with components $h(i) = \gamma \tilde{h}(i)$ solve the original problem. It can be anticipated that the structure of this problem, essentially a connection admission control problem, requires a threshold type solution in which upcoming unlicensed users will only be admitted into the system if the number of occupied channels is below certain threshold.

4.2 Auction-based access

As explained in the introduction, public administrations assign the spectrum bands to wireless operators by a license scheme. Generally, operators gain spectrum licenses by bidding for them in public auction processes. We refer to this spectrum assignment framework as primary

market. The increasing demand of spectrum and the existence of spectrum holes have revealed the inefficiency of this mechanism. One practical and economically feasible way to solve this inefficiency is to allow spectrum owners to sell their spectrum opportunities in a secondary market. In contrast to the primary market, the secondary operates in real-time. Secondary users, that may be operators without a spectrum license, submit their bids for spectrum opportunities to the spectrum owner, who determines the winner or winners by giving them access to the band and charging them the bidding price.

The arrival processes are modeled, as in previous subsection, as independent Poisson processes. The arrival rates for licensed and unlicensed users are λ_L and λ_U respectively. The service rates are μ_L and μ_U. Again, it is assumed that each incoming user occupies a single channel. The system state is given by the number of primary users k and secondary users s holding a channel: (k,s) for $0 \leq k \leq N, 0 \leq s \leq N$ and $k+s \leq N$. Each state is mapped into an integer $i \equiv (k,s)$, so that $i = 0,1,\ldots N_T$, where N_T is given by 4. For mathematic tractability, the bidding prices are classified into a finite set of values: $\mathbb{B} = \{b_1, b2, \ldots b_m\}$ given in money charged per unit of time. Each price on this set has a probability p_i, $i = 1 \ldots m$ to be offered by an incoming user. Obviously $\sum_{i=1}^{m} p_i = 1$. Figure 2 depicts the model described.

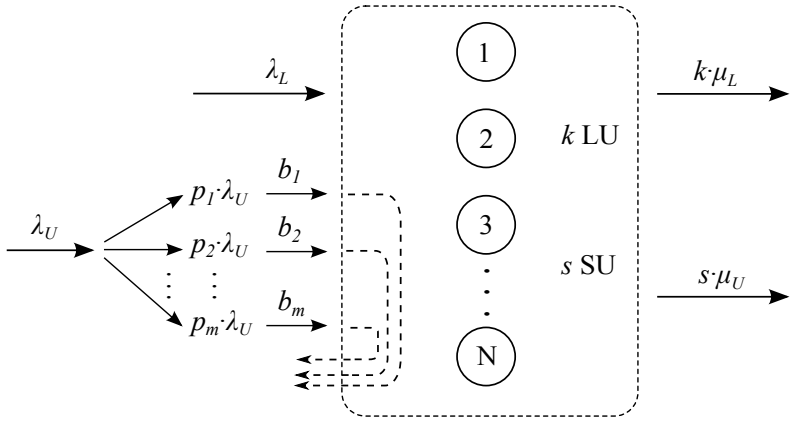


Fig. 2. Diagram of the auction based access model. Secondary users (SU) can offer up to m different bid prices. Each bid offer is assigned a probability. The access policy decides upon each offer according to the price offered and the system's state.

In this case, the objective of the MDP is to obtain the maximum economic profit with the minimum impact on the licensed users. The control u at each stage determines the admitted and rejected bidding prices. Logically, the control should be defined as a threshold, *i.e.* when $u = i$ only bids equal or above p_i are admitted. For notation convenience, the control $u = m+1$ indicates that no bid is accepted. The per-stage reward function $g(i,u)$ is given by the linear combination of $g_L(i,u)$ (defined in previous subsection) and $g_U(i,u)$ defined, in this model, as the expected benefit at stage i when decision u is made. Therefore $g(i,u) = \alpha g_L(i,u) + \beta g_U(i,u)$ where the scalars α and β are weighting factors. Note that $\beta < 0$ since the objective is to minimize the average expected cost given by $g(i,u)$. Let B_i denote the expected income when an unlicensed user whose bidding price is b_i is accepted. Since the average channel holding time for unlicensed users is $1/\mu_U$, then $B_i = b_i/\mu_U$. Given a control u, $P(r|u)$ denotes

the conditional probability that the bidding price of the next accepted secondary user is b_r.

$$P(r|u) = \begin{cases} \frac{p_r}{\sum_{j=u}^{m} p_j} & \text{, if } r \geq u \\ 0 & \text{, otherwise} \end{cases} \tag{11}$$

Let us define $\tilde{g}_U(i,u,j)$ as the average benefit associated to the transition from state i to state j. Its expression is

$$\tilde{g}_U(i,u,j) = \begin{cases} p_U \sum_{r=1}^{m} B_r P(r|u) & \text{, if } j = i+1 \\ 0 & \text{, otherwise} \end{cases} \tag{12}$$

where $p_U = \lambda_U/(\lambda_U + \lambda_L)$ denotes the probability that the next arrival corresponds to a secondary user. Therefore, the per-stage benefit $g_U(i,u)$ is given by

$$\begin{aligned} g_U(i,u) &= \sum_{j=1}^{N_T} \tilde{g}_U(i,u,j) p_{i,j}(u) \\ &= p_{i,i+1}(u) p_U \sum_{j=1}^{N_T} B_r P(r|u). \end{aligned} \tag{13}$$

We can formulate the auxiliary discrete-time average cost problem for the model described. The equation providing the optimum average cost λ is

$$\tilde{h}(i) = \min_{u \in \{0,1\}} \left[\alpha g_L(i,u) + \beta g_U(i,u) v_i(u) - \lambda + \sum_{j=1}^{N_T} \tilde{p}_{ij}(u) \tilde{h}(j) \right] \tag{14}$$

for $i = 1, \ldots, n$. The structure of this problem also anticipates a threshold-type solution. In this case, there will be a set of thresholds, one per bidding price. By properly adjusting the weighting factors α and β we can also compute a Pareto front allowing us to determine the maximum possible benefit for a given blocking objective for the licensed users.

4.3 Constrained MDP

So far, the approach to merge several objectives consisted on combining them into a single objective by means of a weighted sum and solving the problem as a conventional MDP. However, as explained in Section 2, when several objectives concur in an MDP problem, the formulation strategy may consist on optimizing one of them subject to constraints on the other objectives. This strategy results in a CMDP formulation of the problem. Solving MDPs by iterative methods such as policy or value iteration allows us to find deterministic policies, *i.e.* policies that associate each system's state $i \in S$ to a single control $u \in U(i)$, where $U(i)$ is a subset of U containing the controls allowed in state i. However, these policies do not, in general, solve CMDP problems. Instead, the solution of CMDPs is a randomized policy, defined as a function that associates each state to a probability distribution defined over the elements in $U(i)$.

There are mainly two approaches to solve CMDPs, linear programming (LP) and Lagrangian relaxation of the Bellman's equation. This paper follows the former one. Each feasible LP formulation relies on the use of the *dual* variables $\phi(i,u)$, defined as the stationary probability that the system is in state i and chooses action u under a given randomized stationary policy. The problems addressed in this paper result, under every stationary policy, in a truncated birth-death process, since primary users are always accepted. In consequence, every resulting Markov chain is *irreducible*, in other words, it is recurrent and there are not transient states.

Moreover, the state and action spaces are finite. Under these circumstances, as shown in Puterman (2005), every feasible solution of the LP problem corresponds to some randomized stationary policy. Therefore, if the constrained problem is feasible, then there exists an optimal randomized stationary policy.

The LP approach consists of expressing the objective and the constraints in terms of $\phi(i,u)$. Once the problem is discretized, the average cost is defined as

$$\lambda = \lim_{K\to\infty} \frac{1}{K} E\left\{ \sum_{k=0}^{K} g(x_k, u_k) \right\} \tag{15}$$

where k denotes the decision epoch of the process. The objective is to find the policy μ solving

$$\min_{\mu} \lambda \tag{16}$$

The constraints are defined similarly to the main objective: each constraint impose a bound on an average cost related to different per-stage cost. Each constraint has the following form:

$$c = \lim_{K\to\infty} \frac{1}{K} E\left\{ \sum_{k=0}^{K} c(x_k, u_k) \right\} \leq \beta \tag{17}$$

where $c(x(t), u(t))$ is the real-valued function providing the per-stage cost associated to the constraint β. Therefore the constrained average reward MDP with one constraint is defined as

$$\begin{array}{c} \min \lambda \\ \text{s.t.} \\ c \leq \beta \end{array} \tag{18}$$

Given the characteristics of the problem (finite state and action spaces and recurrent Markov chain under every policy), the limits in (15) and (17) exist and are equal to

$$\lambda = \sum_{i\in S} \sum_{u\in U(i)} g(i,u)\, \phi(i,u) \tag{19}$$

and

$$c = \sum_{i\in S} \sum_{u\in U(i)} c(i,u)\, \phi(i,u) \tag{20}$$

respectively. In addition, the following conditions must be hold by the *dual* variables:

$$\sum_{u\in U(j)} \phi(j,u) = \sum_{i\in S} \sum_{u\in U(i)} p_{i,j}(u)\, \phi(i,u) \tag{21}$$

for all $j \in S$, which is closely related to the balance equations of the Markov chain and

$$\sum_{i\in S} \sum_{u\in U(i)} \phi(i,u) = 1, \tag{22}$$

which, together with $\phi(j,u) \geq 1$ for $i \in S$ and $u \in U(i)$ correspond to the definition of $\phi(i,u)$ as a limiting average state action frequency. In consequence, the LP for the CMDP has the

following formulation

$$\begin{gathered}\min_{\phi} \sum_{i\in S}\sum_{u\in U(i)} g(i,u)\,\phi(i,u) \\ \text{s.t.} \\ \sum_{i\in S}\sum_{u\in U(i)} c(i,u)\,\phi(i,u) \le \beta \\ \sum_{u\in U(j)} \phi(j,u) - \sum_{i\in S}\sum_{u\in U(i)} p_{i,j}(u)\,\phi(i,u) = 0 \\ \sum_{i\in S}\sum_{u\in U(i)} \phi(i,u) = 1 \\ \phi(j,u) \ge 1\end{gathered} \tag{23}$$

Assuming that the problem is feasible and ϕ^* is the optimal solution of the LP problem above, the stationary randomized optimal policy μ^* is generated by

$$q_{\mu^*(i)}(u) = \frac{\phi^*(i,u)}{\sum_{u'\in U(i)} \phi^*(i,u')} \tag{24}$$

for cases where the sum in the denominator is nonzero. Otherwise, the state is transitory and the control is irrelevant. Note that $q_{\mu^*(i)}(u)$ denotes the probability of choosing action u at state i under policy μ^*.

Using the approach above in the problems described in previous section is straightforward:

- *Priority-based access*: in the LP problem (23) replace $g(i,u)$ by $g_U(i,u)$ defined in (7), and $c(i,u)$ by $g_L(i,u)$ defined in (6). For each value of β we obtain the point in the Pareto front corresponding to a blocking probability β for the licensed users.
- *Auction-based access*: in the LP problem (23) replace $g(i,u)$ by $g_U(i,u)$ defined in (13), and $c(i,u)$ by $g_L(i,u)$ defined in (6). As in previous case, for each value of β we obtain a point in the Pareto front.

5. Numerical results

In this section we provide examples of the Pareto front computation procedures described in previous section for each DSA type.

5.1 Priority based access

For this DSA scheme we will consider three scenarios characterized by the asymmetry between the traffic intensity of licensed and unlicensed users. In every scenario, the average holding time is equal for every user, independently of their type. Therefore the service rate $\mu_L = \mu_U = 5$. Assuming that the time unit is an hour, this results in an average holding time of 12 minutes per connection. The total traffic ($\lambda = \lambda_L + \lambda_U$) is 40 calls/h, which results in a total incoming traffic of 8 Erlangs. In a wireless cell covering 2.5 km^2 of urban area (cell radius equal to 400 m), with 2000 people per km^2 and a 10% aggregate market penetration (licensed and unlicensed users), the number of covered users is around 500, and the resulting traffic intensity is 0.016 Erlangs per user. The number of available channels is set to $N = 10$, in order to evaluate the system in a relatively congested situation. With the assumed traffic intensity we can estimate the blocking probability of the system for the aggregate traffic by means of

the well-known Erlang's B formula (see Kleinrock (1975)):

$$E(n,\rho) = \frac{\frac{\rho^n}{n!}}{\sum_{j=0}^{j=n} \frac{\rho^j}{j!}} \tag{25}$$

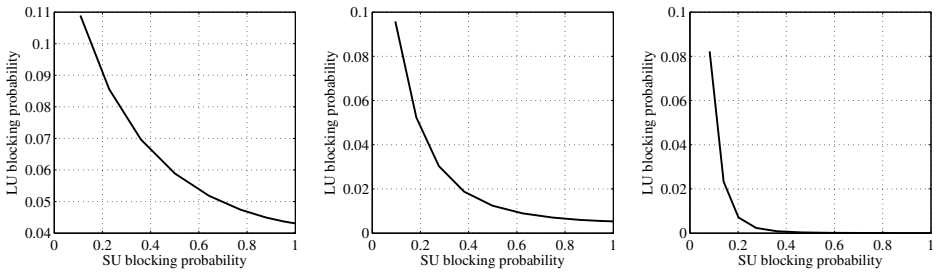
where n is the number of channels and ρ denotes the utilization factor. In our case $\rho=\lambda/\mu_L = \lambda/\mu_U$. According to this formula, if the system accepted every incoming user, the total blocking probability would be $E(10,8)$=0.12. As we will see, this probability is an upper bound for the blocking probability of the primary users, which are always accepted if the system has any available channel, and a lower bound for the secondary users.

The three scenarios are summarized in Table 1.

parameter	**scenario 1**	**scenario 2**	**scenario 3**
λ_L (calls/h)	30	20	10
λ_U (calls/h)	10	20	30
μ_L=μ_U (calls/h)	5	5	5
N	10	10	10

Table 1. Parameters values at the three scenarios of the priority based access problem.

First, we show in Fig. 3 the Pareto front obtained by means of an MDP where the blocking costs of licensed and unlicensed users were merged by means of a convex combination. The Pareto front was obtained by solving each MDP problem for 10000 values of the α parameter ranging from 0.01 to 1.



(a) scenario 1 (b) scenario 2 (c) scenario 3

Fig. 3. Pareto fronts obtained for the priority-based access in scenario 1 (a), scenario 2 (b) and scenario 3 (c)

All the three scenarios receive the same total traffic intensity. However, when the traffic intensity of the primary users is smaller, the Pareto front is closer to both axes, *i.e.* the performances of both the primary and secondary users improve. This is an expectable result since only the traffic of secondary users is controlled by the access policy. When the optimization affects to a higher portion of the total amount of traffic the improvement is also more noticeable, showing the benefits of the MDP formulation.

The Pareto fronts obtained by means of the CMDP formulation in previous scenarios are identical to those shown in Fig. 3, showing that both formulations are equivalent in terms of finding the Pareto front for the priority-based access problem. The only difference relies on practical considerations. The CMDP approach allows us to find a policy with a predefined

blocking probability for primary users while the MDP formulation implies the exploration of the Pareto front, since there is no a priori relationship between α and this blocking probability. On the other hand, implementing the policy solving the CMDP problem implies to randomize at least one control (it can be shown that the number of required randomized controls equals the number of constraints). While this is technically feasible, a stationary deterministic policy is simpler to implement.

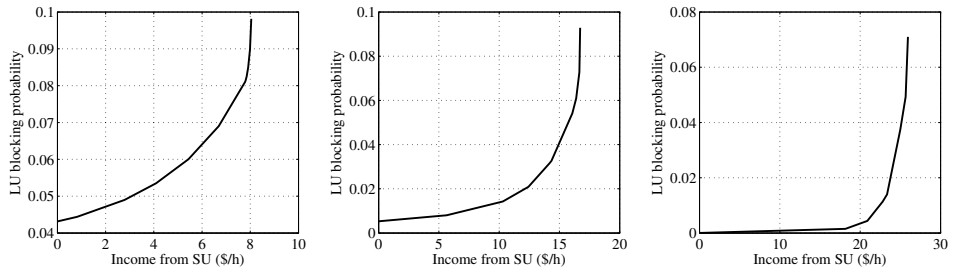
5.2 Auction based access

For the auction-based access we consider again the three scenarios defined in previous section. Additionally we define three classes of secondary users (SU), characterized by the price that they offer per minute of channel occupation. The bid offers per class are: class 1: 0.01 \$/m, class 2: 0.02 \$/m and class 3: 0.03 \$/m. Additionally, we define the probability of an SU incoming call being of each class. The SU class probability distribution is: class 1 probability: 0.5, class 2 probability: 0.3 and class 3 probability: 0.2. We summarize SU class definition in Table 2.

SU class	**class 1**	**class 2**	**class 3**
offered price (\$/m)	0.01	0.02	0.03
probability	0.5	0.3	0.2

Table 2. Classification of SU in terms their bid offers and their probabilities.

Note that both the offered prices and their probability distributions are static, *i.e.* they do not change over time and are independent of the system occupation. It is not completely unrealistic taking into account typical tariff policies of wireless operators. In this environment the class structure and the probability distribution may be seen as types of contracts for secondary users and market penetration of each type of contract respectively. However, for a more dynamical auction process, where bidders are able to change their bid offers adaptively, the model should be revised. One possibility would be to define one probability distribution for each state. More detailed modeling strategies would increase the complexity of the MDP solving algorithm or even make them intractable. This is a classic problem of MDP formulation, known as the *curse of dimensionality* and is typically addressed by means of the heuristic approach of approximate dynamic programming.



(a) scenario 1 (b) scenario 2 (c) scenario 3

Fig. 4. Pareto fronts obtained for the auction-based access in scenario 1 (a), scenario 2 (b) and scenario 3 (c)

Figure 4 shows the Pareto fronts for the auction-based system in the three scenarios. As in previous subsection, the MDP and the CMDP approaches provided similar results. It

can be observed that, for the same traffic intensity (the three scenarios receive 40 calls per unit of time) when the traffic share of the secondary users is higher (scenarios with higher number) the Pareto front moves away from the y-axis, *i.e.* the income obtained from secondary users increases and it also approaches the x-axis, *i.e.* the blocking probability of the licensed users diminishes. It is interesting to check that, especially in scenarios 2 and 3, a very small increment of the blocking probability of licensed users can multiply the benefit obtained from spectrum leasing by a factor of 2 or 3. On the other hand, these figures also indicate that once the income surpasses certain threshold, Pareto-optimal policies can only produce small increments of the income by dramatically rising the blocking probability.

6. Conclusions

This chapter has surveyed the use of MDP formulation within the framework of cognitive radio. We have reviewed the fundamentals of MDP and its generalizations, such as CMDP, POMDP and constrained POMDP. While most previous works focus on decentralized access, we focus on centralized access. The main difference between them is that when the access relies on a central controller or *spectrum broker*, it generally has full knowledge of the spectrum occupation, while in decentralized access decision have to be taken with partial and sometimes unreliable information about channel occupation. Therefore, centralized schemes are more suitable to MDP or CMDP modeling, while decentralized ones generally require POMDP or constrained POMDP which are intractable in many cases and require approximated or heuristic algorithms. We consider two types of access: one where only one type of secondary user tries to access the licensed spectrum and other where users are classified according to the price they are willing to pay for the use of the spectrum. The first one is referred to as priority-based access and the second one as auction-based access. The main issue of the problems addressed is that two contrary objectives coexist. In priority-based access, the controller tries to reduce the blocking probability of both types of users. In auction-based, the objectives are to reduce blocking probability for licensed users and to increase the income received from spectrum leasing. For these problems there does not exists an *optimal* policy, but a set of *Pareto optimal* policies. The performance of these policies lie on the Pareto front, defined as the set of points where one objective cannot be improved without worsening the other one. We have shown how to compute these Pareto fronts for each access scheme by weighting the objectives in an MDP problem and by formulating a CMDP. The first approach requires solving Bellman's equation and the second requires solving a linear program. We have obtained the Pareto fronts for several scenarios, showing the influence of traffic share on system's performance. The Pareto front is a very usual tool to determine the performance threshold for each objective upon which further increments on this objective require excessive degradation of the other one. MDP and CMDP are useful tools for developing centralized access policies for cognitive radio systems. One drawback is the so-called *curse of dimensionality*, that may render computationally intractable the problem as the sizes of the state and action spaces increase. In addition, although policies can be computed off-line, alleviating the computational overhead of the access controller, the system's parameters may be variable, requiring many pre-computed policies and thus imposing large memory requirements.

7. Acknowledgments

This research has been supported by the MICINN/FEDER project grant TEC2010-21405-C02-02/TCM (CALM) and it was also developed in the framework of

"Programa de Ayudas a Grupos de Excelencia de la Region de Murcia, Fundacion Seneca, Agencia de Ciencia y Tecnologia de la RM (Plan Regional de Ciencia y Tecnologia 2007/2010".

8. References

FCC Spectrum policy Task Force "Report on the spectrum efficiency group,", FCC Report, Nov. 2002.

D. P. Bertsekas, "Dynamic Programming and Optimal Control, vol. 2," Third Edition *Athenea Scientific*, 2007.

M. L. Puterman "Markov Decision Processes: Discrete Stochastic Dynamic Programming," First Edition *Wiley-Interscience*, 2005.

W. B. Powell "Approximate Dynamic Programming: Solving the Curses of Dimensionality," First Edition *Wiley-Interscience*, 2007.

Q. Zhao, L. Tong, A. Swami and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, April 2007.

S. Geirhofer, L. Tong and B.M. Sadler, "Cognitive Medium Access: Constraining Interference Based on Experimental Models," *Selected Areas in Communications, IEEE Journal on* , vol.26, no.1, pp.95-105, Jan. 2008.

Y. Chen, Q. Zhao and A. Swami, "Joint Design and Separation Principle for Opportunistic Spectrum Access in the Presence of Sensing Errors," *Information Theory, IEEE Transactions on* , vol.54, no.5, pp.2053-2071, May 2008.

X. Li, Q. Zhao, X. Guan and L. Tong, "Optimal Cognitive Access of Markovian Channels under Tight Collision Constraints," *Selected Areas in Communications, IEEE Journal on*, vol.29, no.4, pp.746-756, April 2011.

O. Yu, E. Saric and A. Li, "Dynamic control of open spectrum management," in *Proceedings of IEEE Wireless Communications and Networking Conference* (WCNC), March 2007, pp. 127-132.

P. K. Tang, Y. H. Chew, W.-L. Yeow and L. C. Ong, "Performance Comparison of Three Spectrum Admission Control Policies in Coordinated Dynamic Spectrum Sharing Systems," *Vehicular Technology, IEEE Transactions on* , vol.58, no.7, pp.3674-3683, Sept. 2009.

A. Hoang, Y.-C. Liang and Y. Zeng., "Adaptive joint scheduling of spectrum sensing and data transmission in cognitive radio networks," *Communications, IEEE Transactions on* , vol.58, no.1, pp.235-246, Jan. 2010.

U. Berthold,F. Fangwen, M. van der Schaar and F.K. Jondral, "Detection of Spectral Resources in Cognitive Radios Using Reinforcement Learning," in Proc. *New Frontiers in Dynamic Spectrum Access Networks, 2008. DySPAN 2008. 3rd IEEE Symposium on* , vol., no., pp.1-5, 14-17 Oct. 2008.

A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *Vehicular Technology, IEEE Transactions on* , vol.59, no.4, pp.1823-1834, May 2010.

L. Kleinrock, "Queuing Systems, Volume 1: Theory," *John Wiley & Sons*, New York, 1975.

6

Call Admission Control in Cellular Networks

Manfred Schneps-Schneppe[1] and Villy Bæk Iversen[2]
[1]*Ventspils University College*
[2]*Technical University of Denmark*
[1]*Latvia*
[2]*Denmark*

1. Introduction

The service area of a cellular network is divided into cells. Users are connected to base stations in the cells via radio links. Channel frequencies are reused in cells that are sufficiently separated in distance so that mutual interference is below tolerable levels. When a new call is originated in a cell, one of the channels assigned to the base station of the cell is used for communication between the mobile user and the base station (if any channel is available for the call). If all the channels assigned to this base station are in use, the call attempt is assumed to be blocked and cleared from the system (blocked calls cleared). When a new call gets a channel, it keeps the channel until either the call is completed inside the cell or the mobile station (user) moves out of the cell. When the call is completed, the channel is released and becomes available to serve another call.

When a mobile station moves across the cell boundary and enters a new cell, a handover is required. Handover is also named handoff. If an idle channel is available in the destination cell, a channel is assigned to it and the call stays on; otherwise the call is dropped. Two commonly used performance measures for cellular networks are dropping probability of handover calls and blocking probability of new calls. The *dropping probability* of handover calls represents the probability that a handover call is dropped during handover. The *blocking probability* of new calls represents the probability that a new call is denied access to the network.

Call admission control (CAC) algorithms are used in order to keep control on dropping probability of handover calls and blocking probability of new calls. They determine whether a call should be accepted or rejected at the base station. Both the blocking probability of new calls and the dropping probability of handover calls are affected by the call admission algorithm used. The call admission algorithms must give priority to handover calls as compared to new calls.

Dedicated to the memory of Janis Sedols, Doctor of Mathematics (Dr.sc.comp.), 24.03.1939–11.08.2011. Dr. Sedols was active in writing this chapter.

Paper financed from EDRF's project SATTEH (No. 2010/0189/2DP/2.1.1.2.0./10/APIA/VIAA/019) being implemented in Engineering Research Institute "Ventspils International Radio Astronomy Centre" of Ventspils University College (VIRAC).

Various priority based call admission algorithms have been reported in the literature, as for example (Beigy & Meybodi, 2003);(Ahmed, 2005);(Ghaderi & Boutaba, 2006). They can be classified into two basic categories: (1) reservation based, and (2) call thinning schemes.

1. Reservation based schemes:
   In these schemes, a subset of channels is reserved for exclusive use by handover calls. Whenever the number of calls (new calls) exceeds a certain threshold, these schemes reject new calls until the number of simultaneous calls (new calls) decreases below the threshold. These schemes accept handover calls as long as the cell has idle channels. When the number of calls is compared with the given threshold, this scheme is called call bounding (Ramjee et al., 1997); (Hong & Rappaport, 1986); (Oh & Tcha, 1992); (Haring et al., 2001); (Beigy & Meybodi, 2005). When the current number of new calls is compared with the given threshold, the scheme is called new call bounding scheme (Fang & Zhang, 2002). Equal Access sharing with reservation schemes reserve an integral number of channels or a fractional number (Ramjee et al., 1997) of channels for exclusive use by handover calls. Schemes with fractional number of guard channels have better control of the blocking probability of the new calls and the dropping probability of the handover calls than schemes with integral number of guard channels.
2. Call thinning schemes:
   These schemes accept new calls with a certain probability that depends on the number of ongoing calls in the cell (Ramjee et al., 1997); (Beigy & Meybodi, 2004). New call thinning schemes accept new calls with a probability that depends on the number of ongoing new calls in the cell (Fang & Zhang, 2002); (Cruz-Pérez et al., 2011). Both schemes accept handover calls whenever the cell has free channels.

In Sections 2 and 3, we compare four basic CAC strategies by examining new call and handover call blocking probabilities for the following schemes:

1. Dynamic reservation,
2. Fractional dynamic reservation,
3. Static (fixed) reservation,
4. New call bounding scheme.

Section 4 deals with Dynamic reservation and Static reservation in two-tier networks. To a great extent, our purpose is a tutorial one because there are many papers on CAC schemes, but they usually contain incomparable numerical results developed by computer simulations. Similar research as our is done by (Ramjee et al., 1997). They show that the guard channel scheme is optimal for minimizing a linear objective function of call blocking and dropping probabilities. The scheme studied below appeals also to network providers in terms of maximizing the revenue obtained by simple mathematical means.

1.1 Dimensioning of multi-tier networks

We have many types of multi-tier cellular networks all around us.

1.1.1 Mobile networks

A multi-tier cellular network is a network that has different types of cells overlaying each other. Each type of cell differs from others by the size. The smaller the size of the cells in a

certain area is, the more channels are available for users (since the number of channels per cell is fixed). We may consider four types of cells:

1. Pico-cells (range 10 – 50 m) are used inside buildings and lifts. The cell antennae are placed in corners of a room or in hallways. Pico-cells are used when the number of users in a building is large and signals from the outside cells cannot penetrate the building. A new type having almost the same features as pico-cells are called femto-cells.
2. Micro-cells (range 50 m – 1 km) are cells used mainly in cities where there are a lot of users.
3. Macro-cells (range 1 km – 20 km) are used in rural areas since the number of users is small, and in populated areas where micro-cells are too small to handle frequent handovers of users that are moving fast while making calls. For example, if you are in a high speed car and connected to a micro-cell, and the car moves too fast for the call to be handed over from one base station (cell antenna) to another, then the call will be dropped.
4. Satellites (world wide coverage).

Having multi-tier cellular networks increases the number of cells, which means that more users are able to use the network without being blocked, and that users in cars or any high speed vehicles are able to talk without worrying about their calls being disconnected.

1.1.2 Hybrid networks

The proliferation of computer laptops, personal digital assistants (PDA), and mobile phones, coupled with the nearly universal availability of wireless communication services is enabling the goal of ubiquitous wireless communications (Beigy & Meybodi, 2003); (Ramjee et al., 1997); (Leong & Zhuang, 2002); (Guerin, 1988). Unfortunately, to realize the benefits of omni-present connectivity, users must contend with the challenges of a confusing array of incompatible services, devices, and wireless technologies. Rice University, USA, is developing RENÉ (Rice Everywhere NEtwork) (Aazhang & Cavallaro, 2001), a system that enables ubiquitous and seamless communication services. The key innovations are a first-of-its-kind multi-tier network interface card, intelligent proxies that enable a new level of graceful adaptation in unmodified applications, and a novel approach to hierarchical and coarse-grained quality of service provisioning. The design of RENÉ requires a coordinated, collaborative effort across traditional layers and across different time scales of the system to maintain uninterrupted user connectivity.

1.1.3 Battlefield networks

Future battlefield networks will consist of various heterogeneous networking systems and tiers with disparate capabilities and characteristics, ranging from ground ad hoc mobile, sensor networks, and airborne-rich sky networks to satellite networks. It is an enormous challenge to create a suite of novel networking technologies that efficiently integrate these disparate systems (Ryu et al., 2003).

The key result of this chapter is the application part (Section 5) with the extension of the Equivalent Random Traffic method for estimation of throughput for networks with traffic splitting and correlated streams. The excellent accuracy (relative error less than 1%) is shown by numerical examples. The ERT-method has been developed for planning of alternate routing in telephone systems by many authors: (Wilkinson, 1956); (Bretschneider, 1973); (Fredericks, 1980) and others. In this chapter we propose an extension of the ERT-method

to take account of correlated streams. Sections 5.7 and 5.8 contain next step in ERT-method extension, namely the application of Neal's theory (Neal, 1971), and his formulas for covariance of correlated streams.

2. Analysis of CAC strategies in single tier networks: Two channel case

We compare four basic CAC schemes by examining new call and handover call blocking probabilities:

- Strategy 1 - Dynamic reservation: the cutoff priority scheme is to reserve some channels for handover calls. Whenever a channel is released, it is returned to the common pool of channels.
- Strategy 2 - Fractional dynamic reservation: the fractional guard channel scheme (the new call thinning scheme) is to admit a new call with a certain probability which depends on the number of busy channels.
- Strategy 3 - Static (fixed) reservation: all channels allocated to a cell are divided into two groups: one to be used by all calls and the other for handover calls only (the rigid division-based CAC scheme).
- Strategy 4 - New call bounding scheme: limitation of the number of simultaneous new calls admitted to the network.

We consider a N-channel cell without waiting positions and two Poisson call flows: handover call flow of intensity A and new call flow of intensity B. The holding times are exponentially distributed with mean value equal one. The exponential distribution simplifies the formulae and the fact that handover calls already has been served for some time before entering the cell considered, does not influence the remaining service time, as the exponential distribution is without memory. Our optimization criteria is the same for all schemes: to get the maximum revenue if each served A-call costs K units ($K > 1$) and each served B-call costs one unit.

2.1 Dynamic reservation strategy is better than fractional dynamic reservation strategy

Let us start from the 2*nd* strategy: fractional dynamic reservation. This strategy seems to be more general than the dynamic reservation strategy, but we shall prove that such statement is not true. The system is modeled by a three-state Markov process (Fig. 1) having the following parameters: Number of channels $N = 2$, A and B = call flow intensities, p_i = the probability of accepting B-calls for service in state i.

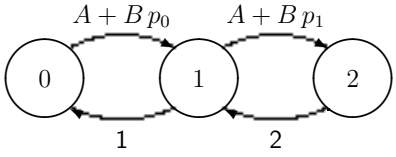


Fig. 1. Fractional dynamic reservation : two-channel state transition diagram.

The stationary state probabilities P_i are defined by equations (up to a normalization factor):

$$P_0 = 1$$
$$P_1 = A + Bp_0$$
$$P_2 = \frac{(A + Bp_0)(A + Bp_1)}{2}$$

The average revenue H equals:

$$H = \frac{A(P_0 + P_1) \cdot K + B(p_0 P_0 + p_1 P_1) \cdot 1}{P_0 + P_1 + P_2} = \frac{2(AK + Bp_0 + (A + Bp_0)(AK + Bp_1))}{2 + (A + Bp_0)(2 + A + Bp_1)} \quad (1)$$

This expression is the ratio of two polynomials, each one with probabilities p_0 and p_1 included in the first power. Consider the expression (1) as a function of one of the probabilities p. After multiplying by the relevant constant we can get it into the form $(p + a)/(bp + c)$. The derivative of this expression has the form $(ac)/(bp + c)^2$. Consequently, the expression (1) has invariable sign in the range of values $(0, 1)$, and its extreme values are located at the ends of the interval, i.e. probabilities p_0 and p_1 can only take values 0 or 1. Therefore, the dynamic reservation strategy has advantage over fractional dynamic reservation strategy.

2.2 Dynamic reservation strategy is better than static reservation strategy

This model (strategy 1) is a special case of the previous one: you can reserve 0, 1 or 2 channels, corresponding to choice of probability (p_0, p_1) in the form of (1,1) (1,0) or (0,0), which, in own order, corresponds to the values of R of 0, 1 or 2. Accordingly, the revenue from formula (1) takes the form:

$$H_0 = \frac{2(AK + B)(1 + A + B)}{2(1 + A + B) + (A + B)^2} \quad (2)$$

$$H_1 = \frac{2(B + AK(1 + A + B))}{2 + (A + B)(2 + A)} \quad (3)$$

$$H_2 = \frac{2AK(1 + A)}{2 + 2A + A^2}$$

How many channels should be reserved? This depends on the parameters K, A and B. To find the optimal value of R, one should solve two equations pointing to the boundary of K:

$$H_0 = H_1 \quad \text{and} \quad H_1 = H_2 \,.$$

We get:

$$K_1 = 1 + \frac{2 + A + B}{A(1 + A + B)} \quad (4)$$

$$K_2 = 1 + \frac{2(A + 1)}{A^2}$$

It is easy to verify that for any values of A and B, the inequality $K_1 < K_2$ is true since:

$$K_2 - K_1 = \frac{2 + 2A + 2B + A^2 + AB}{A^2(1 + A + B)} \,.$$

Hence we have the following solution for optimal reservation R at $N = 2$ channels:

$$\begin{aligned} R &= 0 \quad \text{if} \quad K < K_1 \\ R &= 1 \quad \text{if} \quad K_1 < K < K_2 \\ R &= 2 \quad \text{if} \quad K_2 < K \end{aligned}$$

2.3 Static reservation

In the cases $R = 0$ and $R = 2$, this strategy does not differ from the strategy 1 above. Therefore, it remains to consider the case $R = 1$. This Markov model has four states (Fig. 2):

(00) – both channels are free,
(01) – the common channel is engaged,
(10) – the guard channel is engaged,
(11) – both channels are engaged.

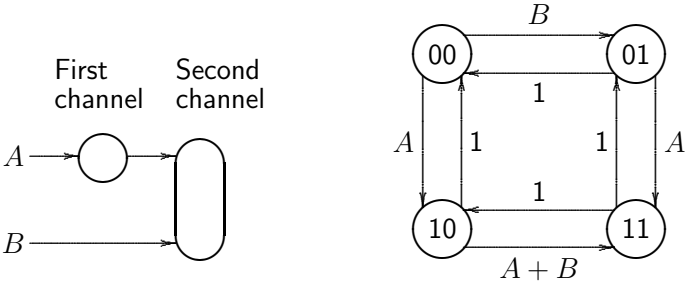


Fig. 2. Two channels, static reservation: the model and the state transition diagram.

From linear balance equations under the assumption of statistical equilibrium we obtain the state probabilities (up to a normalizing factor):

$$\begin{aligned} P_{00} &= 2 + 2A + B \\ P_{01} &= (A + B)^2 + 2B \\ P_{10} &= A(2 + A + B) \\ P_{11} &= A((A + B)^2 + A + 2B) \end{aligned}$$

The average revenue takes the form:

$$H_4 = \frac{A\,(2 + 4A + 3B + 2A^2 + 3AB + B^2) \cdot K + B\,(2 + 4A + B + A^2 + AB) \cdot 1}{2 + 4A + 3B + 3A^2 + 5AB + B^2 + A^3 + 2A^2B + AB^2}$$

Our aim is to prove that the static reservation strategy cannot be more profitable than the dynamic reservation strategy. That is, we must prove that for any values of A, B and K at least one of the values of H_0 and H_1 are not smaller than H_4.

It is easy to verify that by replacing K in this formula by K_1 from expression (4) we obtain $H_4 = 2$. We get the same result by substituting this value of K_1 for K in expressions (2) and (3), i.e. the equalities $H_0 = H_1 = H_4 = 2$ are true for any A, B and given $K = K_1$. Furthermore, we note that H_0, H_1 and H_4 are linear functions of K, i.e. straight lines. Note that for $K < K_1$ the inequalities $H_0 < H_4 < H_1$ are true. Hence these three straight lines intersect at one point, and the straight line H_4 is located between the two others. This means that for any values of A, B and K, at least one of the values of H_0 and H_1 is not less than the value H_4, q.e.d.

2.4 New call bounding scheme (strategy 4)

In case of a 2-channel system, the only nontrivial variant of strategy 4 (Restriction on number of B-calls admitted) is: no more than one B-call. The state transition diagram of this model

looks similar to that in Fig. 2, and the expression for average revenue is:

$$H_5 = \frac{2(A\,(A+B+1)\cdot K + B(A+1))}{2(A+B+1)+A(A+2B)}$$

This strategy is similar to strategy 3 (Static reservation). As shown in Fig. 3, the straight lines H_4 and H_5 are close: up to point $K < 1.25$ strategy 4 is a little more profitable, and from $1.25 < K$ strategy 3 is more profitable. But for any K, strategy 4 is worse than the optimal strategy 1, since $H_5 < H_0$ up to $K < 1.25$ and $H_5 < H_2$ from $1.25 < K$.

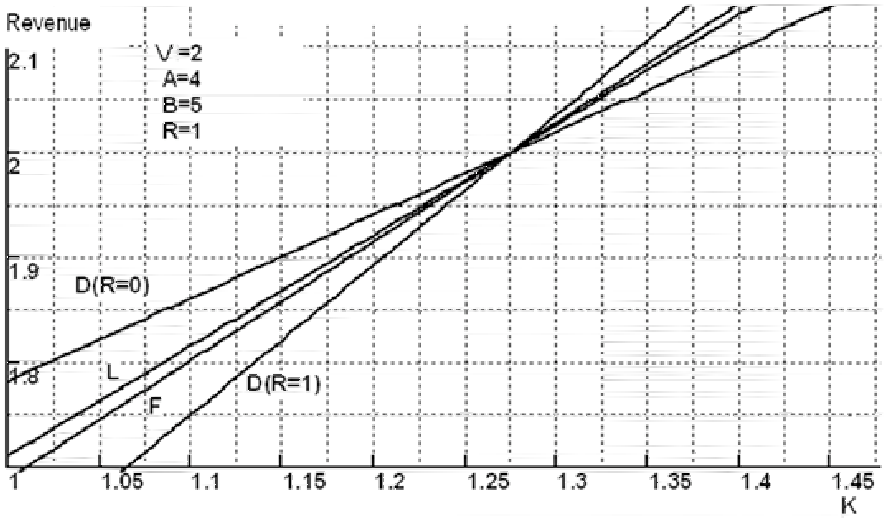


Fig. 3. Dependence of the revenue from handover call cost K for three models: D – dynamic reservation, F – static reservation, and L – restriction on number of admitted B-calls.

Conclusion. Hence it has been proved mathematically that the optimal service strategy in a two-channel system is dynamic reservation. Graphically, this fact is illustrated in Fig. 3 :

- reservation $R = 0$, optimal for values of $K \leq 1.25$, reservation line $D(R = 0)$,
- reservation $R = 1$, optimal for values of $K \geq 1.25$, reservation line $D(R = 1)$.

Problem 1. It is noteworthy that all four straights in Fig. 2 intersect at one point. This experimental fact deserves further study for systems with more than two channels.

3. Comparison of four strategies in single tier network: Common case

3.1 Fractional dynamic reservation

Theorem. For a N-channel loss system where B-calls are accepted with probability p_i, depending on the number of busy channels i $(i = 0, 1, \ldots, N)$, the optimal fractional dynamic reservation is limited to probabilities p_i equal to 0 or 1.

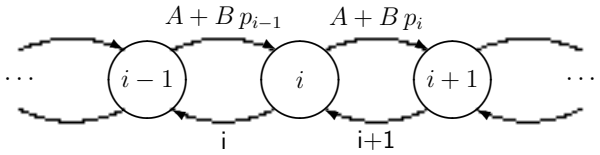


Fig. 4. Fractional dynamic reservation: the common case.

The stationary state probabilities F_i are defined by equations (up to normalization factor):

$$F_0 = 1,$$

$$F_1 = A + Bp_0,$$

$$F_2 = (A + Bp_0)(A + Bp_1)/2,$$

$$\ldots \quad \ldots$$

$$F_N = (A + Bp_0)(A + Bp_1)\ldots(A + Bp_{N-1})/N!$$

The lost revenue is equal to:

$$C = \frac{B \cdot (F_0 p_0 + F_1 p_1 + \ldots + F_{N-1} p_{N-1}) + (A\,K + B) \cdot F_N}{F_0 + F_1 + \ldots + F_N}$$

We should maximize the average revenue:

$$H = A \cdot K + B \cdot 1 - C\,. \tag{5}$$

Note that this expression is the general case of formula (1). As above, the expression (5) is the ratio of two polynomials, each of which includes the probability p_i in first power. Consider the expression (5) as a function of the probability p_i for any i. By the same argument as above we prove the Theorem.

Problem 2. The previous theorem does not imply which of the probabilities p_i are equal to 0 or 1. Common sense is that $p_i = 1$ for $i = 1, ..., R$, and $p_i = 0$ for $i = R + 1, ..., N$. How to prove this mathematically?

3.2 Dynamic reservation as maximum revenue strategy

We compare two strategies: dynamic and static reservation. On the basis of numerical results we have shown that with the optimal reservation R the expected revenue is always higher for the model with the dynamic reservation (Fig. 5). Naturally, when the handover call cost increases, then the number of reserved channels will increase. For $K = 2$ the optimal dynamic reservation is $R = 2$, and for $K = 4$ it is $R = 4$. The curves, of course. coincide at the ends of the definition interval when R is equal to 0 or N.

Numerical calculations (Fig. 6) show that strategy 4, restriction of number of B-calls admitted, is similar to strategy 3. For large values of R these strategies are almost identical, but even with the optimal value of R, strategy 4 has only a slight advantage over strategy 3.

Conclusion. The results of numerical analysis confirm that the optimal strategy is dynamic reservation. This statement is strictly proved in the case of a two-channel system.

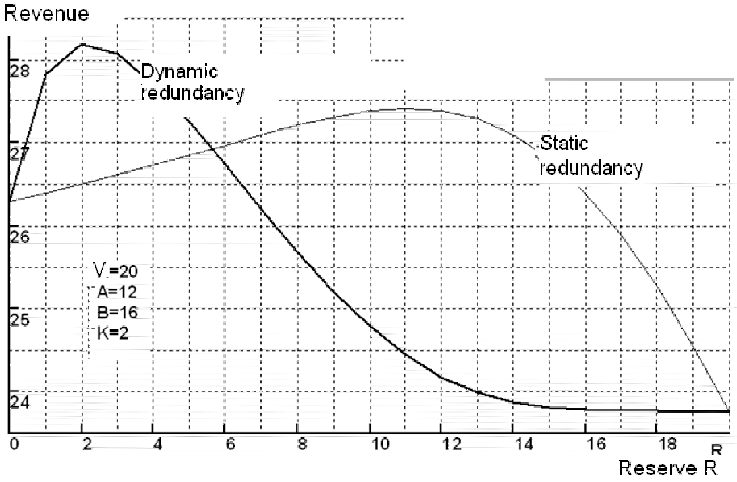


Fig. 5. Dependence of revenue on the size of the reservation R for the dynamic reservation D and fixed reservation F. The cost of handover call is $K = 2$.

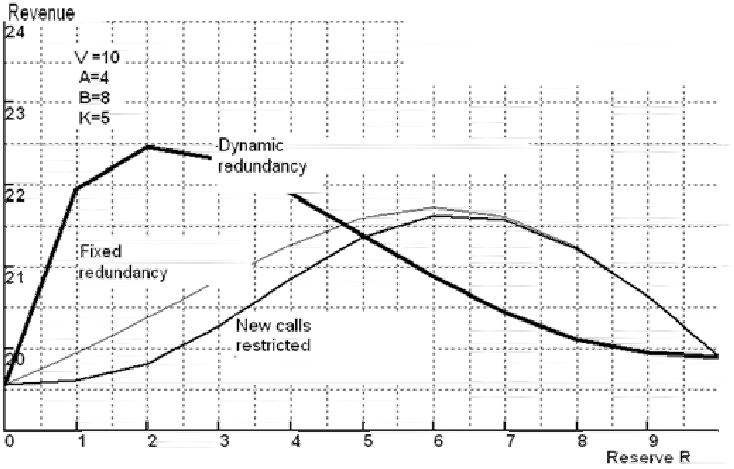


Fig. 6. Dependence of revenue on the size of the reservation for three strategies: $D =$ dynamic reservation, $F =$ static (fixed) reservation, and $L =$ restriction on the number of admitted B-calls.

3.3 On queueing effect

In queuing priority schemes, new calls and handover calls are all accepted whenever there are idle channels for that type of calls. When no idle channels are accessible, calls may be queued or blocked (i.e cleared from the system). Queueing priority schemes can be divided into three groups: new call queuing schemes (Chang et al., 1999), handover call queuing schemes (Yoon & Kwan, 1993); (Tian & Ji, 2001); (Agrawal et al., 1996), and all calls queuing schemes (Yoon & Kwan, 1993); (Chang et al., 1999).

Computational analysis has shown that waiting positions do not change the advantage of dynamic reservation strategy. Fig. 7 displays two pairs of curves: one pair is the same as in Fig. 6, the second one relates to a case with three waiting positions. Of course, the revenue is growing, but the preference of dynamic reservation keeps the place.

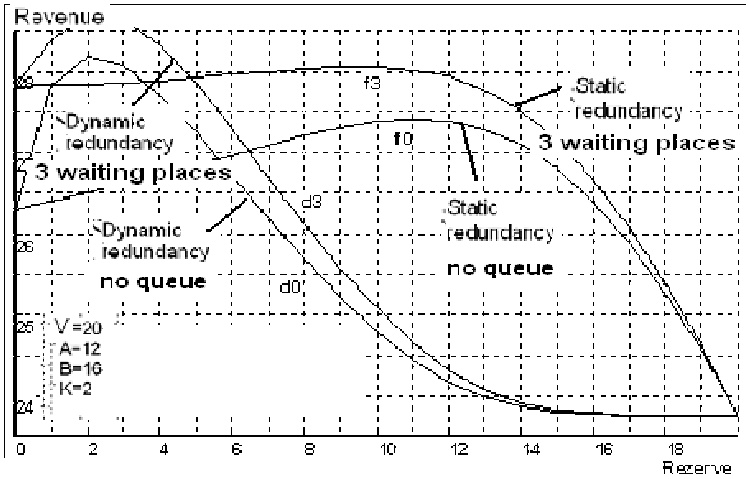


Fig. 7. Waiting positions do not change the advantage of dynamic reservation strategy. One pair of curves is the same as in Figure 6, the other pair relates to the case with three waiting positions.

4. Two-tier network

4.1 Dynamic reservation versus static reservation

There are several reasons for designing multi-tier cellular networks. One is to provide services for mobile terminals with different mobility and traffic patterns. The required performance measures can be met if the traffic and mobility patterns can be classified into more homogeneous parts and treated separately. Consider a system where there are two mobility classes. If the cell radii are optimized for low-mobility terminals, then the high-mobility terminals will have to make a lot of handovers during a communication session. On the other hand, if the optimization is made regarding the handover performance of high-mobility terminals, then the traffic load in each cell may exceed acceptable limits.

In multi-tier cellular networks, different layers offer the designer the opportunity of class based optimization. In case of a two-tier network we consider three or seven micro-cells overlaid by one large macro-cell offered calls from two mobility classes (Fig. 8). High-mobility

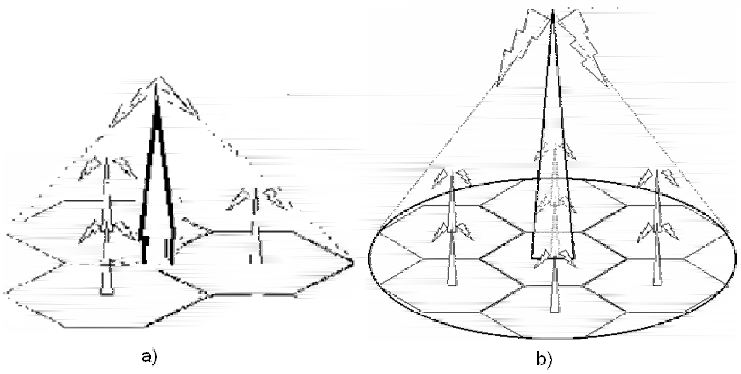


Fig. 8. A macro-cell covering (a) three and (b) seven micro-cells.

calls of intensity A are served by macro-cell only. Low-mobility calls of intensity B are served by the micro-cells as first choice and, if the reservation strategy admits it, by the macro-cell as second choice. Arriving calls are served as follows. The mobility class of the call is identified. High-mobility calls of intensity A are served by macro-cell only. Low-mobility calls of intensity B are served by the appropriate micro-cell as first choice, and if reservation strategy allows it by macro-cell as second choice. In both cases, our optimization criteria is the same: to maximize the revenue when each served A-call costs K units and each served B-call costs one unit ($K > 1$). When calls reach the macro-cell level, they are no longer

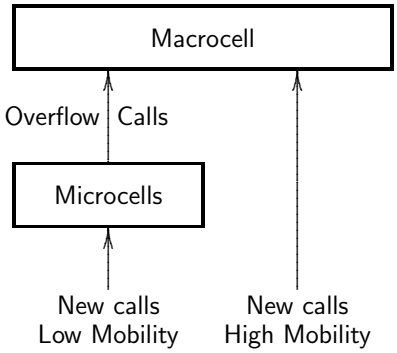
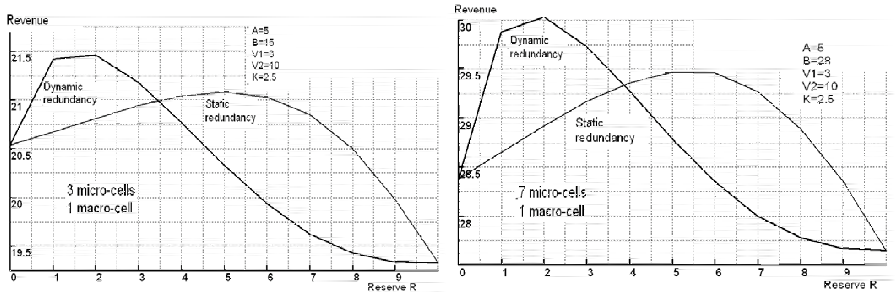


Fig. 9. A two-tier cellular network call flow model.

differentiated according to their mobility classes. Therefore, the calls of the high mobility class terminals and the overflowed handover calls from micro-cells are treated identically. New calls from micro-cells may not use the guard (reserved) channels upon their arrival. If no non-guard channel is available, then new calls are blocked. High mobility calls are blocked if all macro-cell channels are busy. Fig. 9 shows schematically how calls are served and what order is followed when serving them. As above in the case of single-tier network we compare two reservation strategies:

a) Dynamic reservation: The cutoff priority scheme is to reserve a number of channel for high-mobility calls in the macro-cell. Whenever a channel is released, it is returned to the common pool of channels.

b) Static reservation: Divide all macro-cell channels allocated to a cell into two groups: one for the common use by all calls and the other for high-mobility calls only (the rigid division-based CAC scheme).



(a) (b)

Fig. 10. Dependence of the revenue on the reserved number of channels for two-tier networks: (a) Three micro-cells and one macro-cell, (b) Seven micro-cells and one macro-cell.

Numerical results for two two-tier network examples are obtained. They are not qualitatively different from the results of the one-tier model discussed above. Fig. 10.a (three micro-cells and one macro-cell) and Fig. 10.b (seven micro-cells and one macro-cell) show that the dynamic reservation strategy gives the higher maximum revenue in both cases if the reserved number of channels R is properly chosen. The parameters are as follows: $A =$ high-mobility call flow, $B =$ low-mobility call flow, $N_1 =$ number of micro-cell channels for each cell, $N_2 =$ number of micro-cell channels.

Conclusion: In case of two-tier network, the results of numerical analysis confirm that the optimal strategy is dynamic reservation.

4.2 Channel rearrangement effect

In hierarchical overlaying cellular networks, traffic overflow between the overlaying tiers is used to increase the utilization of the available capacity. The arrival process of overflow traffic has been verified to be correlated and bursty. This characteristic has brought great challenges to performance evaluation of hierarchical networks. In most published works, the discussion is focussed on traffic loss analysis in homogenous hierarchical networks, e.g. micro/macro cellular phone systems as in the numerical analysis below. In the paper (Huang et al., 2008), the authors address the problems of performance evaluation in more complicated scenarios by taking account of heterogeneity and user mobility in hierarchical networks. They present an approximate analytical loss model. The loss performance obtained by our approximated analytical model is validated by simulation in a heterogeneous multi-tier overlaying system.

Fig. 12 shows the dependence of revenue on channel rearrangement from macro-cell to micro-cell. We are looking for maximum revenue when low-mobility calls cost one unit and high-mobility calls cost $K = 3$ units.

4.3 On optimal channel distribution (future study)

Fig. 13 shows two arrangements each of 18 radio channels for use by 4 call streams. Fig. 13.a shows a two-tier network with 3 individual channels per stream in the first tier and 6 common

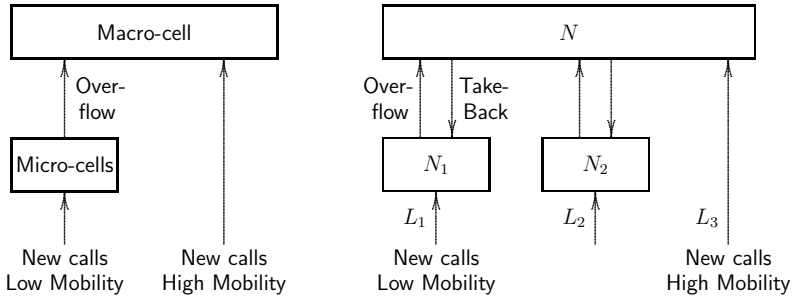


Fig. 11. An illustration of channel rearrangement.

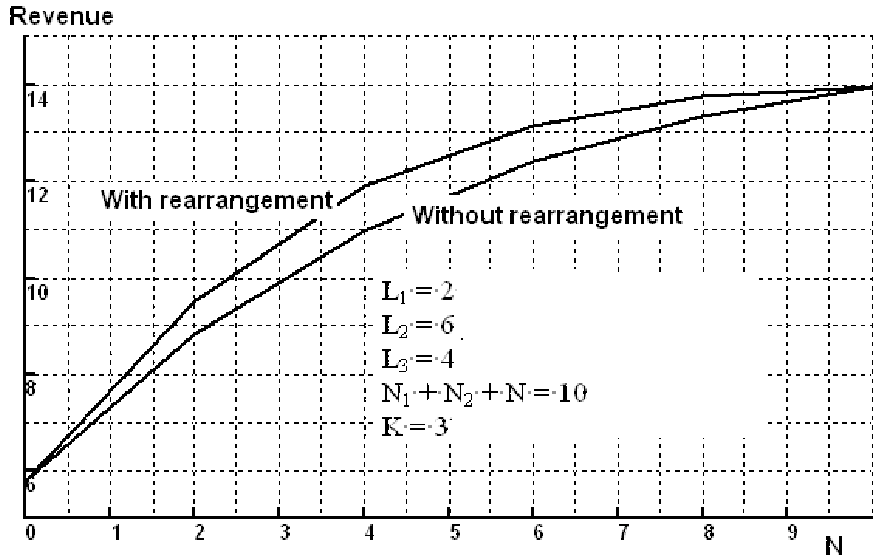
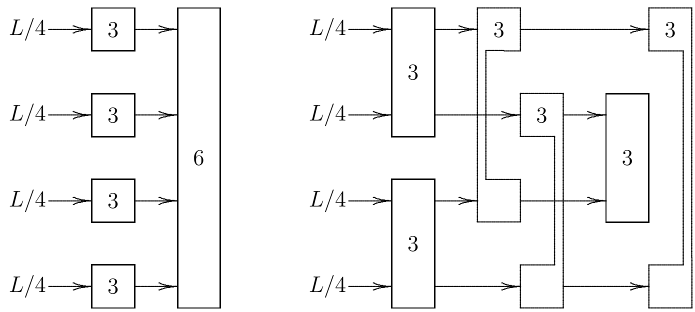


Fig. 12. Dependence of revenue on channel rearrangement from macro-cell to micro-cell.

channels in the second tier. In Fig. 13.b some kind of a homogeneous single tier network is depicted: each call has access to 9 channels equally distributed between streams. Such kind of arrangement could be implemented by modern DSP techniques.

Fig. 14 depicts the loss probability curves for these two schemes. Case (a) relates to pure loss system, case (b) relates to scheme with one waiting positions per stream. What is surprising? In case (a), beginning with a loss probability as low as 0.56% (less than 1%), it is advantageous to use the equally distributed scheme. Therefore, the traditional two-tier network could be recommended here at a very low call rates. Table 1 contains more detailed data on loss probability. When a single waiting position is added, the advantage of the equally distributed scheme increases even more and the cross point of curves occurs at the loss probability equal to 0.025%.

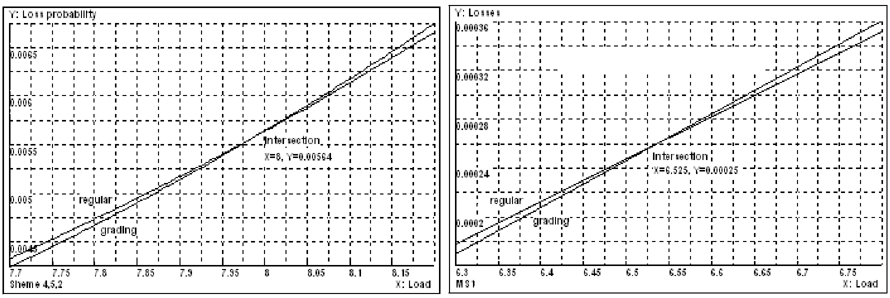


(a) Grading (b) Regularn scheme

Fig. 13. (a)Grading type two-tier network, and (b) single-tier network with equally distributed channels.

Load	Grading	Regular
1	1.0028 10-E11	6.8573 10E-11
3	5.3389 10E-7	1.4922 10E-6
7	1.9967 10E-3	2.1671 10E-3
11	4.0731 10E-2	3.6815 10E-2
15	0.13967	0.12676
20	0.27585	0.25885
25	0.38662	0.37100

Table 1. Loss probabilities for the two schemes of Fig. 13 (no waiting positions).



(a) (a) No waiting positions (b) (b) Single waiting position

Fig. 14. (a) Comparison of channel arrangement between cells for loss system: grading type two-tier network (Fig. 14.a) is preferable for low rates only, but as load grows the scheme (Fig. 14,b) becomes preferable; (b) the same for the case of one waiting position per stream.

A historical remark regarding this strange phenomenon. It goes almost 50 years back, to the results of V.E. Beneš, a distinguished mathematician from Bell Laboratories, who worked on modeling so-called crossbar telephone switches. In (Beneš, 1966) he writes: "*The question has arisen, whether there are examples of pairs of networks, with the same number of cross-points, the first of which is better than the second at one value of L, while the second*

is better than the first at another value of L". The same type of problem was a goal of the studies (Schneps-Schneppe, 1963), but only for the earlier telephone exchange generation with step-by-step switches. In any case, the question remains related to preference of multi-tier networks in comparison with equally distributed schemes, taking modern digital signal processing (DSP) techniques into account.

5. Multi-tier network dimensioning by Equivalent Random Traffic Method

5.1 Introduction

Let us recall some ITU-T documents relating to the dimensioning of circuit groups in traditional telephone networks. These documents deal with dimensioning and service protection methods taking traffic routing methods into account. Recommendations E.520, E.521, E.522 and E.524 deal with the dimensioning of circuit groups with high-usage or final group arrangements.

Recommendation E.520 deals with methods for dimensioning of single-path circuit groups based on the use of Erlang's formula (Fig. 15.a).

Recommendations E.521 and E.522 provide methods for the dimensioning of simple alternative routing arrangements as the one shown in Fig. 15.b, where only first and second-choice routes exist, and where all the traffic overflowing from a circuit group is offered to the same circuit group. Recommendation E.521 provides methods for dimensioning the final group satisfying GoS (Grade-of-Service) requirements for a given capacity of the high-usage circuit groups. Recommendation E.522 advises on how to dimension high-usage groups to minimize the cost of the whole arrangement. Fig. 15.b shows a two-tier network. For the dimensioning of such simple alternative routing arrangements the ERT-method is applicable.

Recommendation E.524 provides overflows approximations for non-random traffic inputs which allows for the dimensioning of more complex arrangements, e.g. three-tier network with correlated streams as shown in Fig. 15.c). The extended ERT-method described below relates to this case and could serve as a basis for a revision of Recommendation E.524.

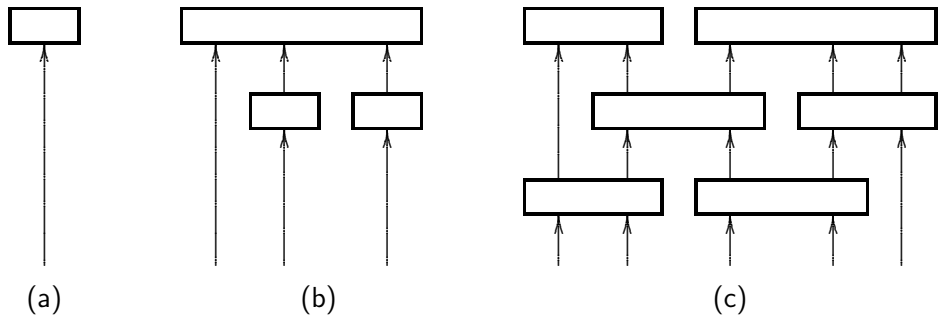


Fig. 15. Examples of three network arrangements: single-tier, two-tier, and three-tier with correlated streams.

5.2 Erlang formula and its generalization for non-integer number of channels

We consider a system of N identical fully accessible channels (servers, trunks, slots, call center agents, pool of wavelengths in the optical network, etc) offered Poisson traffic L and operating as a loss system (blocked calls cleared). The probability that all N channels are busy at a random point of time is equal to:

$$B = E(N, L) = \frac{\frac{L^N}{N!}}{\sum_{i=0}^{N} \frac{L^i}{i!}} \tag{6}$$

This is the famous Erlang-B formula (1917) (Iversen, 2011). For numerical analysis of (6) we use the well-known recurrent formula:

$$E(N+1, L) = \frac{L \cdot E(N, L)}{N + 1 + E(N, L)} \tag{7}$$

with initial value$E(0, L) = 1$. For the ERT-method we need Erlang-B formula for non-integral number of channels. How to get solution for a non-integral N? The traditional approach is based on the incomplete gamma function using:

$$E(N, L) = \frac{L^N \cdot e^{-L}}{\Gamma(N+1, L)} \tag{8}$$

where

$$\Gamma(N+1, L) = \int_L^\infty t^N e^{-t} \mathrm{d}t$$

We propose a new approach for engineering applications. Let the value N be from the interval (0,1). We introduce a parabolic approximation for $\ln R = lnE(N, L)$ at points $N = 0$, $N = 1$, and $N = 2$:

$$\begin{aligned} \ln E(0, L) = \ln(1) &= 0\,, \\ B = \ln E(1, L) &= \ln \frac{L}{L+1}\,, \\ C = \ln E(2, L) &= \ln \frac{L^2}{L^2 + 2L + 2}\,. \end{aligned}$$

Then we get the requested approximation:

$$\ln E(N, L) = \left(\frac{C}{2} - B\right) N^2 + \left(2B - \frac{C}{2}\right) N$$

or in a more convenient form

$$E(N, L) = L^N \, (L+1)^{N^2 - 2N} (L^2 + 2L + 2)^{\frac{N - N^2}{2}} \tag{9}$$

Thus we have an initial value of $E(N, L)$ for N inside the interval (0,1) and we may calculate $E(N, L)$ at any N by means of recurrent formula (7). In (Schneps-Schneppe & Sedols, 2010) the proposed approximation (6) is compared numerically with earlier known Erlang-B formula approximations (Rapp, 1964); (Szybicki, 1967); (Hedberg, 1981) and it is shown to be more accurate.

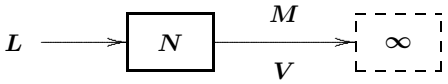


Fig. 16. Kosten's model: N fully accessible channels and a channel group of infinite capacity.

5.3 Kosten's model and its new interpretation

The basic idea the Equivalent Random Traffic (ERT) method is to use Erlang-B formula for overflow traffic offered to a secondary channel group of infinite capacity (Fig. 16), the so-called Kosten model (Kosten, 1937). Kosten's paper contains formulae for all binomial moments of number of busy channels in secondary overflow group. In practice, only the two first moments are used for characterization of the overflow traffic: mean traffic intensity M and variance V (reference is usually made to Riordan's paper (Riordan, 1956)):

$$M = L \cdot E(N, L) \tag{10}$$

$$V = M \cdot \left(1 - M + \frac{L}{N + 1 - L + M}\right) \tag{11}$$

From these two parameters one introduce a new parameter Z, the so-called peakedness:

$$Z = \frac{V}{M} = \left(1 - M + \frac{L}{N + 1 - L + M}\right) \geq 1 \tag{12}$$

Experience shows that peakedness Z is a very good measure for the relative blocking probability a traffic stream with given mean value and variance is subject to.

We offer a new interpretation of Kosten's results. We consider the scheme in Fig. 17. There are N common channels and one separate channel. From (10 and (11 we get a new formula for the variance V when both mean M and mean $M_1 = L \cdot E(N + 1, L)$ are known. From recurrence

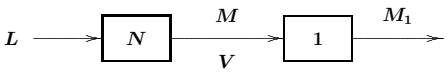


Fig. 17. An illustration of the ERT-method extension.

formula (7) follows:

$$M_1 = \frac{L\,M}{N + 1 + M}$$

or

$$M + N + 1 = \frac{L\,M}{M_1}$$

After substitution of $(M + N + 1)$ into (11) we get:

$$V = M \cdot \left(1 - M + \frac{L}{\frac{L\,M}{M_1} - L}\right) = M\left(1 - M + \frac{M_1}{M - M_1}\right) \tag{13}$$

We can reduce this expression to a simpler form:

$$V = M^2\left(\frac{1}{M - M_1} - 1\right) = M\left(\frac{M}{M - M_1} - M\right) \tag{14}$$

Note that $M - M_1$ is the load carried by a single channel and therefore it is always less than one. Formula (14) is useful for applications of the ERT-method in case of traffic splitting.

5.4 ERT-method

The ERT-method has been developed for planning of alternate routing in telephone networks by several authors: (Wilkinson, 1956);(Bretschneider, 1973);(Fredericks, 1980); and others. Fig. 18 explains the essence of the method. In (Fig. 18.a) g traffic streams which may for example be overflow traffic from other exchanges are offered to a transit exchange. As it is non-Poisson traffic, it cannot be described by classical traffic models. We do not know the distributions (state probabilities) of the traffic streams, but we are satisfied (most often the case in applications of statistics) by describing the i-th traffic stream by its mean value M_i and variance V_i. The aggregated overflow process of the g traffic streams is said to be equivalent to

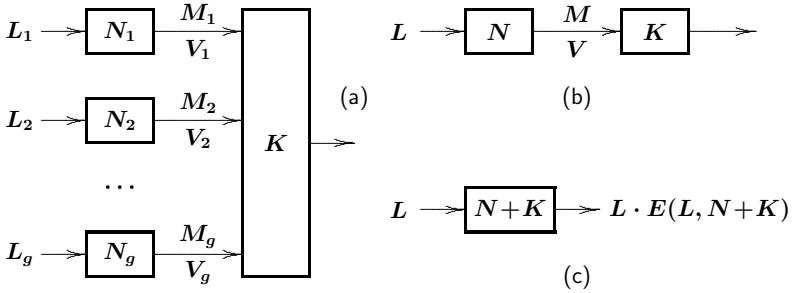


Fig. 18. Application of the ERT-method: (a) g independent traffic streams offered to a common group of K channels, (b) equivalent group, (c) Erlang-B formula applied to a common group with $N + K$ channels.

the traffic overflowing from a single full accessible group with the same mean and variance as the total overflow traffic. The total traffic offered to the group with K channels has the mean value:

$$M = \sum_{i=1}^{g} M_i$$

We assume that the traffic streams are independent (non-correlated), and thus the variance of the total traffic stream becomes:

$$V = \sum_{i=1}^{g} V_i$$

Therefore, the total traffic is described by M and V. We now consider this traffic to be equivalent to a traffic flow which is lost from a full accessible group and has same mean value M and variance V (Fig. 18.b). For given values of M and V, we therefore solve equations (10) and (11) with respect to N and L. Then it is replaced by the equivalent system (Fig. 18.c) which is a full accessible system with $(N + K)$ channels offered the traffic L.

5.5 On accuracy of the ERT-method

Let us give a computational analysis of the classical ERT-method by a three-tier network shown in Fig. 19. There are four streams each offering a traffic equal to 5 erlang traffic. On first tier there are two servers per stream, on second tier there are three servers, and on third tier two servers. Application of the ERT-method to dimension the alternate routing networks consists of three steps.

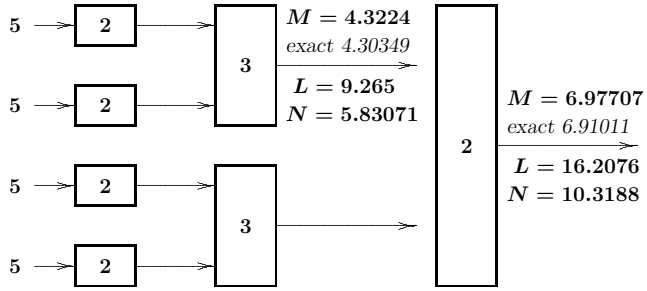


Fig. 19. Three-tier network.

1. First step is a direct application of formulae (10) (11) for two streams and two individual lines (parameters: $L = 5, N = 2$).
2. The second step, we apply the formulae (10) (11) to a three-channel group. We get the equivalent parameters $L = 9.265$, $N = 5.83071$ and the lost traffic $M = 4.3224$. The exact value given in brackets is obtained by solving the system of equations of the Markov process, and is equal to 4.30349, i.e. the relative error is less than 1%.
3. Third step: The two overflow streams are fed into the two lines. We get the equivalent parameters: $L = 16.2076$, $N = 10.3188$, and the lost traffic $M = 6.97707$. The exact value (in brackets) is 6.91011, i.e. the relative error again is less than 1%.

The results of calculations show the excellent accuracy of the method. However, such accuracy is not preserved when the number of channels in the third step increases. If instead of two channels we have $(2 + g)$ channels in the common group, then Table 2 shows values of the loss for different g values. It is obvious the accuracy drops when increasing g. For a value of $g = 10$, the relative error is bigger than 3%, but always on the safe side, and the absolute loss probability is very small.

g	Loss probability Exact	Loss probability ERT-method	Relative error %
0	0.4083	0.4105	0.539
2	0.3239	0.3277	1.173
4	0.2459	0.2506	1.911
6	0.1765	0.1812	2.663
8	0.1181	0.1218	3.133
10	0.0724	0.0747	3.177

Table 2. On accuracy of the classical ERT-method.

5.6 Fredericks & Hayward's ERT-method

In (Fredericks, 1980) an equivalence method is proposed which is simpler to use than Wilkinson-Bretschneider's method. The motivation for the method was first put forward by W.S. Hayward. For given values of (M, V) of a non-Poisson flow, Frederick & Hayward's approach implies direct use of Erlang's formula $E(N, M)$, but with scaling of its parameters as $E(N/Z, M/Z)$. The scaling parameter $Z = V/M$ is the peakedness (12) (Fig. 20).

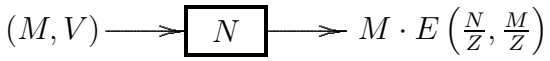


Fig. 20. An illustration of Fredericks & Hayward's approach.

The accuracy of Fredericks & Hayward approach is numerically compared with ERT and with exact values. The calculations were performed for different variants of the scheme shown in Fig. 21. In general, its accuracy is comparable to that of the Wilkinson approach (Table 3). However, in our opinion Wilkinson's approach is more reliable and always yields worst-case values.

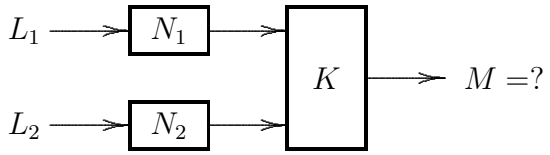


Fig. 21. On accuracy of Fredericks & Hayward's approach.

L_1	N_1	L_2	N_2	K	M (Wilkinson)	M (Hayward)	M (exact)
1	2	2.5	0	1	1.9740	1.9723	1.9721
6	2	2.5	0	1	5.9635	5.9620	5.9589
8	3	4	4	5	2.8717	2.8514	2.8498
3	3	4	4	5	0.2614	0.2268	0.2608
2	3	4	4	5	0.1122	0.0859	0.1140
4	3	2	4	5	0.1636	0.1386	0.1640
2.5	7	8.25	4	9	0.30268	0.2884	0.30273

Table 3. On accuracy of Fredericks & Hayward's approach.

5.7 Correlation of overflow streams

As it is, the ERT-method is not applicable to the analysis of multi-tier networks with correlated streams as shown in Fig. 15.c. In 1960's, this type of problem appeared when dimensioning so-called gradings, the basic structural block in step-by-step exchanges. An important result was developed independently by (Descloux, 1962) and (Lotze, 1964). They determined mean M and variance V of the overflow stream components when split up after a first choice group as shown in Fig. 22.b. On the basis of Kosten's model with two parameters M and V, they developed a 5-parameter model for the two stream case: mean values M_1 and M_2, variances V_1 and V_2, and covariance Cov.

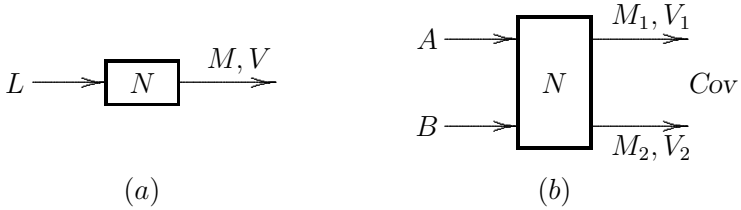


Fig. 22. (a) Kosten's model, (b) two-stream model.

If $A = p_1\,L$ and $B = p_2\,L$, where $p_1 + p_2 = 1$, then the peakedness of partial stream V_i/M_i is defined by total peakedness V/M:

$$\frac{V_i}{M_i} - 1 = p_i \left(\frac{V}{M} - 1\right),$$

and covariance

$$Cov = p_1\,p_2(V - M)\,.$$

The covariance formula is proved according to the theorem of variance for mutually dependent variables

$$V = V_1 + V_2 + 2 \cdot Cov\,.$$

5.8 Correlated streams: Neal's formulae

During early 1970's, Scotty Neal from Bell Labs studied the covariance of correlated streams in alternative routing networks (Neal, 1971). Below we use results from Neal's paper to develop some formulae in notations of Fig. 23. We are looking for covariance between two overflow streams after groups with D and F channels, respectively. The key to Neal's solutions is the original work (Kosten, 1937). Neal (Neal, 1971) extended the ERT-method to mutually dependent streams. On basis of the extended Kosten' model (Fig. 23) he developed a technique for taking correlation into account when combining dependent streams of overflow traffic. More precisely, Neal has considered a 5-parameter Markov model (Fig. 23) with 5 parameters:

1. Number of busy channels in the first choice group (up to N),
2. Number of busy channels i in the first alternate group $(0 \le i \le D)$,
3. Number of busy channels j in the second alternate group $(0 \le j \le F)$,
4. Number of busy channels in the first imaginary infinite channel group,
5. Number of busy channels in the second imaginary infinite channel group.

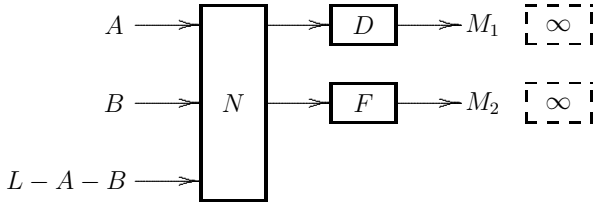


Fig. 23. An illustration to Neal's formulas.

After rather sophistical derivations of two-dimensional binomial moment generating functions and using Kosten's approach, Neal obtains linear equations for the two-dimensional probabilities $\beta(i,j)\,, (0 \le i \le D\,,\ 0 \le j \le F)$. The initial value is:

$$\beta(0,0) = E(N,L) \tag{15}$$

and other probabilities are defined by linear balance equations:

$$(i+j)v_{i+j}\beta(i,j) \;= A\,\beta(i-1,j) + B \cdot \beta(i,j-1) - A\binom{D}{i-1}\beta(D,j) - B\binom{F}{j-1}\beta(i,F)\,,$$

$$\text{where} \quad 0 \le i \le D\,, \quad 0 \le j \le F, \quad i+j > 0\,, \tag{16}$$

with the following recurrent formulas for coefficients ν_i:

$$\nu_i = \frac{L}{i \cdot \nu_{i-1}} + 1 + \frac{N-L}{i}, \quad i > 0, \quad \text{where} \quad \nu_0 = \frac{1}{E(N,L)}. \tag{17}$$

Then

$$M_1 = A \cdot \beta(D,0) \tag{18}$$

$$M_2 = B \cdot \beta(0,F)$$

$$Cov = \frac{A\,Q_1 + B\,Q_2}{2} - M_1\,M_2 \tag{19}$$

where

$$Q_1 = \frac{\displaystyle\sum_{j=0}^{F} \left(\beta(D,j) \prod_{k=j+1}^{F+1} \frac{B}{k\,\nu_k} \right)}{1 + \displaystyle\sum_{j=1}^{F} \left(\binom{F}{j-1} \prod_{k=j+1}^{F+1} \frac{B}{k\,\nu_k} \right)}$$

$$Q_2 = \frac{\displaystyle\sum_{j=0}^{D} \left(\beta(j,F) \prod_{k=j+1}^{D+1} \frac{A}{k\,\nu_k} \right)}{1 + \displaystyle\sum_{j=1}^{D} \left(\binom{D}{j-1} \prod_{k=j+1}^{D+1} \frac{A}{k\,\nu_k} \right)}$$

Based on Neal's formulae (15) – (19 we get the lost stream intensities M_1 and M_2 and the variance V. From (18) follows that loss probability of the first stream is $\beta(D,0)$.

5.9 Comments on Neal's results

In the following discuss the applicability of Neal's results.

5.9.1 Algorithm

We extract the equations which have $j = 0$ from the equation system (16). Eliminating members with zero coefficients and considering that $\beta(0,0)$ is known, we obtain the system with D equations referring to $\beta(i,0)$:

$$i\,\nu_i\,\beta(i,0) = A\,\beta(i-1,0) - A \binom{D}{i-1} \beta(D,0). \tag{20}$$

By solving this system of linear equations we get expressions for $\beta(i,0)$:

$$\beta(D,0) = \frac{1}{\displaystyle\sum_{i=0}^{D} \frac{D!}{(D-i)!\,A^i} \prod_{j=0}^{i} \nu_j}, \tag{21}$$

$$\beta(i,0) = \beta(D,0) \binom{D}{i} \left(1 + \sum_{j=1}^{D-i} \prod_{k=1}^{j} \frac{\nu_{i+k}(D+1-i-k)}{A} \right), \qquad 0 < i < D \tag{22}$$

Values ν_j are obtained by formula (17). Using direct test we can ascertain that (21) and (22) together with statement (15) indeed satisfies the system of equations (20).

5.9.2 The modified Erlang formula

Formula (21) in the form

$$\beta(D,0) = \frac{\dfrac{A^D}{D!}}{\sum\limits_{i=0}^{D} \dfrac{A^{D-i}}{(D-i)!} \prod\limits_{j=0}^{i} \nu_j} \tag{23}$$

is an obvious analogy to the Erlang formula for the scheme shown in Fig. 23. This formula is applicable to the ERT-method. It allows for non-integral number of channels.

From this the mean intensity M_1 follows:

$$M_1 = \frac{\dfrac{A^{D+1}}{D!}}{\sum\limits_{i=0}^{D} \dfrac{A^{D-i}}{(D-i)!} \prod\limits_{j=0}^{i} \nu_j} \tag{24}$$

Formula (24) is easily implemented and allows for non integer values of N.

5.9.3 Variance

By analogy of (13) we get the variance:

$$V_1 = M_1 \left(1 - M_1 + \frac{M_{1+}}{M_1 - M_{1+}}\right) \tag{25}$$

where

$$M_{1+} = A\,\beta(D+1,0)$$

Substituting A by B and D by F in (24) and (25) we get M_2 and V_2 in similar way.

5.10 Extended ERT-method. Numerical example

Consider an example where the extended ERT-method can be used and the covariance obtained by using formulas (18) - (19). Let us calculate the mean value M of the overflow traffic for the following scheme (Fig. 24).

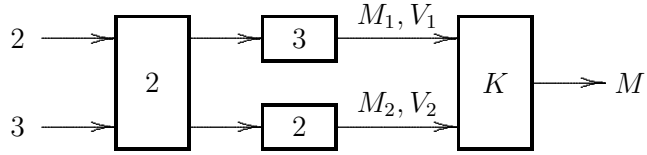


Fig. 24. Scheme with correlated streams.

Using formulas (24) – (25) we find:

$M_1 = 0.180018,\ \ M_2 = 0.873221,\ \ V_1 = 0.252930,\ \ V_2 = 1.25284\,.$

Using (18) - (19) we calculate the covariance of the two streams:

$Cov = 0.0282596\,.$

We now can calculate the intensity of the flow which is overflowing to the group with K

channels:

$$M^* = M_1 + M_2 = 1.053239, \quad V^* = V_1 + V_2 + 2\,Cov = 1.56229\,.$$

Using the extended ERT-method we get the equivalent group:

$$L^* = 3.33306, \quad N^* = 3.44900\,.$$

Therefore, using Erlang-B formula

$$M = L^* \cdot E(N^* + K, L^*)\,.$$

We can obtain mean intensity M of overflow stream for various values of K as shown in Table 4. The results of calculations show the excellent accuracy of the extended ERT-method.

K	**M** exact	**M** by ERT	Rel. error %
1	0.63322	0.63801	0.756
2	0.34583	0.34936	1.019
3	0.17067	0.17128	0.357
4	0.07616	0.07492	1.634
5	0.03091	0.02929	5.215

Table 4. Accuracy of the Extended ERT-method for correlated streams.

However, such accuracy is not preserved when the number of channels K in the final group increases. Table 4 shows values of the loss for different K values. It is obvious that for increasing K the accuracy drops. For the value $K = 5$ the relative error is greater than 5%. The same effect one observes in Table 1 and Table 2. For decreasing (very small) blocking probabilities the accuracy increases, but the absolute error decreases.

6. Conclusions

There are two parts in this chapter: a theoretical one and an application one. In the theoretical part, we consider four strategies for call admission control (CAC) in single and two-tier cellular networks, which are designed to ensure priority of handover calls, namely: dynamic reservation (cutoff priority scheme), fractional dynamic reservation (fractional guard channel scheme), static reservation (fixed division-based CAC scheme) and restriction on the number of simultaneous new calls admitted (new call bounding scheme). We show the advantage of dynamic reservation by numerical analysis and strictly prove it in the case of two-channel system with blocking.

The application part deals with Equivalent Random Traffic method for multi-tier networks dimensioning. The key result is an extension of the Equivalent Random Traffic method for estimation of the throughput for networks with traffic splitting and correlated streams. The excellent accuracy (relative error less than 1%) is illustrated by numerical examples.

7. References

Aazhang, B. & Cavallaro,J.R. (2001). Multitier Wireless Communications. *Wireless Personal Communications*, Vol. 17 (2001), 323–330.

Agrawal, P.; Anvekar, D.K. & Naredran, B. (1996). Channel management policies for handovers in cellular networks, *Bell Labs Technical Journals*, 1996, 96–110.

Ahmed, M.H. (2005). Call admission control i wireless networks: a comprehensive survey. *IEEE Communications Surveys*, Vol. 7 (2005), 50–69.

Beigy, H. & Meybodi, M.R. (2003). User based call admission control policies for cellular mobile systems: a survey. *CSI Journal on Computer Science and Engineering*, Vol. 1 (2003), 45–58.

Beigy, H. & Meybodi, M.R. (2004). A new fractional channel policy. *Journal of High Speed Networks*, Vol. 13 (2004), 25–36.

Beigy, H. & Meybodi, M.R. (2005). A general call admission policy for next generation wireless networks. *Computer Communications*, Vol. 28 (2005), 1798–1813.

Beneš, V.E. (1966). Some examples of comparisons of connecting networks. *Bell System Techn. J.*, Vol. 45 (1966), No. 10, 1829–1935.

Bretschneider, G. (1973). Extension of the equivalent random method to smooth traffics. *Proceedings of Seventh International Teletraffic Congress*, Stockholm, June 1973, paper 411. 9 pp.

Chang, C.; Chang, C.J & Lo, K.R. (1999). Analysis of hierarchical cellular systems with reneging and dropping for waiting new calls and handover calls. *IEEE Transactions on Vehicular Technology*, Vol. 48 (1999), 1080–1091.

Cruz-Pérez, F.A.; Toledo-Marín, T. & Hernández-Valdez, G. (2011). Approximated Mathematical Analysis Methods of Guard-Channel-Based Call Admission Control in Cellular Networks. *Cellular Networks - Positioning, Performance Analysis, Reliability*. Edited by: A. Melikov, ISBN 978-953-307-246-3, InTech, 2011.

Descloux, A. (1962). On the components of overflow traffic. *Internal Memorandum. Bell Telephone Laboratories Inc.*, December 1962. 6 pp.

Fang, Y. & Zhang, Y. Call admission control schemes and performance analysis in wireless mobile networks. *IEEE Transactions on Vehicular Technology*, Vol. 51 (2002), 371–382.

Fredericks, A.A. (1980). Congestion in blocking systems – a simple approximation technique. *The Bell System Tech. J.*, Vol. 59 (1980), No. 6, 805–827.

Ghaderi; M. & Boutaba, R. (2006). Call admission control in mobile cellular networks: a comprehensive survey. *Wireless Communications and Mobile Computing*, Vol. 6 (2006), 69–93.

Guerin, R. (1988). Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications*, Vol. 36 (1988), 153–163.

Haring, G.; Marie, R.; Puigjaner, R. & Trivedi, K. (2001) Loss formulas and their application to optimization for cellular networks. *IEEE Transactions on Vehicular Technology*, Vol. 50 (2001), 664–673.

Hedberg, I. (1981). A simple extension of the Erlang loss formula with continuous first order partial derivatives. *L.M. Ericsson, Internal Report* XF/Sy 81 171 (1981), 4 pp.

Hong, D. & Rappaport, S. (1986). Traffic modelling and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handover procedure. *IEEE Transactions on Vehicular Technology*, Vol. 35 (1986), 77–92.

Huang, Q.; Ko, K.-T.; Chan, S. & Iversen, V.B. (2008). Loss performance evaluation in heterogeneous hierarchical networks. *Mobility Conference 2008*, Vol. 16, doi>10.1145/1506270.1506291.

Iversen, V.B. (2011). *Teletraffic Engineering and Network Planning*. 382 pp. 2011 http://www.com.dtu.dk/education/34340/.

Kosten, L. (1937). Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen. *Elktr. Nachr.-Techn.*, Vol. 14, No. 1, 1937, 5–12.

Leong, C.W. & Zhuang, W. (2001). Call admission control for voice and data traffic in wireless communications. *Computer Communications*, Vol. 25, (2002), 972–979.

Lotze, A. (1964). A traffic Variance Method for Gradings of Arbitrary Type. *Proceedings of Fourth International Teletraffic Congress*, Document 80, London, 1964.

Neal, S. (1971). Combining correlated streams of nonrandom traffic. *The Bell System Tech. J.*, Vol. 50 (1971), No. 6, 2015–2037.

Oh, S. & Tcha, D. (1992). Prioritized channel assignment in a cellular radio network. *IEEE Transactions on Communications*, Vol. 40 (1992), 1259–1269.

Ramjee, R.; Towsley, D. & Nagarajan, R. (1997). On optimal call admission control in cellular networks. *Wireless Networks*, Vol. 3 (1997), 29–41.

Rapp, Y. (1964). Planning of junction network in a multi-exchange area. *Ericsson Technics* 1964, pp. 77–130.

Riordan, J. (1956). Derivation of moments of overflow traffic. Appendix 1 (pp. 507–514) in (Wilkinson, 1956).

Ryu, B.; Andersen, T.; Elbatt, T. & Zhang, Y. (2003). Multi-Tier Mobile Ad Hoc Networks: Architecture, Protocols, and Performance. *Proceedings of 2003 IEEE Military Communications Conference (MILCOM 2003)*, Vol. 2, pp. 1280–1285, , Monterey, California, October 2003.

Schneps-Schneppe, M. (1963). New principles of limited availability scheme design. *Elektrosviaz*, No. 7, 1963, 40–46 (in Russian).

Schneps-Schneppe, M. & Sedols, J. (2010). On Erlang B-formula and ERT-Method Extension. *ICUMT-2010*, Moscow, Oct 2010. (EDAS paper 1569334223).

Szybicki, E. (1967). Numerical methods in the use of computers for telephone traffic theory applications. *Ericsson Technics* 1967, 439–475.

Tian, X. & Ji, C. (2001). Bounding the performance of dynamic channel allocation with QoS provisioning for distributed admission control in wireless networks. *IEEE Transactions on Vehicular Technology*, Vol. 50 (2001), 388–397.

Wilkinson, R.I. (1956). Theories for toll traffic engineering in the U.S.A. *The Bell System Tech. J.*, Vol. 35, (1956), 421–514.

Yoon, C.H. & Kwan, C. (1993). Performance of personal portable radio telephone system with and without guard channels. *IEEE Journal on Selected Areas in Communications*, Vol. 11 (1993), 911–917.

7

Femtocell Performance Over Non-SLA xDSL Access Network

H. Hariyanto[1], R. Wulansari[1], Adit Kurniawan[2] and Hendrawan[2]
[1]TELKOM R&D Centre, Bandung
[2]Bandung Instittue of Technology, Bandung
Indonesia

1. Introduction

Femtocells are low-power wireless access points; operate in licensed spectrum and use residential or office DSL, cable or other broadband connections. Most mobile network operators (MNOs) offer femtocell access point (FAP) to retain their customers by improving indoor coverage and capacity. However in order to leverage massive femtocells deployment and generate new revenue, there should be business cases beyond the connectivity. Offering various femto services are crucial to strengthen customer value preposition and create new revenue generator for operators. Each service certainly requires a specific amount of bandwidth and QoS treatment. Therefore the study of bandwidth and QoS requirement for different traffic types are important, in order to design an optimum backhaul requirement for femtocell.

Heavy Reading in its report [1] stated that the cost of leased line for macrocell backhaul counted 25% of total MNO's capex. The need of small cells are paramount important in delivering high speed wireless broadband data. However the cost of new carrier-grade backhaul to support indoor base stations (IBSs) may increase depend on the new IBSs numbers and availability of leased lines. Femtocells utilize the existing broadband connection in the customer side. By this approach, the cost of backhaul can be reduced with the trade off fluctuation on the backhaul quality; if there is no specific service level agreement (SLA) setup between MNO and internet service provider (ISP).

Bear in-mind that femtocell is a CPE with self configure capabilities, so that it will impose less interaction with mobile operators. For residential users, they may buy the femtocell from the mobile operator or electronic store and instantly plug it to the existing broadband connection at home. The users may not be aware of how the fixed-wireline operator will treat the femto traffic compared to other best-effort internet traffic. They may not be alert to that other broadband traffic traversed via the same home gateway will affect femtocell service performance. Bottleneck may occur anywhere in the network and affect femtocell performance.

A comprehensive femtocell deployment guideline considering backhaul quality for 3G femtocell was addressed in [2]. The guideline describes the quality issue of VoIP services over 3G femtocell networks. VoIP services were observed as representation of real time

traffic. It was assumed that the users may complain to the cellular operator (instead of broadband IP provider) when they experience delay or poor Mean Opinion Score (MOS) during a voice call. As in 3G cases, users may wait for FTP data transfer or surfing the internet web site. In the latter case, higher latency or packet loss will not create a question from users than if the same situation experienced by users use VoIP or video services.

According to [2] and [3], most femtocell technologies provide good quality voice calls and sufficient support to data services when the broadband IP link provides a minimum performance of:

- Less than 150 ms round-trip delay (more than 200 ms will not be practical for two ways conversation);
- Less than 40 ms jitter;
- A general packet loss of 3% or less is acceptable; however, packet loss is typically "bursty" by nature, and, as such, average rates below 0.25% should be maintained;
- At least 1 Mbps in downlink, i.e. from the broadband IP provider network to the FAP GW;
- At least 256 kbps in uplink, i.e. from the FAP GW to the broadband IP provider network.

This chapter describes the femtocell performance over xDSL access network as the backhaul. This work has been conducted as part of TELKOM contribution to FREEDOM Project (www.ict-freedom.eu) which consists of two phases of measurements and analysis. The first phase addressed the performance of ADSL2, ADSL2+ as a function of distance. It also observes the population of user's density enjoying certain attainable rate or less. Furthermore it also addresses transmission delay of xDSL over different bandwidth profiles. DSL backhaul quality model is derived in order to address different qualities of backhaul. The model can be used in elaboration of RRM, scheduling and system level simulation which need to take into account the backhaul quality.

While in the first phase characterization, the measurement was conducted without femtocell, in the second phase we observed femtocell bandwidth requirement to support various basic services including HTTP, FTP, voice and video streaming. We limit the study for residential case where xDSL access network is used. In this case mobile network operator and xDSL provider do not have agreement to maintain end-to-end QoS, hence non-SLA terminology is used. It should be understood that the bottleneck is not always occurred in the low speed backhaul, but it may occur event in high speed backhaul link; if the wireline broadband service requires a huge amount of bandwidth (for example high definition IPTV, video surveillance for home monitoring, etc).

The study of femtocell bandwidth requirement aims to observe the individual bandwidth consumption according to basic communication traffic types including HTTP, voice, video and FTP. For this purpose, we measured the bandwidth for 4-unit FAP and calculate bandwidth requirement for higher capacity FAP types such as 8-unit FAP which may be used in apartment deployment case. The bandwidth requirement study will give some insight to customer as well as operator, who deal with limited backhaul bandwidth, in order to understand how far their backhaul is capable of delivering basic communication services.

Based on the measurement result reported in FREEDOM, we will show the effect of background traffic in xDSL modem to the femtocell performance. Increasing the bandwidth

may cope with the performance degradation but with the cost of adding more bandwidth to the existing broadband line. We also give some comments to the FREEDOM study outcome; how femtocell performance in non-SLA network can be improve by implementing backhaul aware scheduling and admission control [4].

2. xDSL characterization

xDSL technology offers fix broadband services over the existing copper twisted pair infrastructure. According to Organisation for Economic Co-operation and Development (OECD) as shown in Figure 1, the xDSL access technology has more subscribers compared to the other access technologies including fiber.

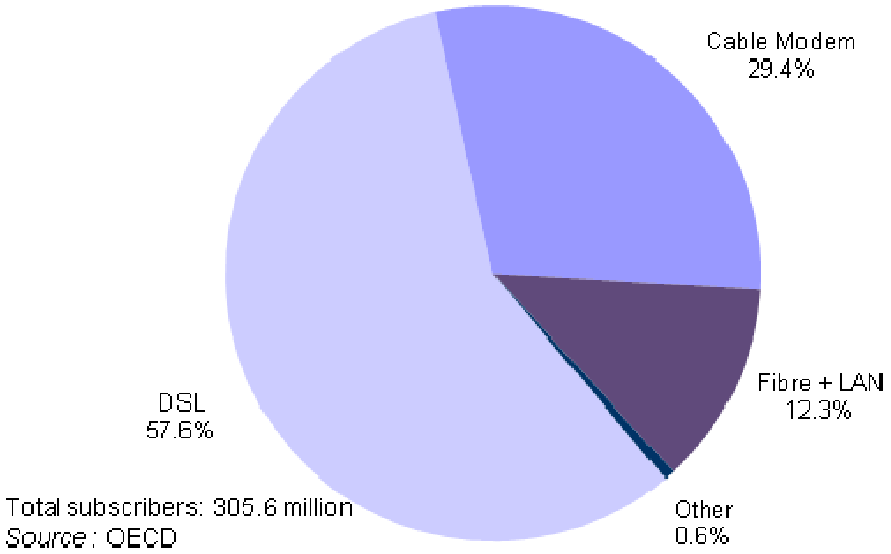


Fig. 1. OECD Fixed (wired) broadband subscriptions, by technology, Dec. 2010

The end user gets a dedicated link from xDSL modem to aggregation node called DSLAM (Digital Subscriber Line Access Multiplexer) or MSAN (Multi Switch Access network). There are several standards of xDSL which mostly asymmetric such as ADSL, ADSL2 and ADSL2+. The DSL also support symmetrical upstream and downstream ratio as in SDSL, SHDSL; however in terms of commercial penetration rate, asymmetrical DSL is higher compared to the symmetrical one, hence this paper pay more focus on asymmetrical DSL.

The ADSL2 standard, a recent version of ADSL, adopts enhanced modulation to reduce noise effect on the signals for higher coding gain and higher rate of the line. The ADSL2 system works at 50 K faster than the ADSL system and transmits signals 200 m farther, amounting to 6% more coverage. The newer version of ADSL is ADSL2. This standard issued in 2003 which referred to ITU-T G.992.5 standard. According to this standard, an ADSL2+ system shall work at up to 24 Mbps or a higher rate on downstream with downstream frequency around 2.2 MHz.

ADSL and ADSL+ are deployed using the existing PSTN infrastructure. The DSLAM node is located in the central office and defined as aggregation nodes. The existing copper cable is used so that the broadband service can be evenly distributed from Local Exchange, Street Cabinet and household's area. The size of access network zone is determined by the maximum copper cable length [5]. At DSLAM side, traffic is multiplexed and transmitted over fiber based transmission to the IP backbone.

An ISP may implement xDSL technology using different approach including DSLAM in the local exchange (fiber to the exchange, FTTE), DSLAM/MSAN in the cabinet (fiber to the cabinet, FTTC) and DSLAM/MSAN in the building/house (fiber to the house/building, FTTH/FTTB). For this study, we focus on FTTE and FTTC deployment where MSANs are located in the central exchange (FTTE) and street cabinet respectively to reach customer residential with the cable length less than 4 km.

2.1 xDSL attainable rate

The maximum attainable rate and allocated bandwidth per ISP plan (Mbps) will affect the femtocell performance. While the bandwidth allocation depends on subscription profile, maximum attainable rate are determined by corresponding xDSL technologies and physical copper cable quality (mainly characterized based on its attenuation and SNR).

In order to discuss about attainable rate and to derive xDSL quality model as femtocell backhaul, TELKOM conducted a study on its copper cable performance in supporting several xDSL technologies including ADSL2, ADSL2+, and VDSL2. Even though the study initially performed to assess IPTV implementation, the study result is relevant to support femtocell implementation.

The study of xDSL performance could not be separated by FTTx technologies implementation, since both are complementary to each other. Performance of transmission technologies over copper have been evaluated for the following reference architectures:

a. FTTE (DSLAM or MSAN in Exchange) which use technology: ADSL2/2+ technology as the last mile access
b. FTTC (MSAN in street cabinet) which uses technology: ADSL2/2+, VDSL2 (profile 8b, use the same Tx level as ADSL2/2+)
c. FTTB (ONU or MSAN at building) which uses technology: VDSL2 (profiles 17a and 30a)

The studies are performed by using simulation under Telecom Italy supervision. The cable models have been defined according to TELKOM's installed cables. Only FTTE and FTTC configuration will be described in this paper. While FTTE can provide internet service for subscriber located 3-5 away from Local Exchange, the FTTC can provide adequate and reasonable performance for residential in order to support IPTV service over xDSL broadband.

Structure of Telkom Indonesia cables can be summarized as following:

- Diameter of conductors: 0.6mm-0.8mm
- Insulation: polyethylene
- Basic structure: quad (2 pairs)

- First level of aggregation:
- For cables up to 100 cp: Unit, consists of 5 quad (10 cp)
- For cables with 200 cp or more: Super Unit of 25 quad (50 cp)

Based on those information, cable model were constructed based on the following parameter:

- Diameter of conductors: 0.6mm (worst case)
- Attenuation: as of CT 1341 Italian cable, polyethylene insulated quad cable
- Reference cable binder:
  - Primary cable: 50 pairs (1 super unit, 25 quads)
  - Secondary cable: 50 pairs (5 units, 5x5 quads)
  - Drop cable: 20 pairs
- Number of boxes per access binder: 3
- Number of drops per building cable: 4

Crosstalk model is based on following assumptions:

- NEXT reference value (@ 1 MHz) is considered the value @ 1% of the estimated statistical distribution
- FEXT reference value (@ 1 MHz and 1 km) is assumed as in standard (Recommendation ITU-T G996.1, ETSI 101 524)
- Crosstalk (statistical value @ 99% of confidence):
  - KNEXT (@ 1 MHz) = -52.2 dB,
  - KFEXT (@ 1 MHz @ 1 km) = -45.0 dB

Noise mix follows the assumption;

- Present broadband systems: 95% ADSL2/2+ (over POTS), 5% SHDSL, no regeneration allowed in access network
- Medium Term Broadband Services penetration assumed for performance estimation: 30%
- Long Term Broadband Services penetration assumed for performance estimation: 50%
- Noise mix composition for scenario (FTTC):
  - Mix 30%BB: 1 SHDSL (@2.3Mbit/s), 14 ADSL2/2+, 35 POTS (or vacant)
  - Mix 50%BB: 1 SHDSL (@2.3Mbit/s), 24 ADSL2/2+ or VDSL2, 25 POTS (or vacant)

xDSL Physical layer setting for performance evaluation follows

- Frequency Plan for VDSL2: 998 Hz
- Physical layer setting for ADSL2/2+ ; Internet access services: NM=6dB, Channel Mode FAST ; IPTV services: NM=9dB, INP=2, Max Delay=8ms
- Physical layer setting for VDSL2, IPTV services: NM=9dB, INP=2, Max Delay=8ms
- ADSL2/2+ performance curves refer to systems implementing extended framing.

Based on the xDSL modeling, the simulation can provide the following results:

- ADSL2/ADSL2+ Performance in FTTE
- ADSL2/ADSL2+/VDSL2 Performance in FTTC

Based on above data and the xDSL modeling, the performance of ADSL2/ADSL2+ and VDSL2 can be seen in Figure 2 and Figure 3.

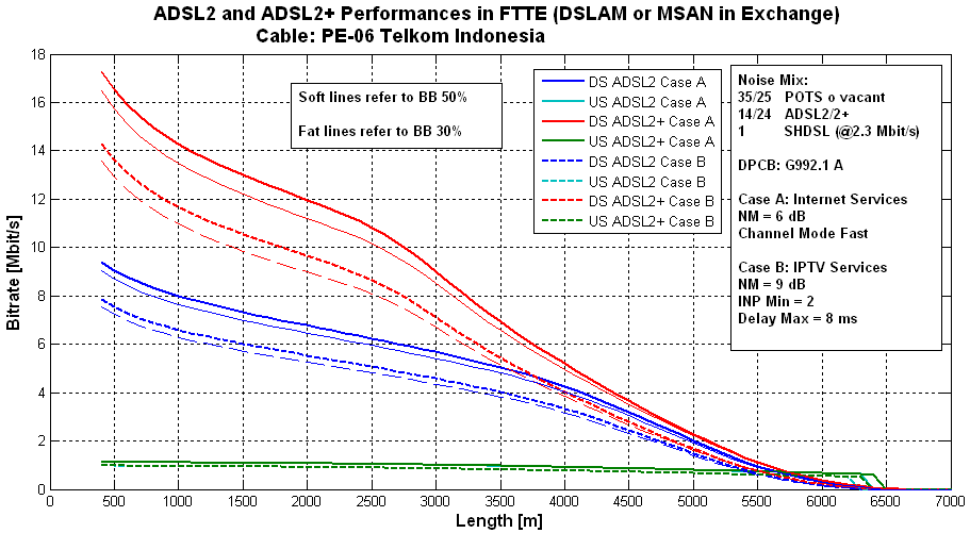


Fig. 2. Performance of ADSL2 and ADSL2+ using FTTE (MSAN at Local Exchange)

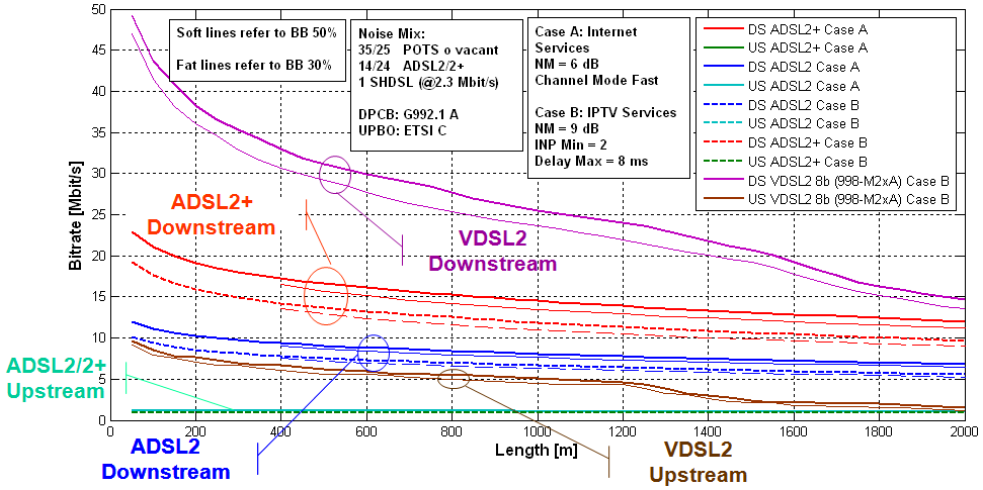


Fig. 3. Performance of ADSL2, ADSL2+ and VDSL in FTTC (MSAN at street cabinet)

In general, attainable data rate depend on xDSL technology. VDSL2 deliver higher data rate since it uses wider frequency plan up to 12 MHz compared to ADSL2+ which uses 2.2 MHz and ADSL2 uses 1.1 MHz. As the cable length increases the total impedance will also increase, it will add more attenuation to the signal. As a result the data rate will decrease as the cable length increase. Higher frequency will experience more sever attenuation as a function of distance, so that the higher data rate can only be maintained in shorter distance.

As can be seen from the graph, the maximum attainable rate also depends on xDSL channel mode. DSL is designed to deliver internet access. By default it uses fast channel

mode, since it will offer a higher efficiency (more data, less error correction redundancy code in each packet). The fast channel mode will allow users to have faster and smaller ping times. However as the real time applications such as IPTV was introduced to the market, interleave channel mode is required. In video applications, there is no time to re-transmit data if errors are detected. In order to limit the impact of long burst errors, an interleaver device is used to spread the data out or shuffles the data after encoded by the Reed-Solomon code [6]. By using Reed Solomon and interleaver as in ADSL and VDSL technology, long error bursts will be equally distributed, so that the errors can be corrected more easily using forward error correction. Since there are bits used for codeword, bits number for data in the interleave mode will be less, hence it will affect the total data rate.

ADSL2/ADSL2+ has a limitation in the upstream bandwidth which is up to 1.1 Mbps, when FAST mode is used or up to 900 kbps when INTERLEAVED is used. However this relatively high upstream bandwidth is only available if the cable length is less than 600 meter from MSAN location. To ensure the stability, 512 kbps bandwidth should be considered for both modes, since the bandwidth is available even when the cable length 6 km away from MSAN location.

For higher upstream bandwidth, VDSL2 should be considered under FTTC configuration. With VDSL2, the upstream data rate can reach around 5 Mbps at the range up to 1 Km.

2.2 xDSL transmission delay

In order to characterize xDSL as femtocell backhaul, we also test transmission delay of ADSL and ADSL2+ from modem to the DSLAM. The experiment has been done in TELKOM R&D Centre Test Bed (called OASIS). Due to the fact that the measurement conducted in the testbed, it is impossible to varying the cable length to have exact transmission delay observation. For the reason that delay is the inverse of frequency carrier, we measured the delay transmission by varying bandwidth profile. By limiting bandwidth profile, the frequency carrier is set and affects the transmission delay. We did not consider delay due to lost in cable as in the commercial DSL deployment.

The network configuration for this observation can be seen in Figure 4.

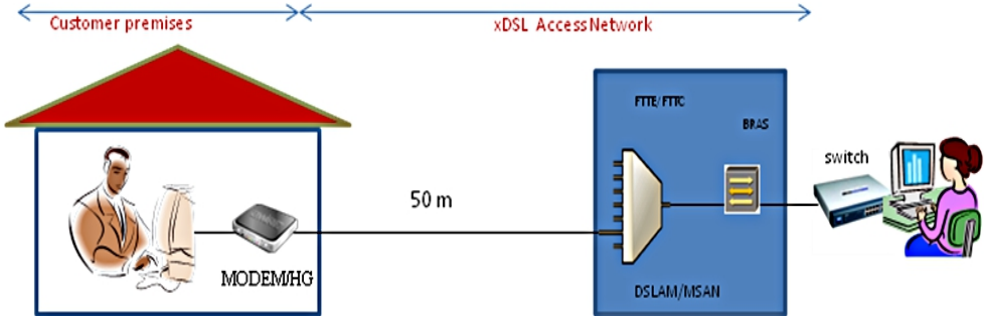


Fig. 4. Transmission Delay Measurement in TELKOM R&D Centre's Testbed

The transmission delay of xDSL over various bandwidth profiles can be seen in Figure 5.

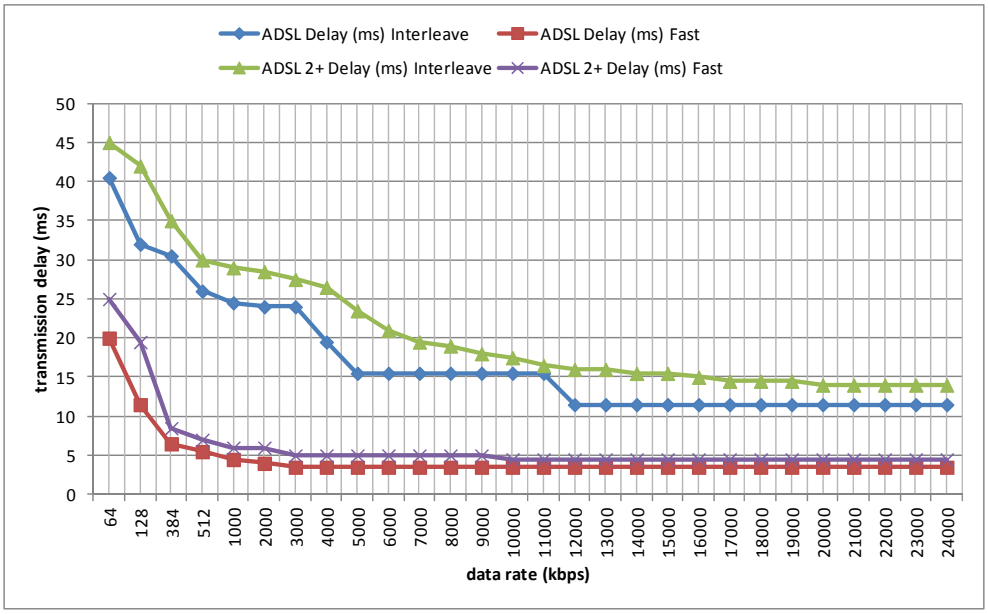


Fig. 5. Average delay transmission of ADSL/ADSL2+ over various bandwidth profile

It can be seen from the graph that the delay at 64 kbps bandwidth is about 40 ms for fast mode and about two times for interleave. As the bandwidth increases the delay decreases since wider frequency bandwidth is required to produce more throughput.

2.3 xDSL quality model

We have discussed the maximum attainable rate which highly depends on DSL technologies, copper quality, relative distance from DSLAM/MSAN to the modem in user side and the channel modes implementation. However in the commercial point of view, the xDSL data rate is further limited by subscription profiles. An ISP usually offers various lines speed to the customer along with the broadband services. The penetration of fix broadband and line speed may vary from country to country depend on the penetration rate and the purchase power for individual line speeds.

We use Europe market as an example. According to Figure 6, as January 2010, there were about two-thirds of fixed broadband lines in the European region offered line speeds between 2 – 10 Mbps. While low speed broadband lines ranging from 144 kps – 2 Mbps represent only 16% of all fixed broadband lines, the penetration of high speed broadband link above 10 Mbps is about 23% of all fixed broadband. In terms of growth, most net fixed broadband additions in 2009 were for high speeds above 10+ Mbps. Most EU countries experienced a reduction in the proportion of low-speed fixed broadband lines.

Indonesia represents a development country. The growth rate of broadband Indonesia projected up to 49%, but the penetration rate is estimated the lowest in Asia. In this broadband business the majority market (67%) is served by TELKOM Group with its

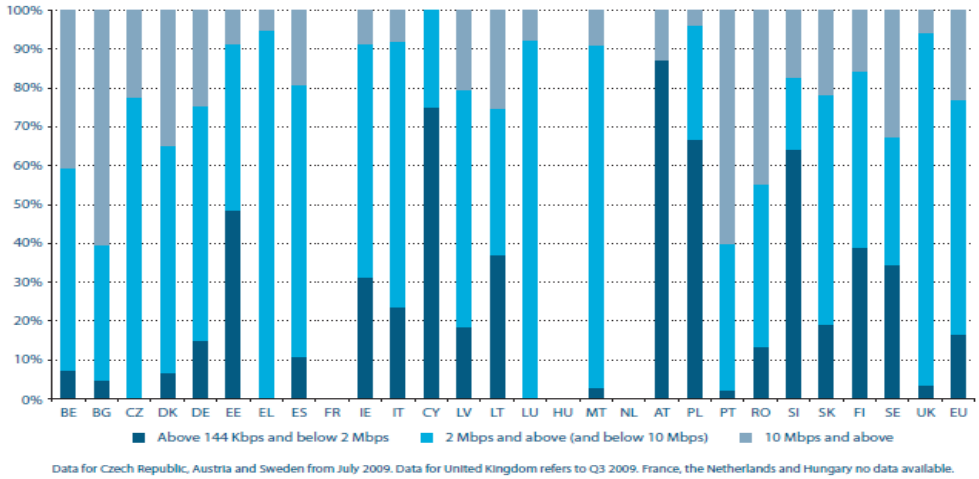


Fig. 6. Fixed Broadband Lines by Technology, January 2010 [7]

Telkom xDSL and Telkomsel mobile broadband. According to the market survey conducted in Jakarta 2011, the fixed broadband line subscription lower than 1 Mbps represent about 72.1 % of population. The line speeds above 3 Mbps represents 9.1% of population and the subscription of 2 Mbps link is about 18.8%.

By considering both technical characteristics and line speed penetration, we propose xDSL backhaul quality model for reference purpose for femtocell deployment. The model can be summarized in Table 1.

Low speed backhaul has relatively high penetration rate especially in Telkom Indonesia which is above 60%. By using fast mode, it will allow to offer high speed internet access up to 4 Mbps. In terms of line speed, this type of backhaul can accommodate almost 100% subscribers with line speed below 3 Mbps. In European region, it will address only 16% subscribers with line speed below 2 Mbps.

Backhaul Types	DL att. rate	UL att. Rate	FTTx/DSL Conf.	Copper Length
Low Speed Backhaul	< 2 Mbps	Up to 512 kbps	FTTE/FTTC with ADSL2 /ADSL2+, fast mode	1 – 4 km
Medium Speed Backhaul	3-10 Mbps	Up to 1 Mbps	FTTC with ADSL2+, fast and interleave mode	≤ 1 km
High Speed Backhaul	11-24 Mbps	2 - 5 Mbps	FTTC with VDSL2, fast and interleave mode	≤ 800 m

Table 1. xDSL Quality Model for Femtocell Deployment

The medium bandwidth category (3 – 10 Mbps in downstream direction) is usually used to accommodate IPTV, internet and hosted home video surveillance services. In the upstream direction, 1 Mbps line speed can be considered when the copper length can be maintained less than 600 meter from MSAN location. By this medium speed backhaul, we assumed more than 50% DSL subscribers in most European region can have this line speed. While in Jakarta, IPTV service is still emerging, however in term of the attainable rate, it will potentially cover 30% DSL penetration and it will further increase as the deployment of FTTC is progressing.

Figure 7 illustrates the femtocell backhaul model based on the Table 1. In distance 4 km, backhaul quality will be limited. Theoritically ADSL2+ in 4 km can deliver 9 Mbps, however according to the real implementation it can only deliver 4-5 Mbps. By shortening the copper length using FTTC configuration where now copper length is maximum 1 km, the backhaul quality is much better.

High speed backhaul (up to 24 Mbps) is offered to the customer who requires more bandwidth in both downstream and upstream direction. Currently this type of backhaul addresses more than 23% of all fixed broadband in most European Region. Even though VDSL2+ can offer more than 30 Mbps data rate, the upstream data rate limits the overall performance. We limit the range of this backhaul type up to 800 meter distance in order to achieve 2-5 Mbps speed in the upstream.

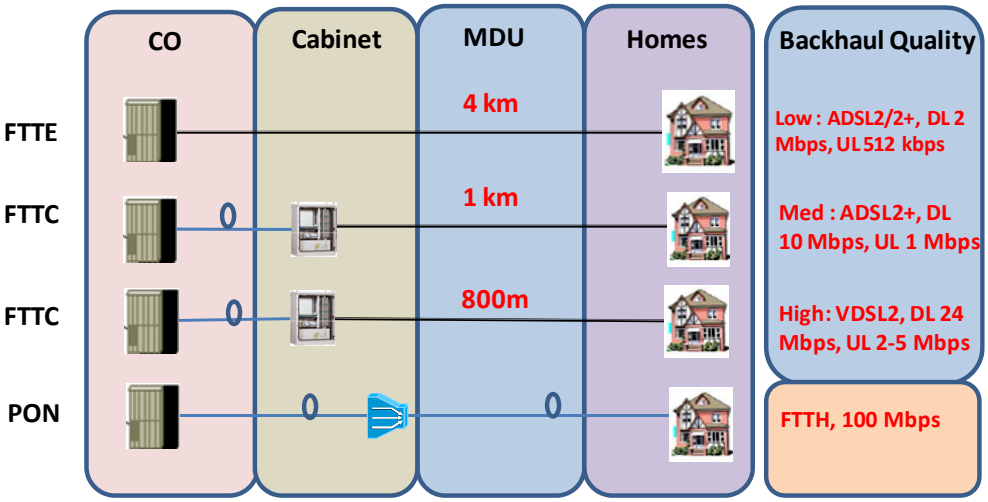


Fig. 7. Backhaul Quality Model

3. Femtocell bandwidth requirement

As been discussed in the previous section, femtocell utilizes the existing broadband IP access available in the customer home. As ADSL2/ADSL2+ has highest penetration rate compared to other fixed wireline access technologies, logically the initial femtocell deployment would have used this type of access. HSDPA Femtocell, for instance, has radio capacity about 7.2 Mbps downlink and 1.4 Mbps uplink. Low speed backhaul, as in Table 1,

cannot completely accommodate the HDSPA full buffer in the air interface, so there may be bottleneck in the xDSL link. However if operator carefully analyze the individual bandwidth and QoS requirement for the basic communication services such as voice, http, ftp and video streaming, one can still hope that even low speed backhaul is able to support femtocell.

During connected mode, the bandwidth required by a FAP depends on service or application. Each service has its own traffic behavior. The aggregate traffic from different users will determine the total traffic occupied by the femtocell. However, in case Indonesian mobile operators, where the market is very competitive, 3G packet data offerings are based on unlimited data schemes. The offering may vary from 64kbps, 128kbps, 384kbps, 1.8Mbps, 3.6 Mbps and 7.2 Mbps. The average bandwidth required by a FAP to support these offerings can be seen in Table 2. The bandwidth is determined in the backhaul interface both for xDSL and Ethernet. As it can be seen from the table, that unlimited data packages of 64kbps, 128kbps from four different UEs can be supported by low speed backhaul category (below 2 Mbps downlink, 512 kbps uplink, according to xDSL case as in Table 1). Offers providing up to 384kbps, 1.8 Mbps, 3.6 Mbps and 7.2 Mbps should consider medium to high bandwidth quality. The constraint will be in uplink streams if the offering is symmetrical between uplink and downlink.

Services	xDSL		Ethernet	
	Downlink	Uplink	Downlink	Uplink
12.2k CS voice	84.8 kbps	84.8 kbps	62.4 kbps	62.4 kbps
64k CS video	212 kbps	212 kbps	163.2 kbps	163.2 kbps
64k data	83.9 kbps	83.9 kbps	74.5 kbps	74.5 kbps
128k data	167.7 kbps	167.7 kbps	149.1 kbps	149.1 kbps
384k data	503.2 kbps	503.2 kbps	447.3 kbps	447.3 kbps
HSDPA 1.8	2.3588 Mbps		2.0967 Mbps	
HSDPA 3.6	4.7176 Mbps		4.1934 Mbps	
HSDPA 7.2	9.4352 Mbps		8.3868 Mbps	
HSUPA 1.4		1.8346 Mbps		1.6308 Mbps

Table 2. Femtocell bandwidth estimation over xDSL and Ethernet

Those connectivity data offerings are indeed valid for macrocell. The MBS has been designed to anticipate peak data rate by utilizing carrier class backhaul. In case of femtocell, the backhaul is depending on broadband connectivity subscription in the customer side. The worse case situation is that, the backhaul bandwidth may be far below supported peak data rate of associate radio technology (HSPA, WiMAX, LTE, etc). Therefore we propose more realistic bandwidth requirement for a FAP by observing the individual bandwidth consumption of selected services. For this purpose, we measured the bandwidth for 4-unit-calls FAP (FAP which supports 4 simultaneous calls) which is commonly used for residential femtocell deployment.

3.1 Mesurement methodology

Figure 8 shows xDSL and femto system installed in TELKOM RDC where ADSL2/ADSL2+ is used. The DSLAM node connected to metro-ethernet located in TELKOM OASIS building located in Bandung. The SecGW and FAP GW are connected to metro-ethernet (ME) backbone to reach SGSN in 3G core network (through Iu-PS interface). Iu-CS connection (FAP GW to MSC in other city, 200km away from OASIS testbed) is also available for CS voice call. The connection between xDSL and Femto System and between Femto System and SGSN are considered as controlled environment. Since GGSN is a commercial network, the network load may affect the femtocell performance, especially during busy hours. Another drawback, it is impossible to put measurement tool (such as end-point) in the GGSN side limiting the impact to commercial network performance. For this reason, the measurement tool or application server as end point is put in the internet cloud. In positive way, the femtocell trial performed in this activity, will reflect the real condition.

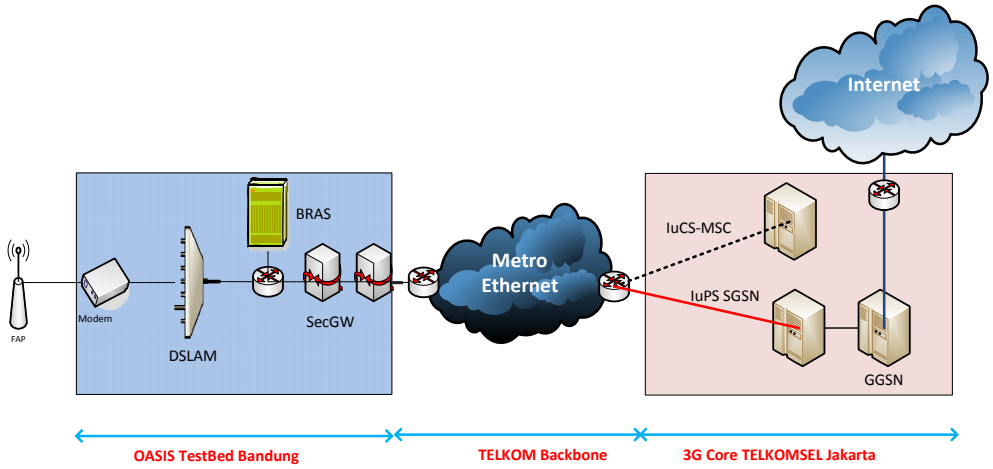


Fig. 8. xDSL as Femtocell Backhaul, A Reference Architecture

In order to derive bandwidth requirements (scenario 1) and femtocell performance (scenario 2), we define mix traffic composition which consists of HTTP, FTP, voice and video streaming. Individual content is defined according to the survey result conducted by TELKOM. There is several internet content accessed by subscribers of two major xDSL providers in Indonesia. As can be seen in Figure 9; facebook.com, detik.com, youtube.com, 4shared.com are among the most popular contents in Indonesia. Among top 10 internet content we choose detik.com or sometime facebook.com to represent HTTP traffic, youtube.com for streaming and 4shared.com to download files from internet. Traffic mix definition being used in the measurement is written in Table 3.

In order to capture bandwidth required by a single FAP, 4 simultaneous UEs are setup to access the FAP. This scenario will be used as a basis to observe minimum xDSL bandwidth requirement and also as a reference of configuring traffic models in STC for observing femtocell performance. Since our focus on backhaul bandwidth, we limit the impact of radio channel fluctuation due to interference and mobility. In this case, FUEs will access the FAPs in such away; the FAPs's signal quality is very good and stable.

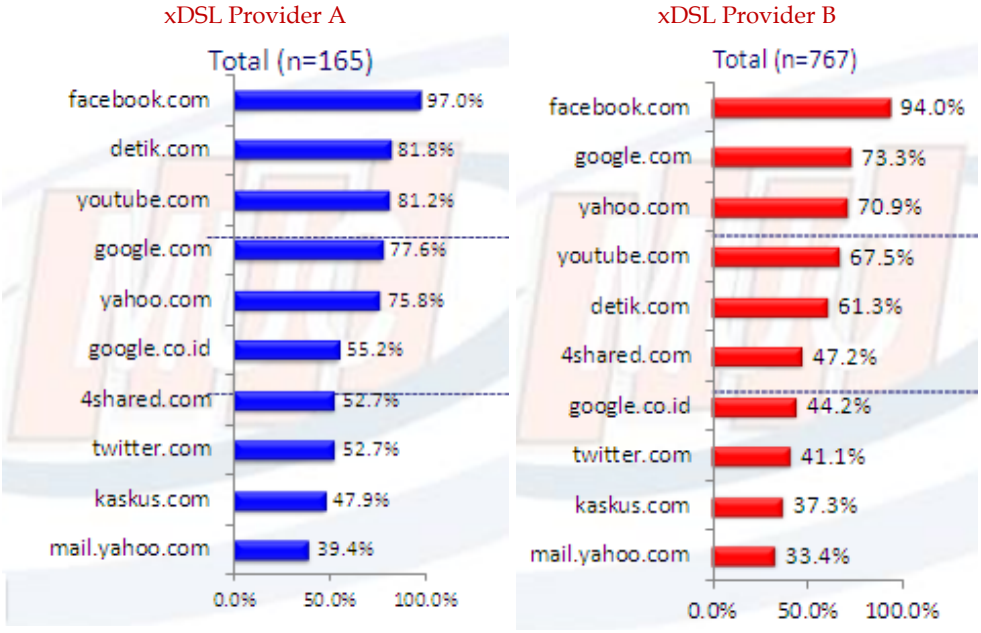


Fig. 9. Top 10 most visited internet content by subscriber of two ISPs in Indonesia

Traffic Mix	Traffic Source	HTTP	File transfer	Video streaming	Voice	Notes
Smartphone mix 4 CS Call	Real	-	-	-	AMR	scenario 1
Smartphone mix all traffic	Real	m.detik.com	4shared.com (5 MB)	youtube 240x360p	AMR	scenario 1
Tablet PC mix all traffic	Real	detik.com & facebook.com	4shared.com (13 MB)	youtube 240x360p	-	scenario 1,2
PC Background	Real	Detik.com	FTP up to 40 MB	youtube	-	Scenario 2
Mix Traffic	Generated and Real	detik.com & facebook.com	4shared.com (13 MB)	Video conference	AMR	scenario 1, 2

Table 3. Summary of traffic mix for femtocell performance using xDSL as the backhaul

The applications are accessed through various UE types including, smartphone and tablet PC. It is assumed that each user accesses only one service type. The traffic traversing through a FAP is recorded by Femto NMS. For this purpose, xDSL profile is set to maximum (20 Mbps), so that the original bandwidth required to send traffic will go smoothly without any congestion or queuing in the access network. Since we use commercial GGSN, the traffic will be affected by the GGSN load, however we anticipated by using 3G SIM Card with the highest priority QoS profile defined in HLR/GGSN. Furthermore the individual test is

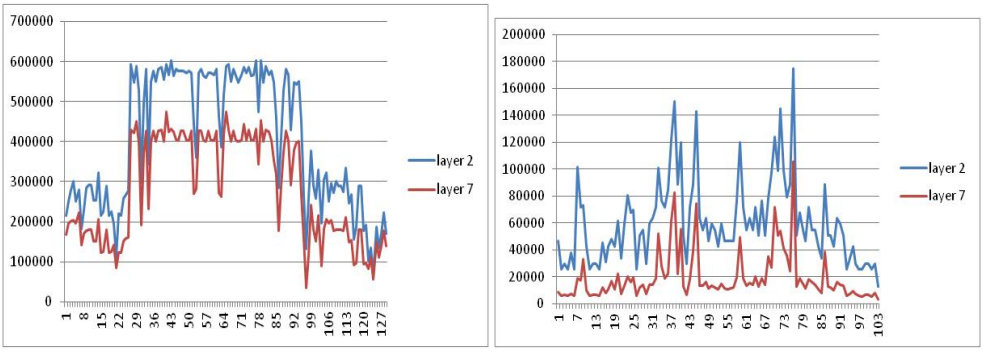
repeated several times, i.e. 30 times effectively, but only the best 3 retries with similar statistical properties will be shown.

The measurement process was divided into two phases. Firstly, we captured individual bandwidth consumed by http, ftp, voice (AMR) and youtube streaming. The service was accessed from a smartphone connected to a FAP. Secondly, we observed the throughput for four simultaneous voice calls and mix traffic (HTTP, FTP, youtube, voice) from four different smartphone at the same time.

We repeated the similar observation for iPad. Since iPad is designed to access packet data only, hence voice AMR call cannot be tested. We replaced the voice call with facebook. We understood that most iPad users frequently download new applications, audio-video podcasts, mp3 music, ebooks from iTunes or Apple Store. For similarity with the observation in smartphone, we used ftp traffic from 4shared.com to download bigger file size compared to the one from smartphone. We also used the same page from detik.com. While from smartphone, mobile page version was displayed; in the iPad, full web page was displayed, hence the consumed bandwidth is different.

3.2 Bandwidth requirement based on measurement result

Bandwidth occupation by Mix traffic from smartphones and iPads can be seen in Figure 10 and Figure 11 respectively.

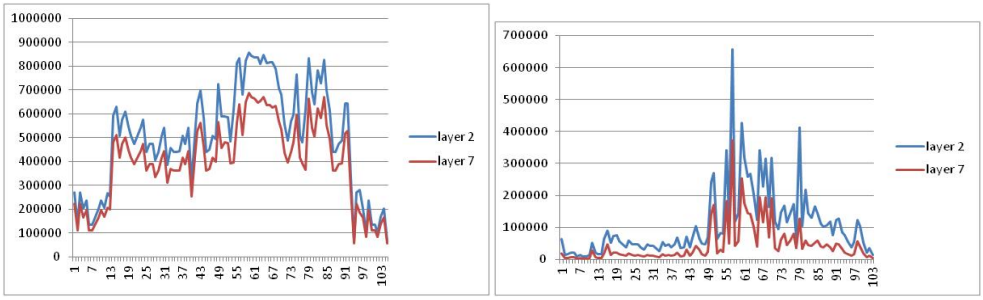


a) downlink b) uplink

Fig. 10. Femtocell mix traffic bandwidth from 4 smartphones, measured in xDSL (layer 2) and application (layer 7)

The summary of statistical properties for individual traffic and mixed traffic accessed by both smartphone and iPad are shown in Table 4 and Table 5 respectively.

Since both DL and UL traffic follow lognormal distribution, maximum value is used so that all traffic can traverse smoothly. According to Table 4, we can see that the bandwidth required for a femtocell depends on the type of traffics. In downlink side it requires about 602 kbps to perfectly handle mix traffic from 4 smartphones while in uplink is about 175 kbps. The uplink traffic in smartphone contains voice AMR (12.2kbps) so ideally the bandwidth should be preserved above 84.8 kbps.



a) downlink b) uplink

Fig. 11. Femocell mix traffic bandwidth from 4 iPads, measured in xDSL (layer 2) and application (layer 7)

In case of iPad being used as shown in Table 5, the downlink bandwidth consumption is about 857 kbps and the uplink is about 632 kbps. The uplink traffic for iPad is higher than smartphone because mix traffic mainly used by youtube. However as can be seen from Figure 11 peak throughput of 632 kbps is occurred only small portion compared to overall throughput shape, the rest of throughput is below 400 kbps.

Traffic	Content	Strm	Max (bps)	Min (bps)	Average (bps)	Variance	Distribution
Cs voice	AMR 12,2 Kbps	DL	84,800	8,480	77,372.7	466701399	Lognormal
		UL	84,800	8,480	77,372.7	466701399	Lognormal
HTTP	m.detik.com, first page	DL	224,720	4,240	42,995.5	3116322837	Lognormal
		UL	106,000	0	29,400.4	886953056	Lognormal
FTP	www.4shared.com "05.The Lazy Song", 5MB	DL	576,640	4,240	185,196.8	4640833694	Normal
		UL	42,400	0	16,178.7	43352265	Lognormal
Streaming	m.youtube.com, "If you sleep in at my house, you are "Doom"ed", 41s, 240p, 380 Kbps	DL	339,200	29,680	278,521.7	7042225332	Normal
		UL	71,016	8,480	27,326.68	148142570	Lognormal
Mix	4 CS call	DL	339,200	8,480	298,168.5	8044648128	Lognormal
		UL	339,200	8,480	298,168.5	8044648128	Lognormal
Mix	All traffic	DL	602,080	67,840	400,326.4	19280386460	Lognormal
		UL	174,896	12,720	60,522.33	768972191	Lognormal

Table 4. Statistical properties of individual and mix traffic in smartphone case

So far we have derived the bandwidth requirement for femtocell by monitoring the throughput from Femtocell NMS.

- BR for smartphone = 607 kbps (DL) and 175 kbps (UL)
- BR for iPad = 857 kbps (DL) and 400 kbps (UL)

Traffic	Content	Strm	Max (bps)	Min (bps)	Average (bps)	Variance	Distribution
HTTP	www.detik.com, first page	DL	1,094,976	12,720	282,005.8	2,24269E+11	Normal
		UL	168,536	0	94,768.0	2549473195	Normal
HTTP	www.facebook.com, home page	DL	561,800	0	109,618.1	18271227072	Lognormal
		UL	210,936	0	45,605.9	3675590487	Lognormal
FTP	www.4shared.com "divxim.net-kLite Codec Pack 4.9.5 FULL", 13,6MB	DL	407,040	4,240	213,622.6	3926226329	Normal
		UL	59,360	0	22,047.5	65333480	Lognormal
Streaming	m.youtube.com, "If you sleep in at my house, you are "Doom"ed", 41s, 240p, 380 Kbps	DL	339,200	4,240	280,032.6	4284588520	Normal
		UL	48,760	12,720	27,491.1	61155798	Lognormal
Mix	All traffic	DL	857,536	67,840	501,702	10293926961	Normal
		UL	658,256	8,480	106,523,7	13009342028	Lognormal

Table 5. Statistical properties of individual and mix traffic in iPad case

4. Femtocell performance

In this section, the performance of a femtocell service is observed in the presence of background traffic in xDSL modem. This observation will effectively address the nature of FAP which is customer premises equipment. It is most likely that the user will plug the FAP to xDSL modem or home gateway on top of the existing broadband access in the home. Without prior notice, the femtocell service will be mixed with traffic from PC or other devices connected to the same modem.

4.1 Measurement methodology

Scenario 2 was defined to verify femtocell performance in the existence of background traffic in xDSL modem. This scenario is designed to show that if MNO and xDSL provider does not sign an agreement, the internet traffic and FAP traffic will be mixed into a single PVC (Physical Virtual Connection), so that regardless the QoS setting in the modem, both traffic will have a same priority and compete each other as best effort.

The femtocell and PC are connected to xDSL modem using single PVC (Physical Virtual Connection) so that the traffic will mix each other. We set the DSLAM to interleave mode, while the modem uses default UBR (Universal Bit Rate) type of service. The PC generated mixed traffic HTTP (www.detik.com), FTP (rapidshared 40 Mbps) and youtube. The background will use real DSL traffic generated from the measurement tools. We use the same network reference architecture as in Figure 8, except that in the modem/home gateway there is a PC and a femtocell connected to the modem. The PC generates traffic mix (called PC background) as defined in Table 3.

Femtocell is attached to serve 4 FUEs simultaneously. We are referring to iPad case to inline with the maximum BR obtained from previous observation. By using iPad case which has

higher bandwidth consumption, the bandwidth requires for smartphone case logically will also be supported. For femtocell we use generated traffix mix from Spirent Test Centre (STC) and Cisco Telepresence Service. This is important since we need a reference service performance to be monitored in the presence of other traffic from other FUEs as well as the PC. In this case video conference is used since it has several performance metrics including throughput, jitter and packet loss for both video and audio quality. In order to approach femtocell BR obtained from bandwidth requirement observation which uses real traffic, we generated mix traffic from STC is made as close as possible to the iPad mix traffic. The comparison between generated traffic and real iPad mixed traffic can be seen in Figure 12.

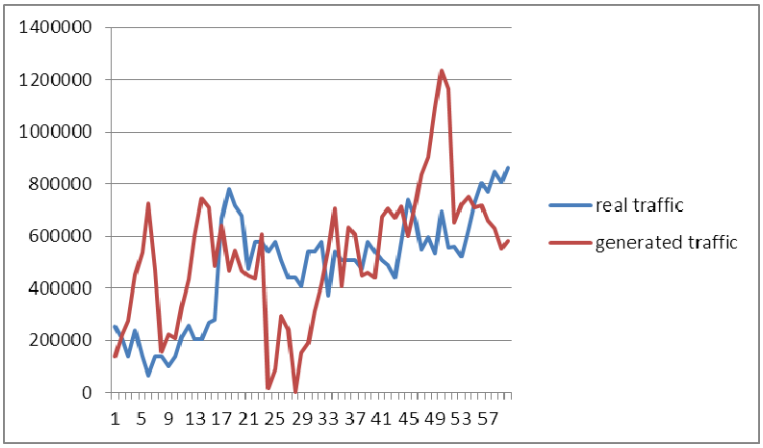


Fig. 12. The comparison between generated traffic and iPad mix traffic

The test scenario is divided into two sub scenarios. The first sub-scenario is to obtain the femtocell performance without background traffic. While in the bandwidth requirement measurement (scenario 1) we set xDSL profile only to 20 Mbps; in this sub-scenario we set the line bandwidth to 20 Mbps, 1 Mbps, and 800 kbps. In each line profile we observed the video-audio performance. It will give a performance reference as well as verification to the BR obtained from scenario 1.

In the second sub-scenario we activate the traffic from PC. We then analized the performance of video conference (packet loss, jitter) with the existence of background traffic. In this case we used reference performance of xDSL set to 1 Mbps without background and compare the new video performance in the presence of background from PC. We increased the bandwidth profile step by step until the video performance below the threshold (packet loss < 3%, jitter < 40ms).

4.2 Performance degradation due to background traffic in access

Table 6 shows the video conference performance without PC background. When the xDSL bandwidth profile set to 800 Kbps or equal to BR for iPad, the video quality for iPad is maintained below the threshold; the packet loss below 3% and the jitter far below than 40 ms. By giving 20% more bandwidth to 1 Mbps it will give additional space to anticipate other header and load during busy hours and also it will give better performance for real time

traffic. In this case, it gives less jitter and increase throughput. TELKOM usually adds 20% on top of bandwidth requirement to engineer xDSl line speed, hence 1 Mbps line speed is used. It can be shown from Table 6 that the performance metrics of 1 Mbps line is closed to 20 Mbps.

Mode	Average Video Quality for 5 minutes observation					
	Bitrate (kbps)		Packet loss (%)		Jitter (ms)	
	Tx	Rx	Tx	Rx	Tx	Rx
20 Mbps	83.58	23.53	0.53	2.45	13.30	19.53
1 Mbps	78.93	22.93	0.34	2.47	12.23	20.75
800 Kbps	84.70	20.26	0.37	2.12	13.45	24.96
Mode	Average Audio Quality for 5 minutes observation					
	Bitrate (kbps)		Packet loss (%)		Jitter (ms)	
	Transmit	Receive	Transmit	Receive	Transmit	Receive
20 Mbps	49.33	25.38	0.14	0.47	7.38	3,3
1 Mbps	49.33	25.38	0.14	0.47	7.38	3,3
800 Kbps	54.81	27.21	0.16	1.25	6.68	9.9

Table 6. H264 Video Conference Performance without PC background

We use video conference as reference in order to see the effect of background traffic from PC to the femto service performance, in this case video conferencing. It can be seen from Table 7, as

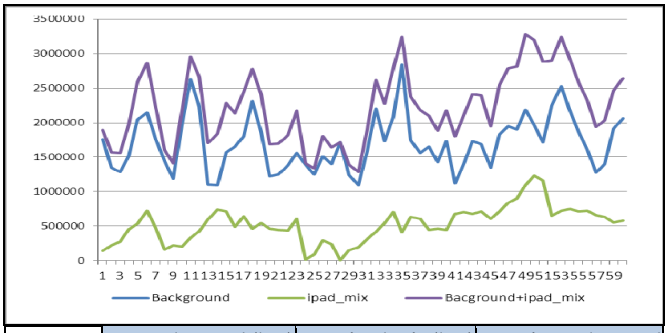
Mode	Average Video Quality					
	Bitrate (kbps)		Packet loss (%)		Jitter (ms)	
	Tx	Rx	Tx	Rx	Tx	Rx
20 Mbps	81.83	24.97	1.25	2.47	13.92	18.05
2 Mbps	84.25	23.73	0.29	2.99	13.62	35.22
1,5 Mbps	85.28	26.48	0.38	4.43	15.70	29.95
1,2 Mbps	82.26	23.19	0.33	5.94	14.64	41.11
1 Mbps	89.07	27.85	0.12	3.96	13.87	45.52
Mode	Average Audio Quality					
	Bitrate (kbps)		Packet loss (%)		Jitter (ms)	
	Transmit	Receive	Transmit	Receive	Transmit	Receive
20 Mbps	56.67	58.92	0.82	0.66	7.02	3.95
2 Mbps	52.00	25.33	0.18	0.95	7.35	14.08
1,5 Mbps	27.43	25.02	0.07	1.22	8.45	16.88
1,2 Mbps	47.08	26.92	0.13	1.67	7.00	16.85
1 Mbps	28.57	25.45	0.05	1.32	10.20	17.57

Table 7. H264 Video Conference Performance in the presence of background traffic from PC

soon the background traffic exist, the packet loss of video quality increase from 2.47% to 3.98% and the jitter double from 20.75 ms to 45.52 ms. In order to improve the video performance we increase the xDSL bandwidth profile from 1 Mbps to 1.2 Mbps, 1.5 Mbps and 2 Mbps. When it set to 2 Mbps, the packet loss is below 3%, and the jitter is also below

40 ms. In this case, when the femto and PC using the same PVC, ones can improve the performance of femtocell by subscribing more xDSL bandwidth about double than the required bandwidth by a FAP. Of course the final bandwidth is sensitive to the amount of internet traffic from local area network (PCs).

Figure 13 shows the mix traffic from femtocell and PC when BW profile set to 20 Mbps. Even though the average total traffic from both femtocell and PC is about 2.2 Mbps, setting up line speed to 2 Mbps has given acceptable performance to the real time femtocell traffic.



	Background (bps)	iPad Mix (bps)	Mix+Background
max	2,839,920	1,234,368	3,280,096
min	1,093,248	4,240	1,285,244
average	1,703,366.8	526,393.8	2,229,760.6
variance	158653351055	66753348160	280358769397

Fig. 13. Aggregate Throughput in xDSL link, BW profile set to 20 Mbps

5. Conclusion

In this chapter the xDSL characteristics has been explained including xDSL attainable rate and transmission delay. Based on TELKOM study result, the performance of ADSL2, ADSL2+ and VDSL2 under FTTE, FTTC and FTTB configuration have been discussed. The backhaul quality model for xDSL has been proposed in order to accommodate both technical limitation of xDSL (in example cable length and constaint in xDSL uplink bandwidth) and current penetration rate of bandwidth subscription profiles both in Indonesia and Europe. DSL backhaul quality model is derived in order to address different qualities of backhaul. The model has been used in FREEDOM project in elaboration of cooperative RRM, scheduling and system level simulation which need to take into account the backhaul quality.

According to the performance measurement result, it can be concluded that the femtocell performance can be affected by internet traffic in the xDSL modem. This case mostly happens when the users instantly plug in the femtocell to their broadband connection without knowing that the traffic from a FAP and PC could compete to each other without any priority or separation. This is also true if the xDSL service provider is a separate company, and there is no service level agreement with the femtocell service provider in order to maintain end-to-end QoS.

In order to achieve better performance for femtocell, the customer should have additional bandwidth to accommodate both traffic. Alternatively an integrated modem-femtocell

solution (home-gateway with femtocell capability) can be introduced in order to have joint-scheduling between xDSL modem and femtocell.

Implementing backhaul aware scheduling (BAS) in the femtocell can be another alternative in order to minimize the impact of bottleneck in the backhaul to the femtocell service performance [4]. This study was driven by the limitation of backhaul capacity (e.g., bottleneck or congestion) caused by other traffic (e.g., IPTV or Internet access in xDSL modem) which affects the performance of FAP in serving requested traffic from femtocell users. The admission control is incorporated with the scheduling method to treat all kinds of traffic served by the FAP. With BAS, the FAP can decide whether the backhaul capacity is enough or not to support existing session. The simulation results show that with BAS, the performance of FAP can be improved especially for peak backhaul conditions compared to FAP without BAS.

6. Acknowledgment

This work has been performed in the framework of the FP7 project FREEDOM IST-248891 STP, which is partially funded by the European Community. The Authors would like to acknowledge the contributions of colleagues from FREEDOM Consortium [8], TELKOMSEL, HUAWEI and CISCO who supported the measurement campaign and gave technical assistant.

7. References

[1] Patrick Donegan, "Backhaul Strategies for Mobile Carriers", Heavy Reading Report Vol. 4, No. 4, March 2006

[2] Epitiro Technologies, Ltd, Femtocell Deployment Guide: An Operator-focused Strategy for a Successful Femtocell Rollout, 2008

[3] Simon R. Saunders, Stuart Carlaw, Andrea Giustina, Ravi Raj Bhat, V. Srinivasa Rao and Rasa Siegberg."*Femtocell Opportunities and Challenges for Business and Technology*", 2009 John Wiley & Sons Ltd.

[4] Hariyanto, Hadi; Noviyanti, Karina W.; Widiawan, A K.; Kurniawan, Adit; Hendrawan; "Backhaul-aware scheduling for WiMAX femtocell with limited backhaul capacity", TENCON 2011 - 2011 IEEE Region 10 Conference, Digital Object Identifier: 10.1109/TENCON.2011.6129013, Page(s): 1280 – 1284, 2011.

[5] H. M. Sigurdsson,K. E. Skouby, "Techno-economic evaluation of broadband access technologies: The BREAD approach", Center for Information and Communication Technologies, CICT Technical University of Denmark

[6] P. Golden, H. Dedieu, K. Jacobsen, "Fundamental of xDSL Technologies", Auerbach Publications, Taylor & Francis Group, New York, 2006.

[7] European Commission, "Europe's Digital Competitiveness Report", Publication Office of the European Union, Luxembourg, 2010

[8] FP7-ICT-2009-4, FREEDOM "Femtocell-based netwoRk Enhancement by intErference management and coorDination of infOrmation for seaMless connectivity", http://www.ict-freedom.eu/.

8

Sum-of-Sinusoids-Based Fading Channel Models with Rician K-factor and Vehicle Speed Ratio in Vehicular Ad Hoc Networks

Yuhao Wang and Xing Xing
School of Information Engineering, Nanchang University
China

1. Introduction

In this chapter, we propose an extended reference model and two novel sum-of-sinusoids (SoS) models (statistical and deterministic simulation models) propagation models considering the Rician K-factor and vehicle speed ratio in Vehicular Ad Hoc Networks (VANETs). Our models consider comprehensive scene of wave propagation, including I2V (infrastructure-to-vehicle) channels with a LOS or NLOS environment, IVC (inter-vehicle communication) channels with a LOS or NLOS environment. The analysis of the statistical properties of the proposed models show that the statistics of the new models match those of the reference model at a large range of normalized time delays. The proposed models show improved approximations to the desired auto-correlation and faster convergence with the increase of Rician K-factor and vehicle speed ratio.

2. Background

VANETs have been envisioned for safety, traffic management, and commercial applications, such as information notices of hazardous road conditions, traffic congestion, or sudden stops, furthermore, commercial services (e.g., data exchange, infotainment, rear-seat multiplayer games) (Boban et al., 2009)-(Abbas et al., 2010). Recently, the Quality of Service (QoS) and network security are used to determine the feasibility of such applications. To achieve the desirable QoS and network security, many techniques focusing on the design and optimization of routing protocol in VANETs have been proposed in (Lwinmuller et al., 2006)-(Saleet et al., 2010). However, the fundamental issue centers on the accurate description of the characteristics and mechanism of wave propagation in VANETs, so it is essential to establish a reasonable radio propagation model for VANET channels, which help to understand the characteristics of the channel and take some effective steps to improve the QoS and network security.

Direct communication between vehicles in VANETs may be supported by the deployment of Mobile Ad Hoc Networks (MANETs), which does not rely on fixed infrastructure and can accommodate a constantly evolving network topology (Boban et al., 2009; Vahid, 2009). More recently, infrastructure-to-vehicle (I2V) and inter-vehicle communication (IVC) links are being evaluated for a variety of applications. I2V channels is similar to the traditional cellular systems, the base station is stationary, and only the mobile terminal are in motion. However,

for IVC channels, a variety of applications, such as intelligent transportation systems and ad hoc networks, are based on mobile-to-mobile communications. Both the base station and mobile terminal are all in motion and the transmitted and received signals are all affected by the surrounded scatterers. So channel modeling in VANETs should be considered the both characteristics in I2V and IVC channels. More recently, infrastructure-to-vehicle (I2V) and inter-vehicle communication (IVC) links are being evaluated for VANETs and a LOS or NLOS environment should also be considered. The I2V and IVC channels can be distinguished by vehicle speed ratio and the difference of LOS and NLOS environment can also be represented by Rician K factor. Therefore, a comprehensive channel is needed to wholly describe the scene of wave propagation for VANETs.

An important factor of a vehicular channel model is the mobility (Gowrishankar et al., 2007; Yoon & Noble, 2006) by including the mobility of nodes and the channel variability. Channel variability, is not well modeled in today's wireless vehicle networks. (Pawlikowski et al., 2002) reports that simplistic models may not be practical and it is also different to draw conclusions on the real performance of the upper layers. Designers require statistical models that can accurately capture the characteristic of propagation behavior observed at both mobile vehicles (Michelson & Chuang, 2006).

Currently, free space and two ray ground channel models are the most popular propagation models for simulation in vehicular wireless networks (NS, 2000). For the free space channel model, it describes an ideal propagation characteristic, and the received power depends on the transmitted power, the gain of antenna, and the distance of transmitter-receiver. While for the two way ground model, it assumes that the received signal is the sum of the direct line of sight path and the reflected path from the ground. However, the model does not take obstacles into consideration. It is also too ideal for short transmitter-receiver separation distances, as it assumes that signals have a perfect 250m radius range. On the other hand, QualNet supports open-space propagation as well as stochastic propagation models such as Rayleigh, Rician and Log-normal fading, in which all models describe the time-correlation of the received signal power. Rayleigh model considers indirect paths between the transmitter and the receiver, while Rician model considers when there is one dominant path and multiple indirect signals. OPNet supports open-space propagation models as well as an enhanced open-space model that accounts for hills, foliage and atmospheric affects(OPNET, 2000). Furthermore, obstacle effects are combined in (Jardosh et al., 2006; Jradosh et al., 2005; Mahajan et al., 2007), but the propagation characteristic is limited to line-of-sight. (Stepanoy & Rothermel, 2008) applies a radio planning tool and validates the evaluation for the impact of a more realistic propagation by a set of measurements.

In a dense urban area, path loss, shadow fading and short-term variants are the main factors affecting the communication quality. Path loss and shadowing fading determine the effective communication distance between two adjacent vehicles, while multi-path and Doppler spectrum caused by the sum of absolute speeds of individual nodes affect the quality of service (QoS) in inter-vehicle networks. However, it is noted that some of these effects can be avoided, such as by increasing the height of the antenna, and the inerratic variations is just relative to the distance between transmitter and receiver. Here, the model is focused on the short-term variants, especially for Doppler spectrum model caused by both high mobile vehicles. The Doppler spectrum model in (Clarke, 1968; Gans, 1972) for wireless cellular network cannot really be used for link between double mobile nodes. Akki and Haberp(Akki & Haber, 1989) consider a Doppler spectrum model for radio link between

double mobile nodes in a two-dimensional (2-D) uniform scattering environment. It is a deterministic channel model without considering the specific characteristics, such as the effect of antenna and dynamic distribution of received multi-path wave.

Motivated by the recently presented IVC fading channel models in (Patel et al., 2005; Zajic & Stuber, 2006; Wang et al., 2009), we propose novel statistical and deterministic SoS models for Rician channels in VANETs. As described in (Patel et al., 2005), the properties, like auto-correlation and cross-correlation of the statistical models, differ for each simulation trial, but converge in a statistical sense to the desired properties over an infinite number of simulations. Therefore, such models are termed statistical models. In contrast, the properties of deterministic model, are identical for all simulation trials, hence, they can be predetermined, such models are called deterministic models. We provide detailed simulation results to verify and compare the performance of the proposed models in next sections.

The remainder of this chapter is organized as follows. Section 3 discusses the related work reported in the literature. In Section 4, we extend Akki and Haber's mathematical reference model for IVC channels by introducing the line-of-sight (LOS) components and derive the statistical properties of the extended model. Section 5 establishes new statistical and deterministic SoS simulation models. Their statistical properties are also derived and verified in this section. In Section 6, performance analysis is carried out through comparisons between the reference and the two SoS models. At last, the conclusion remarks are given in Section 7.

3. Related research work

A number of techniques have been proposed for the modeling and simulation of I2V channels. Among them, Clarke (Clarke, 1968) proposed the statistical theory of mobile-radio reception, and a power-spectral theory of propagation in the mobile-radio was developed by Gans in (Gans, 1972). The Jakes' simulator (Jakes, 1994; Dent et al., 1993), which is a simplified simulation model of Clarke's model (Clarke, 1968), has been widely used for frequency nonselective Rayleigh fading channels. Various modified models (Patzold et al., 1998)-(Li & Huang, 2002) and improvements (Xiao & Zheng, 2002)-(Zheng & Xiao, 2003) of Jakes' simulator for generating multiple uncorrelated fading waveforms needed for modeling frequency selective fading channels and multiple-input multiple-output (MIMO) channels have been reported. It is commonly perceived that Jakes' simulator (and its modifications) is more computationally efficient than Clarke's model since Jakes' simulator needs only one fourth the number of low-frequency oscillators as needed in Clarke's model. However, recently Pop and Beaulieu (Pop & Beaulieu, 2001) put forward a view that Jakes' simulator and its variants are not wide sense stationary (WSS), and that the reduction of simulator oscillators based on azimuthal symmetries lacks sufficient basis (Xiao et al., 2006). They improved the simulator by introducing random phase shifts in the low-frequency oscillators to remove the stationary problem in (Pop & Beaulieu, 2001). But Xiao and Zheng (Zheng & Xiao, 2003) gave a statistical analysis of Clarke's model with a finite number of sinusoids and showed that the Pop-Beaulieu simulator has also deficiencies in some of its higher-order statistics. it was further proved in (Xiao et al., 2002) that second-order statistics of the quadrature components and the envelope do not match the desired ones. Moreover, even in the limit as the number of sinusoids approaches infinity, the auto-correlations and cross-correlations of the quadrature components, and the auto-correlation of the squared envelope of the improved simulator, fail to match the desired correlations. Also, Jakes's

original simulator and published modified versions, have similar problems with these second-order statistics. In (Xiao et al., 2006), Xiao and Zheng proposed a statistical SoS model for I2V channels which employs a zero-mean stochastic sinusoid as the specular LOS component, in contrast to previous Rician fading simulators that utilize a non-zero deterministic specular component. The statistical properties of the new simulators are confirmed by extensive simulation results, showing good agreement with theoretical analysis in all cases.

Channel modeling in VANETs should be considered the both characteristics in I2V and IVC channels. Those I2V channel models may not fully reflect the mobility characteristics of VANET channels. Several works in the open literature have been studied in this area (Akki & Haber, 1989)-(Patel et al., 2003). The theoretical analysis of the IVC channels for urban and suburban land communication channels was first developed by Akki and Haber (Akki & Haber, 1989; Akki, 1994), and was extended by Vatalaro and Forcella in (Vatalaro & Forcella, 1997) to account for scattering in three dimensions (3-D), and by Linnartz and Fiesta in (Linnartz & Fiesta, 1996) to include LOS scenarios. Some channel measurement results for narrowband IVC communications have been presented in (Kovacs et al., 2002; Maurer et al., 2002; Cheng et al., 2007). R.Wang and D.Cox (Wang & Cox, 2002) introduced the discrete line spectrum method to simulate the IVC channels. Whereas the accuracy of this method was assured only for short-duration waveforms, Moreover, the numerical integrations required in determining the discrete set of frequencies and corresponding Doppler spectrum make the implementation complex and not easily reconfigurable for different Doppler frequencies or the Doppler frequency ratio. So it is not always suitable for real time hardware channel emulation or software simulation. A method based on inverse fast Fourier transform (IFFT) was presented by D.J.Young and N.C.Beaulieu (Young & Beaulieu, 2000). This method was more accurate and efficient than the method of discrete line spectrum, but the IFFT-based method requires a complex elliptic integration. The authors in (Patel et al., 2003) proposed a "double-ring" scattering model to simulate the IVC scattering environment and developed modifications of two SoS models (statistical and deterministic SoS models) often used to simulate I2V channels in (Patel et al., 2005). More recently, Wang and Liu (Wang et al., 2009) presented a scattering Rician IVC fading model with a LOS component by SoS method, which is based on the Rayleigh model proposed in (Patel et al., 2005). A new statistical SoS in (Zajic & Stuber, 2006) model is proposed for Rayleigh IVC fading channel to directly generate multiple uncorrelated complex envelope, which shows faster convergence than the model in (Patel et al., 2005) and adequate statistics with small simulation trials.

The statistical properties of Xiao and Zheng's simulators in (Xiao et al., 2006) are confirmed by extensive simulation results, showing good agreement with theoretical analysis in all cases and is a typical model with high quality for I2V channels. But with the development of mobile ad hoc networks, VANET channel modeling often involves the IVC channels, which is generally considered as the common case of the I2V channels. Therefore, in this chapter, we mainly focus on the modeling for IVC channels in VANETs. This motivates us to extend the new statistical SoS model in (Zajic & Stuber, 2006) by employ a LOS component to characterize the IVC channels of VANETs. Furthermore, the deterministic SoS model, proposed in (Patel et al., 2005), are employed to simulate Rayleigh IVC channel for its reduced-complexity and theoretical and simulation results verified the usefulness of the model. Seeking to find a more suitable Rician simulation model for VANET channels, we also introduce a LOS component to extend the deterministic SoS model for comparison.

4. The mathematical reference model

Akki and Haber's simulation model for IVC channels with no LOS component gives the complex channel envelope as (Akki & Haber, 1989)

$$Y(t) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \exp\{j[(2\pi f_1 \cos(\alpha_n)t + 2\pi f_2 \cos(\beta_n)t + \theta_n)]\} \tag{1}$$

where f_1 and f_2 are the maximum Doppler frequencies due to the motion of the transmitter and the receiver, respectively. N is the number of propagation paths, α_n and β_n are the random angle of departure (AOD) and the angle of arrival (AOA) of the nth path measured with respect to the transmitter and the receiver velocity vectors, respectively, and θ_n is the random phase uniformly distributed on $[-\pi, \pi)$, independent of $\alpha_n's$ and $\beta_n's$ for all n.

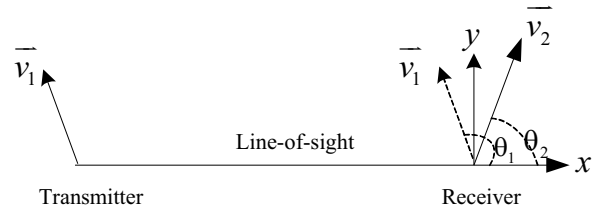


Fig. 1. IVC channel with a LOS component in VANETs

Definitions:

- In Fig. 1, the velocities of transmitter and receiver are $\mathbf{v}_1$,$\mathbf{v}_2$, λ is the carrier wavelength, $a = |\mathbf{v}_1|/|\mathbf{v}_2|$.
- θ_1,θ_2 are the angle between $\mathbf{v}_1$,$\mathbf{v}_2$ and the LOS component, respectively.
- $\mathbf{v}_x$ and $\mathbf{v}_y$ are the relative velocity of receiver to the transmitter in the x-axis and y-axis direction, respectively.
- the Doppler frequency caused by $\mathbf{v}_x$ and $\mathbf{v}_y$ are f_x, f_y.

The angle between $\mathbf{v}_x$ and LOS component is 0° and the direction of $\mathbf{v}_y$ is perpendicular to the LOS component. From (Gregory, 2003), the Doppler frequency caused by LOS component in the IVC environment is $|f_x| = (|\mathbf{v}_2|\cos\theta_2 - |\mathbf{v}_1|\cos\theta_1)/\lambda$. The LOS component is given by

$$L = \sqrt{K}\exp[j\{2\pi(|\mathbf{v}_2|\sin\theta_2 - |\mathbf{v}_1|\sin\theta_1)\}t + \phi_0] \tag{2}$$

where K is the ratio of the specular power to the scattering power, and the initial phase ϕ_0 is uniformly distributed over $[-\pi, \pi)$.

With reference to (1) and (2), the received complex envelope of the IVC fading channel with a LOS component can be expressed as

$$Z(t) = \frac{Y(t) + \sqrt{K}\exp(j2\pi f_0 t + \phi_0)}{\sqrt{1+K}} \tag{3}$$

where $f_0 = (|\mathbf{v}_2|\cos\theta_2 - |\mathbf{v}_1|\cos\theta_1)/\lambda$.

Assuming omnidirectional antennas and isotropic scattering conditions around the transmitter and the receiver, the statistical properties of the reference model are given as follows.

The auto-correlation and cross-correlation functions of the in-phase and quadrature components, and the auto-correlation functions of the complex envelope of fading signal $Z(t)$ are given by

$$R_{Z_cZ_c}(\tau) = R_{Z_sZ_s}(\tau) = \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau) + K\cos(2\pi f_0\tau)}{2(1+K)} \tag{4}$$

$$R_{Z_cZ_s}(\tau) = -R_{Z_sZ_c}(\tau) = \frac{K\sin(2\pi f_0\tau)}{2(1+K)} \tag{5}$$

$$R_{ZZ}(\tau) = \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau) + K\exp(j2\pi f_0\tau)}{(1+K)} = \frac{2J_0(2\pi a f_2\tau)J_0(2\pi f_2\tau) + K\exp(j2\pi f_0\tau)}{(1+K)} \tag{6}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind, $a = f_1/f_2$ is the ratio of two maximum Doppler frequencies (or vehicles speeds), and here assuming $f_1 \leq f_2$. $a = 0$ means that the transmitter is stationary and then equation (6) gives the auto-correlation for I2V channels, which indicates that I2V channels are a special case of IVC channels in VANETs.

Time-averaging is often used in place of ensemble averaging in simulation practice. The auto-correlation function of the real part of $Z(t)$ for one trial (sample of the process) then becomes

$$\begin{aligned}\hat{R}_{Z_cZ_c}(\tau) &= \lim_{T\to\infty}\frac{1}{T}\int_0^T Z_c(t)Z_c(t+\tau)dt \\ &= \frac{1}{N}\sum_{n=1}^{N}\cos(2\pi f_1 cos\alpha_n + 2\pi f_2 cos\beta_n)\tau\end{aligned} \tag{7}$$

Furthermore, the time averaging changes from trial to trial due to the random angle. As pointed out in (Xiao et al., 2006), the variance of the time average $Var[R(\cdot)] = E[|R(\cdot) - \hat{R}(\cdot)|^2]$ provides a measure of the closeness of the model in simulating the desired channel with a finite number of sinusoids. A lower variance indicates that a smaller number of simulation trials are needed to achieve the desired statistical properties, and, the convergence of the corresponding model is faster. The time-averaged variances of the aforementioned correlation statistics are derived as

$$Var[R_{Z_cZ_c}(\tau)] = Var[R_{Z_sZ_s}(\tau)] = [\frac{1 + J_0(4\pi a f_2\tau)J_0(4\pi f_2\tau) - 2J_0^2(2\pi a f_2\tau)J_0^2(2\pi f_2\tau)}{2N}]/(1+K)^2 \tag{8}$$

$$Var[R_{Z_cZ_s}(\tau)] = Var[R_{Z_sZ_c}(\tau)] = [\frac{1 - J_0(4\pi a f_2\tau)J_0(4\pi f_2\tau)}{2N}]/(1+K)^2 \tag{9}$$

$$Var[R_{ZZ}(\tau)] = Var[R_{ZZ}(\tau)] = [\frac{4 - 4J_0^2(2\pi a f_2\tau)J_0^2(2\pi f_2\tau)}{N}]/(1+K)^2 \tag{10}$$

In next sections, we use these statistics to compare the performance of the proposed models with this reference model.

5. Two SoS simulation models for VANETs

This section proposes two Rician channel models with a LOS component for VANETs by introducing the aforementioned LOS component and extends two SoS models (statistical and deterministic simulation models). Important statistical properties for the proposed Rician models are derived and provided for comparison purposes.

5.1 Statistical SoS model

Recently, A new model is proposed in (Zajic & Stuber, 2006) to directly generate multiple uncorrelated complex envelope, which has resolved the difficulty in producing time averaged auto- and cross-correlation functions that match the reference model (Akki & Haber, 1989).

The k^{th} complex faded envelope is given by (Zajic & Stuber, 2006)

$$Y_k(t) = Y_{ck}(t) + jY_{sk}(t) \tag{11}$$

where

$$Y_{ck}(t) = \frac{2}{\sqrt{N_0 M}} \sum_{n=1}^{N_0} \sum_{m=1}^{M} \cos\left[2\pi f_2 t \cos\left(\beta_{mk}\right)\right] \cos\left[\left(2\pi f_1 t \cos\left(\alpha_{nk}\right) + \phi_{nmk}\right]\right. \tag{12}$$

$$Y_{sk}(t) = \frac{2}{\sqrt{N_0 M}} \sum_{n=1}^{N_0} \sum_{m=1}^{M} \sin\left[2\pi f_2 t \cos\left(\beta_{mk}\right)\right] \sin\left[\left(2\pi f_1 t \sin\left(\alpha_{nk}\right) + \phi_{nmk}\right]\right. \tag{13}$$

f_1, f_2, α_{nk}, β_{mk} and ϕ_{nmk} are the maximum radian Doppler frequencies, the random angle of departure, the random angle of arrival, and the random phase, respectively. It is assumed that P independent complex faded envelopes are required $(k = 0, ..., P-1)$ each consisting of NM sinusoidal components. The angles of departures and the angles of arrivals are chosen as follows:

$$\alpha_{nk} = \frac{2\pi n}{4N_0} + \frac{2\pi k}{4PN_0} + \frac{\theta - \pi}{4N_0} \tag{14}$$

$$\beta_{mk} = \frac{1}{2}\left(\frac{2\pi m}{M} + \frac{2\pi k}{PM} + \frac{\psi - \pi}{M}\right) \tag{15}$$

where $n = 1, ..., N_0$, $m = 1, ..M$, $k = 0, ..., P-1$. The angles of departures and the angles of arrivals in the k^{th} complex faded envelope are obtained by rotating the angles of departures and the angles of arrivals in the $(k-1)^{th}$ complex envelope by $2\pi/(4PN_0)$ and $2\pi/(2PM)$, respectively. The parameters ϕ_{nmk}, θ and ψ are independent random variables uniformly distributed on the interval $[-\pi, \pi)$.

With reference to $(2), (11), (12), (13)$, the received complex envelope of the IVC fading channels can be obtained as follows:

$$Z_k(t) = \frac{Y_k(t) + \sqrt{K}\exp(j2\pi f_0 t + \phi_0)}{\sqrt{1+K}} \tag{16}$$

The time-average auto-correlation and cross-correlation function of the in-phase and quadrature components, and the auto-correlation functions of the complex envelope of fading signal $Z_k(t)$ are given by

$$\hat{R}_{Z_{ck}Z_{ck}}(\tau) = \hat{R}_{Z_{sk}Z_{sk}}(\tau) = \frac{2J_0(2\pi f_1 \tau)J_0(2\pi f_2 \tau) + K\cos\left(2\pi f_0 \tau\right)}{2(1+K)} \tag{17}$$

$$\hat{R}_{Z_{ck}Z_{sk}}(\tau) = -\hat{R}_{Z_{sk}Z_{ck}}(\tau) = \frac{K\sin(2\pi f_0 \tau)}{2(1+K)} \tag{18}$$

$$\hat{R}_{Z_k Z_k}(\tau) = \frac{2J_0(2\pi f_1 \tau)J_0(2\pi f_2 \tau) + K\exp\left(j2\pi f_0 \tau\right)}{1+K} \tag{19}$$

Proof: we first prove the equation (17)

$$
\begin{aligned}
&R_{Z_{ck}Z_{ck}}(\tau)\\
&= E[Z_{ck}(t)Z_{ck}(t+\tau)]\\
&= \frac{1}{1+K}\{\frac{4}{N_0M}E[\sum_{n,m=1}^{N_0,M}\cos(2\pi f_1 t\cos\alpha_{nk}+\phi_{nmk})\cos(2\pi f_2 t\cos\beta_{mk})\\
&\quad\cdot\sum_{p,q=1}^{N_0,M}\cos(2\pi f_1(t+\tau)\cos\alpha_{pk}+\phi_{pqk})\cos(2\pi f_2(t+\tau)\cos\beta_{qk})]+KE[\cos(2\pi f_0 t+\phi_0)\\
&\quad\cdot\cos(2\pi f_0(t+\tau)+\phi_0)]+2\sqrt{\frac{K}{N_0M}}E[\sum_{n,m=1}^{N_0,M}\cos(2\pi f_1 t\cos\alpha_{nk}+\phi_{nmk})\cos(2\pi f_2 t\cos\beta_{mk})\\
&\quad\cdot\cos(2\pi f_0(t+\tau)+\phi_0)]+E[\cos(2\pi f_0 t+\phi_0)\\
&\quad\cdot2\sqrt{\frac{K}{N_0M}}\sum_{n,m=1}^{N_0,M}\cos(2\pi f_1(t+\tau)\cos\alpha_{nk}+\phi_{nmk})\cos(2\pi f_2(t+\tau)\cos\beta_{mk})]\}\\
&= \frac{1}{1+K}\{\frac{1}{N_0M}E[\sum_{n,m=1}^{N_0,M}\cos(2\pi f_1\tau\cos\alpha_{nk})\cos(2\pi f_2\tau\cos\beta_{mk})]\}+\frac{K}{2(1+K)}\cos(2\pi f_0\tau)\}\\
&= \frac{1}{1+K}\{\frac{1}{N_0}\sum_{n=1}^{N_0}\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos[2\pi f_1\tau\cos(\frac{2\pi n}{4N_0}+\frac{2\pi k}{4PN_0}+\frac{\theta-\pi}{4N_0})]d\theta\\
&\quad\cdot\frac{1}{M}\sum_{m=1}^{M}\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos[2\pi f_2\tau\cos(\frac{2\pi m}{2M}+\frac{2\pi k}{2PM}+\frac{\psi-\pi}{2M})]d\psi\}+\frac{K}{2(1+K)}\cos(2\pi f_0\tau)\\
&= \frac{1}{1+K}\{\frac{1}{N_0}\sum_{n=1}^{N_0}\frac{1}{2\pi}\int_{\frac{2\pi(n-1)}{4N_0}+\frac{2\pi k}{4PN_0}}^{\frac{2\pi n}{4N_0}+\frac{2\pi k}{4PN_0}}\cos(2\pi f_1\tau\cos\gamma_n)\cdot4N_0d\gamma_n\\
&\quad\cdot\frac{1}{M}\sum_{n=1}^{M}\frac{1}{2\pi}\int_{\frac{2\pi(m-1)}{2M}+\frac{2\pi k}{2PM}}^{\frac{2\pi M}{2M}+\frac{2\pi k}{2PM}}\cos(2\pi f_2\tau\cos\gamma_m)\cdot2Md\gamma_m\}+\frac{K}{2(1+K)}\cos(2\pi f_0\tau)\\
&= \frac{1}{1+K}\{\frac{1}{N_0}\cdot\frac{1}{2\pi}\int_{\frac{2\pi k}{4PN_0}}^{\frac{\pi}{2}+\frac{2\pi k}{4PN_0}}\cos(2\pi f_1\tau\cos\gamma_n)\cdot4N_0d\gamma_n\\
&\quad\cdot\frac{1}{M}\cdot\frac{1}{2\pi}\int_{\frac{2\pi k}{2PM}}^{\pi+\frac{2\pi k}{2PM}}\cos(2\pi f_2\tau\cos\gamma_m)\cdot2Md\gamma_m\}+\frac{K}{2(1+K)}\cos(2\pi f_0\tau)\\
&= \frac{1}{1+K}[\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\cos(2\pi f_1\tau\cos\gamma_1)d\gamma_1\cdot\frac{1}{\pi}\int_0^{\pi}\cos(2\pi f_2\tau\cos\gamma_2)d\gamma_2]\frac{K}{2(1+K)}\cos(2\pi f_0\tau)\\
&= \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau)+K\cos(2\pi f_0\tau)}{2(1+K)}
\end{aligned}
$$

This completes the proof of (17). Similarly, one can prove the (18) and (19), details are omitted for brevity.

The time-averaged variances of above correlation statistics of $Z_k(t)$ are presented as

$$\begin{aligned} Var\{\hat{R}_{Z_{ck}Z_{ck}}(\tau)\} &= Var\{\hat{R}_{Z_{sk}Z_{sk}}(\tau)\} \\ &= [\frac{1+J_0(4\pi f_1\tau)J_0(4\pi f_2\tau)}{4N_0M} - f_c(2\pi f_1\tau, 2\pi f_2\tau)]/(1+K)^2 \end{aligned} \tag{20}$$

$$Var\{\hat{R}_{Z_{ck}Z_{sk}}(\tau)\} = Var\{\hat{R}_{Z_{sk}Z_{ck}}(\tau)\} = 0 \tag{21}$$

$$Var\{\hat{R}_{ZZ}(\tau)\} = [\frac{1+J_0(4\pi f_1\tau)J_0(4\pi f_2\tau)}{N_0M} - 4f_c(2\pi f_1\tau, 2\pi f_2\tau)]/(1+K)^2 \tag{22}$$

where

$$\begin{aligned} f_c(2\pi f_1\tau, 2\pi f_2\tau) = \sum_{k_1=1}^{N_0}[\frac{1}{2\pi}\int_{\frac{2\pi(k_1-1)}{N}+\frac{2\pi k}{4PN_0}}^{\frac{2\pi k_1}{N}+\frac{2\pi k}{4PN_0}}\cos(2\pi f_1\tau cos\gamma_1)d\gamma_1]^2 \cdot \\ \sum_{k_2=1}^{M}[\frac{1}{2\pi}\int_{\frac{2\pi(k_2-1)}{M}+\frac{2\pi k}{2PM}}^{\frac{2\pi k_2}{M}+\frac{2\pi k}{2PM}}\cos(2\pi f_2\tau cos\gamma_2)d\gamma_2]^2 \end{aligned} \tag{23}$$

Proof: We start with the first part of (20) and derive

$$\begin{aligned} &Var\{\hat{R}_{Z_{ck}Z_{ck}}(\tau)\} \\ &= E[|\hat{R}_{Z_{ck}Z_{ck}}(\tau) - \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau)+K\cos(2\pi f_0\tau)}{2(1+K)}|^2] \\ &= E[|\hat{R}_{Z_{ck}Z_{ck}}(\tau)|^2] - \frac{J_0^2(2\pi f_1\tau)J_0^2(2\pi f_2\tau)}{(1+K)^2} - [\frac{K\cos 2\pi f_0\tau}{2(1+K)}]^2 \\ &= E\{\frac{1}{(1+K)^2}\cdot\frac{1}{N_0^2M^2}[\sum_{n,m=1}^{N_0,M}\cos(2\pi f_1\tau\cos\alpha_{nk})\cos(2\pi f_2\tau\cos\beta_{mk}) \\ &\quad\cdot\sum_{p,q=1}^{N_0,M}\cos(2\pi f_1\tau\cos\alpha_{pk})\cos(2\pi f_2\tau\cos\beta_{qk})]\} - \frac{J_0^2(2\pi f_1\tau)J_0^2(2\pi f_2\tau)}{(1+K)^2} \\ &= \frac{1}{(1+K)^2}\cdot\frac{1}{N_0^2M^2}\{E[\sum_{n,m=1}^{N_0,M}\cos^2(2\pi f_1\tau\cos\alpha_{nk})\cos^2(2\pi f_2\tau\cos\beta_{mk})] \\ &\quad\cdot\sum_{n,m=1}^{N_0,M}\sum_{p,q=1}^{N_0,M}E[\cos(2\pi f_1\tau\cos\alpha_{nk})\cos(2\pi f_2\tau\cos\beta_{mk})]E[\cos(2\pi f_1\tau\cos\alpha_{pk}) \\ &\quad\cdot\cos(2\pi f_2\tau\cos\beta_{qk})]\} - \frac{J_0^2(2\pi f_1\tau)J_0^2(2\pi f_2\tau)}{(1+K)^2} \qquad (n\neq p \quad or \quad m\neq q) \\ &= \frac{1}{(1+K)^2}\cdot\frac{1}{N_0^2M^2}\{N_0M\cdot\frac{1+J_0(4\pi f_1\tau)J_0(4\pi f_2\tau)}{4} \\ &\quad+N_0^2M^2[J_0^2(2\pi f_1\tau)J_0^2(2\pi f_2\tau) - f_c(2\pi f_1\tau, 2\pi f_2\tau)]\} - \frac{J_0^2(2\pi f_1\tau)J_0^2(2\pi f_2\tau)}{(1+K)^2} \\ &= [\frac{1+J_0(4\pi f_1\tau)J_0(4\pi f_2\tau)}{4N_0M} - f_c(2\pi f_1\tau, 2\pi f_2\tau)]/(1+K)^2 \end{aligned}$$

Similarly, we can validate the second part of (20) and equation (21). Thus, we have

$$
\begin{aligned}
&Var\{\hat{R}_{Z_kZ_k}(\tau)\} \\
&= E[|\hat{R}_{Z_kZ_k}(\tau) - \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau) + K\exp(j2\pi f_0\tau)}{1+K}|^2] \\
&= E[|2\hat{R}_{Z_{ck}Z_{ck}}(\tau) + j2\hat{R}_{Z_{ck}Z_{sk}}(\tau) - \frac{2J_0(2\pi f_1\tau)J_0(2\pi f_2\tau)}{(1+K)} - \frac{K\exp j2\pi f_0\tau}{(1+K)}|]^2 \\
&= [\frac{1+J_0(4\pi f_1\tau)J_0(4\pi f_2\tau)}{N_0M} - 4f_c(2\pi f_1\tau, 2\pi f_2\tau)]/(1+K)^2
\end{aligned}
$$

This completes the proof.

5.2 Deterministic SoS model

The model proposed above may require several simulation trials to converge to the desired properties. A low-complexity alternative is described in this section. It was recently used for IVC channels with no LOS component (Patel et al., 2005) and called the MEDS model. The complex faded envelope generated by the MEDS model is given by

$$Y(t) = Y_c(t) + jY_s(t) \tag{24}$$

$$Y_c(t) = \sqrt{\frac{2}{N_cM_c}} \sum_{n,m=1}^{N_c,M_c} \cos(2\pi f_{1,n}^c t + 2\pi f_{2,m}^c t + \phi_{nm}^c) \tag{25}$$

$$Y_s(t) = \sqrt{\frac{2}{N_sM_s}} \sum_{n,m=1}^{N_s,M_s} \cos(2\pi f_{1,n}^s t + 2\pi f_{2,m}^s t + \phi_{nm}^s) \tag{26}$$

$$f_{1,n}^{c/s} = f_1\cos(\frac{\pi(n-0.5)}{2N_{c/s}}) \qquad n = 1,2,...N_{c/s} \tag{27}$$

$$f_{2,m}^{c/s} = f_2\cos(\frac{\pi(m-0.5)}{M_{c/s}}) \qquad m = 1,2,...M_{c/s} \tag{28}$$

where the phase $\phi_{nm} \sim U[-\pi,\pi)$ is independent for all n,m and the in-phase and quadrature components.

Similarly, with reference to $(2),(25),(26),(27),(28)$, the complex signal of the IVC channel model are expressed as

$$Z(t) = \frac{Y(t) + \sqrt{K}\exp(j2\pi f_0 t + \phi_0)}{\sqrt{1+K}} \tag{29}$$

As described in (Patel et al., 2005), all the frequencies, $f_{1,n}^c, f_{2,m}^c$ and $f_{1,k}^s, f_{2,l}^s$ must be distinct. In addition, $f_{1,n}^c, f_{2,m}^c$ and $f_{1,k}^c, f_{2,l}^c$ have also to be distinct. From simulations, we found that with $N_c = M_c = N_C$ and $N_s = M_s = N_C + 1$, the Doppler frequencies are indeed distinct for practical ranges varying from 5 to 60. Under these assumptions, it can be shown that the time-average correlations are equal to the statistical correlations.

$$\hat{R}_{Z_cZ_c}(\tau) = \frac{1}{1+K}[\frac{1}{N_C^2}\sum_{n,m=1}^{N_c,N_c}\cos\{2\pi f_{1,n}^c\tau + 2\pi f_{2,m}^c\tau\} + \frac{K\cos(2\pi f_0\tau)}{2}] \quad (30)$$

$$\hat{R}_{Z_sZ_s}(\tau) = \frac{1}{1+K}[\frac{1}{(N_C+1)^2}\sum_{n,m=1}^{N_c+1,N_c+1}\cos\{2\pi f_{1,n}^s\tau + 2\pi f_{2,m}^s\tau\} + \frac{K\cos(2\pi f_0\tau)}{2}] \quad (31)$$

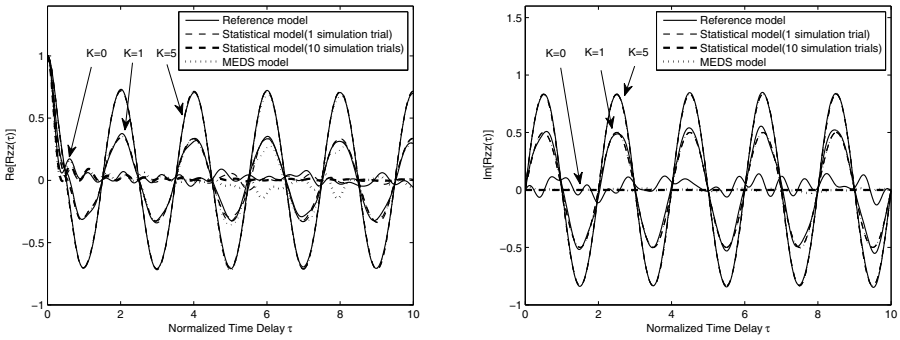
$$\hat{R}_{Z_cZ_s}(\tau) = -\hat{R}_{Z_sZ_c}(\tau) = \frac{K\sin(2\pi f_0\tau)}{2(1+K)} \quad (32)$$

$$\hat{R}_{ZZ}(\tau) = \frac{1}{1+K}[\frac{1}{N_C^2}\sum_{n,m=1}^{N_c,N_c}\cos\{2\pi f_{1,n}^c\tau + 2\pi f_{2,m}^c\tau\} + \frac{1}{(N_C+1)^2}\sum_{n,m=1}^{N_c+1,N_c+1}\cos\{2\pi f_{1,n}^s\tau + 2\pi f_{2,m}^s\tau\} + K\exp(2\pi f_0\tau)] \quad (33)$$

Remark: The expressions for variances of the correlation functions for the proposed MEDS model cannot be obtained in a simplified form, here are not provided.

6. Performance analysis and comparison

This section compares the performance of the proposed simulation models. Unless stated otherwise, all the simulation results presented here are obtained using a normalized sampling period $f_1T_S = 0.01$ (T_s is the sampling period).

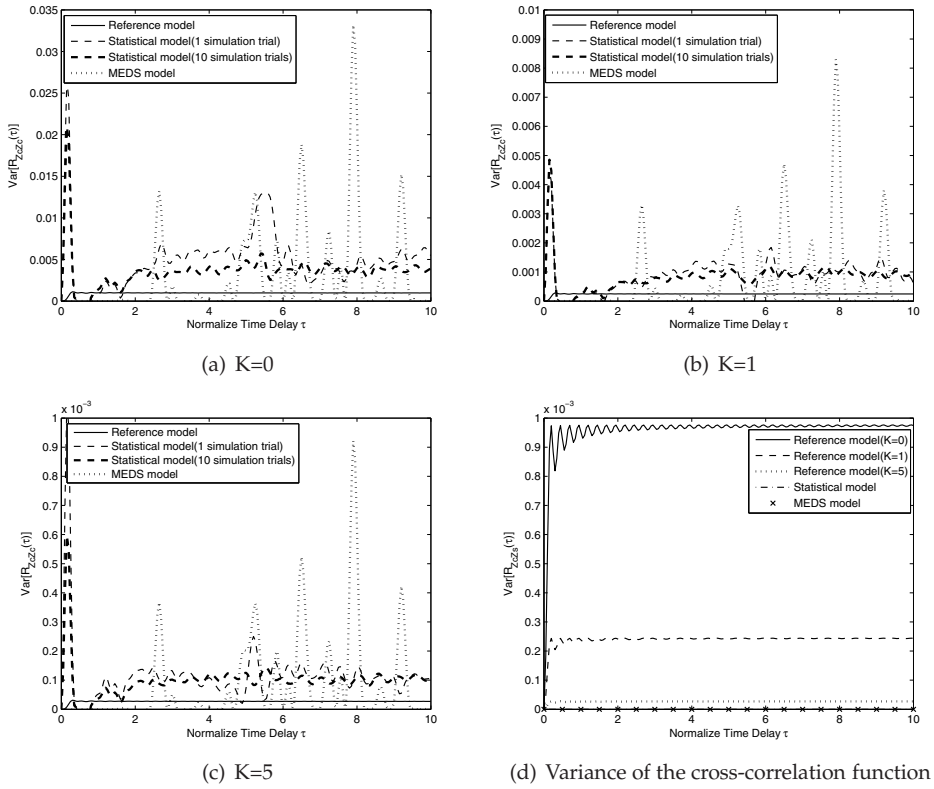


(a) (b)

Fig. 2. The auto-correlation function of the complex envelope with different K

6.1 Effects of Rician factor in VANETs

The results of Figs. 2-3 are obtained using $a = 1, f_1 = f_2 = 50Hz, N_0 = M = N_C = P = 8$. For a fair comparison, we use $N = 4N_0 \times 2M = 512$ sinusoids for simulation of the reference model.



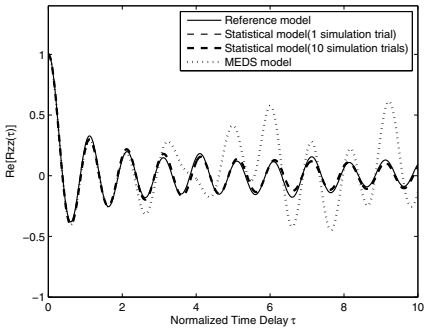
(a) K=0 (b) K=1

(c) K=5 (d) Variance of the cross-correlation function

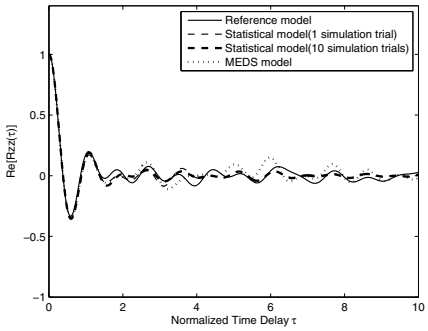
Fig. 3. Variance of the auto-correlation and cross-correlation function with different K

Fig. 2 shows the correlation properties of the aforementioned models with different K factors in VANETs. For a large range of normalized time-delay ($0 \leq f_1 T_S \leq 4$), the proposed simulation models keep good agreement with the reference model, without exhibiting any sort of periodicity as encountered in Wang and Cox's model (Wang & Cox, 2002). For the same time delay τ, the magnitude of the channel correlation tends to be larger. As the K factor increases, the proposed models get closer to the reference model.

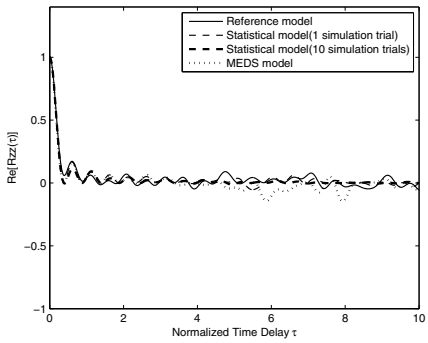
Fig. 3 compares the variances of the auto- and cross-correlation functions for the proposed simulation models. As shown in Figs. 3a-3c, the variances of the auto-correlation functions decrease as K factor increases. It indicates that the simulation models perform better under a larger amount of LOS components. When K is larger, the LOS components become more dominant over the scattering components, which avoids the deviation caused by the finite scatters. We can also observe that the variances of the auto-correlation of our models are higher than the reference model. It is noted that the difference between the statistical model and the reference model becomes smaller when the number of the simulation trials is increased. Simulation results show that the statistical model achieves better convergence by averaging 10 simulation trials. Fig. 3d shows that the variances of the cross-correlation for the proposed models are lower than the reference model.



(a) a=0



(b) a=0.2



(c) a=1

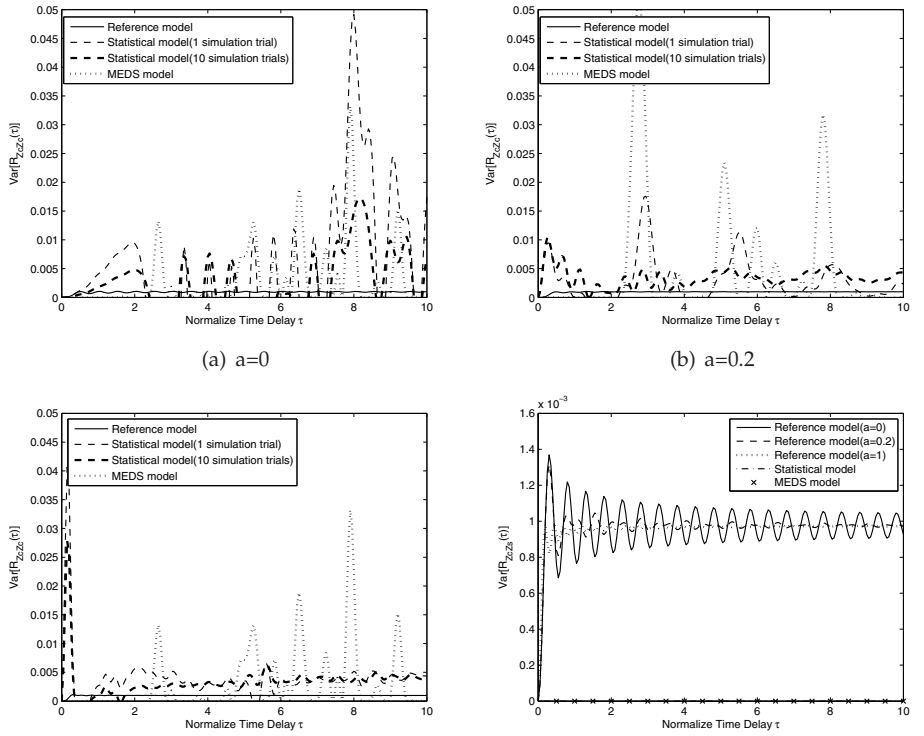
Fig. 4. Real part of the auto-correlation function of the complex envelope with different a

It should be noted that the plot of the auto-correlation function when $K = 0$ keeps agreement with (Patel et al., 2005) and the result of the variance of auto-correlation is almost similar to that in (Zajic & Stuber, 2006), which indicate that our models and performance analysis are more comprehensive than the existing ones.

6.2 Effects of vehicle speed ratio in VANETs

The simulation results presented in Figs. 4-5 are obtained using $K = 0, f_2 = 50Hz, f_1 = 0, 10$ and $50Hz$ when the corresponding value of speed ratio a equals to $0, 0.2$ and 1. The imaginary part of the auto-correlation of the complex envelope for the proposed two SoS models is always equal to 0, which is in line with the ideal situation and shows better performance compared with the reference model.

Fig. 4 shows the real part of correlation properties of the above models with different a in VANETs. It is observed that the proposed models provide a better approximation to the desired auto-correlation when a increases. From Figs. 5a-5c, we found that the variances of the auto-correlation of our models tend to be lower with a larger value of a. So the proposed models perform better with a smaller relative speed difference. A physical interpretation



(a) a=0 (b) a=0.2

(c) a=1 (d) Variance of the cross-correlation function

Fig. 5. Variance of the auto-correlation and cross-correlation function with different a

should be like this: smaller relative difference means that the transmitter and the receiver are in a relatively closer static state, so the randomness of channel tends to be smaller. As plotted in Fig. 5d, the variances of the cross-correlation of the proposed models are considerably lower than the reference model.

It should also be mentioned that our simulation results of variance of the correlation when $a = 0$ represent the case of I2V channels, which shows the comprehensiveness of our models and performance analysis as well.

7. Conclusion

The automotive industry conveniently built powerful and safer cars by embedding advanced materials and sensors. With the advent of wireless communication technologies, cars are being equipped with wireless communication devices, enabling them to communicate with others. Such communications are not plainly restricted to data transfers (such as GPS, video and audio, emails, etc.), but also create new opportunities for enhancing road safety. Some applications only require communication among vehicles, while many others require the coordination between vehicles and road-side infrastructure. Recent technological developments, notably in mobile computing, wireless communication, and

remote sensing are pushing ITS towards a major leap forward. Vehicles become sophisticated computing systems, with several computers and sensors onboard, each dedicated to certain car operations. Interconnected vehicles do not only collect information about themselves and their environments, but they also exchange the information in real time with other nearby vehicles. As radio-communication-based solutions can operate beyond the line-of-sight constraints, they can enable cooperative approaches. Vehicles and infrastructure cooperate to perceive potentially dangerous situations in an extended space and time horizon. Appropriate vehicular communication architectures are necessary to create reliable and extended driving support systems for road safety and transportation efficiency.

A number of technical challenges need to be resolved in order to deploy vehicular networks and to provide related services for drivers and passengers in such networks. Scalability and interoperability are two important issues that should be addressed. The employed protocols and mechanisms should be scalable to numerous vehicles and interoperable with different wireless technologies, such as reliable link performance and MAC protocols, routing and dissemination, IP configuration and mobility management, security etc. As a key component of the ITS, vehicular wireless networks, has attracted research attention from both the academia and industry of US, EU, and Japan. Although many works have been done on communication and routing protocol, only few models have been developed to characterize the fading effect in vehicular wireless network.

In this chapter, we proposed a new statistical and deterministic SoS model for IVC fading channels with a LOS component in VANETs. The properties of the proposed models were derived and verified in terms of the auto-correlation and the cross-correlation by comparison between theoretical and simulation results. The statistics of them match those of the reference model for a large range of normalized time delays ($0 \leq f_1 T_S \leq 4$). When the K factor increases, the proposed SoS models show an improved approximation of the desired auto-correlation and faster convergence. And then we described the curves of the statistics for the simulation models with different vehicle speed ratios, which indicates that the smaller relative speed difference of two vehicle speeds in VANETs contribute to the better performance of the proposed models. More importantly, Based on our models, we provided more comprehensive performance analysis and comparison compared to existing models.

It is observed that the the variances in the cross-correlation for the statistical model and the MEDS model are considerably lower than those of reference model. Meanwhile, For the same time delay τ, the statistical model shows better performance and faster convergence than the MEDS model. Hence, the statistical model may be more suitable for IVC fading channels with a LOS component in VANETs.

8. Acknowledgments

The authors would like to thank the Cognitive Radio Sensor Network research group in Electronic Information Engineering of Nanchang University for continuous support and lively discussions. The authors are also grateful to the anonymous reviewers for their helpful comments.

This work has been supported by the National Natural Science Foundation of China (No.60762 005), the Natural Science Foundation of Jiangxi Province for Youth (No.2010GQS0153 and No.2009GQS0070) and the Graduate Student Innovation Foundation of Jiangxi Province (No.YC10A032).

9. References

M. Boban, O. K. Tonguz & J. Barros (2009). Unicast Communication in Vehicular Ad Hoc Networks: A Reality Check, *IEEE Commun. Lett.*, Vol. 13, No. 12, Dec. 2009, 995-997.

T. L. Willke, P. Tientrakool & N. F.Maxemchuk (2009). A Survey of Inter-Vehicle Communication Protocols and Their Applications, *IEEE Commun. Surv. & Tuto.*, Vol. 11, No. 2, Jun. 2009, 3-20.

A. F. Molisch, F.Tufvesson, J.Karedal & C.F.Mecklenbrauker (2009). A Survey on Vehicle-to-Vehicle Propagation Channels, *IEEE Trans. Wireless Commun.*, Vol.16, No.6, Dec. 2009, 12-22.

A. M. Abbas & O.Kure (2010). Quality of Service in mobile ad hoc networks: a survey, *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. 6, No.2, Jul. 2010, 75-98.

T. Lwinmuller, E. Schoch & F. Kargl (2006). Postion Verification Approaches for Vechicular Ad Hoc Networks, *IEEE Trans.Commun.*, Vol. 13, No. 5, Jan. 2006, 278-285.

T. Taleb, E. Sakhaee, A. Jamalipour, K. Hashimoto, N. Kato & Y. Nemoto (2007). A Stable Routing Protocol to Support ITS Services in VANET Networks, *IEEE Trans. Veh. Technol.*, Vol. 56. No. 6, Nov. 2007, 3337-3347.

D. Kim, J.J. Garcia-Luna-Aceves, K. Obraczka, J. C. Cano & P. Manzoni (2003). Routing Mechanisms for Mobile Ad Hoc Networks Based on the Energy Drain Rate, *IEEE Trans. on mobile computing*, Vol. 2, No. 2, April-June. 2003, 161-173.

X. J. Du & D. P. Wu (2007). Adaptive Cell Relay Routing Protocol for Mobile Ad Hoc Networks *IEEE Trans. Veh. Technol.*, Vol. 55. No. 1, Nov. 2007, 3381-3396.

H. Y. Huang, P. E. Luo, M. L. Li, D. Li, X. Li, W. Shu & M. Y. Wu (2007). Performance Evaluation of SUVnet With Real-Time Traffic Data, *IEEE Trans. Veh.Technol.*, Vol. 56. No. 6,Nov. 2007, 3381-3396.

N. Wisitpongphan, F. Bai, P. Mudalige, V. Sadekar, & O. Tonguz (2007). Routing in Sparse Vehicular Ad Hoc Wireless Networks, *IEEE J. Sel. Areas in Commun.*, Vol. 25, No. 8, Oct. 2007, 1538-1556.

J. Nzouonta, N. Rajgure & G. Wang (2009). VANET Routing on City Roads Using Real-Time Vehicular Traffic Information, *IEEE Trans Veh. Technol.*, Vol. 58, No. 7, Sep. 2009, 3609-3626.

W. J. Wang, F. Xie & M. Chatterjee (2009). Small-Scale and Large-Scale Routing in Vehicular Ad Hoc Networks, *IEEE Trans Veh. Technol.*, Vol. 58, No. 9, Nov. 2009, 5200-5213.

H. Saleet, O. Basir, R. Langar & R. Boutaba (2010). Region-Based Location-Service-Management Protocol for VANETs, *IEEE Trans Veh. Technol.*, Vol. 59, No. 2, Feb. 2010, 917-931.

V. Tarokh (2009). In: *New Directions in wireless Communications Research*, Springer Dordrecht Heidelberg London New York, 2009, 1-25.

C. S. Patel, G.L.Stuber & Thomas G.Pratt (2005). Simulation of Rayleigh-faded mobile-to-mobile communication channels, *IEEE Trans. Commun.*, Vol. 53, No. 11, Nov. 2005, 1876-1884.

A. G. Zajic & G. L. Stuber (2006). A New Simulation Model for Mobile-to-Mobile Rayleigh Fading Channels, in *IEEE Wireless Commun. Netw. Conf., (WCNC'06)*, pp. 1266-1270, Apr. 2006.

L. C. Wang, W. C. Liu & Y. H. Cheng (2009). Statisitical Analysis of a Mobile-to-Mobile Rician Fading Channel Model, *IEEE Trans. Veh. Technol.*, Vol. 58, No. 1, Jan. 2009, 32-38.

R. H. Clarke (1968). A statistical theory of mobile-radio reception, *Bell Syst. Tech. J.*, Aug. 1968, 957-1000.

M. J. Gans (1972). A power-spectral theory of propagation in the mobile-radio environment, *IEEE Trans. Veh. Technol.*, Vol. 21, No. 1, Feb. 1972, 27-38.

W. C. Jakes (1994). *Microwave Mobile Communications*. Wiley, 1974; re-issued by IEEE Press, 1994.

P. Dent, G. E. Bottomley & T. Croft (1993). Jakes fading model revisited, *IEEE Electron. Lett.*, Vol. 29, No. 13, Jun. 1993, 1162-1163.

M. Patzold, U. Killat, F. Laue, & Y. C. Li (1998). On the statistical properties of deterministic simulation models for mobile fading channels, *IEEE Trans. Veh. Technol.*, Vol. 47, No. 1, Feb. 1998, 254-269.

K.-W. Yip and T.-S. Ng, "A simulation model for Nakagami-m fading channels, m < 1," *IEEE Trans. Commun.*, vol. 48, No. 2, Feb.2000, 214-221.

M. F. Pop & N. C. Beaulieu (2001). Limitations of sum-of-sinusoids fading channel simulators, *IEEE Trans. Commun.*, Vol. 49, No. 4, Apr. 2001, 699-708.

Y. X. Li & X. Huang (2002). The generation of independent Rayleigh faders, in *Proc. IEEE ICC*, pp. 41-45, Jun. 2000.

C. Xiao & Y. R. Zheng (2002). A generalized simulation model for Rayleigh fading channels with accurate second-order statistics, in *Proc. IEEE VTC-Spring*, pp. 170-174, May 2002.

C. Xiao, Y. R. Zheng, & N. C. Beaulieu (2002). Second-order statistical properties of the WSS Jakesą́ fading channel simulator, *IEEE Trans. Commun.*, Vol. 50, No.6, Jun. 2002, 888-891.

Y. R. Zheng & C. Xiao (2002). Improved models for the generation of multiple uncorrelated Rayleigh fading waveforms, *IEEE Commun. Lett.*, Vol. 6, No.6, Jun. 2002, 256-258.

Y. R. Zheng & C. Xiao (2003). Simulation models with correct statistical properties for Rayleigh fading channels, *IEEE Trans. Commun.*, Vol. 51, No.6, Jun. 2003, 920-928.

C.Xiao, Y.R.Zheng & N.C.Beaulieu (2006). Novel sum-of-sinusoids simulation models for Rayleigh and Rician fading channels, *IEEE Trans. Wireless Commun.*, Vol.5, No.12, Dec. 2006, 3667-3679.

A. S. Akki & F. Haber (1989). A statistical model for mobile-to-mobile land communication channel, *IEEE Trans. Veh. Technol.*, Vol. 35, No. 1, Feb. 1986, 2-7.

A. S. Akki (1994). Statistical properties of mobile-to-mobile land communication channels, *IEEE Trans. Veh. Technol.*, Vol. 43, No. 4, Nov. 1994, 826-831.

F. Vatalaro & A. Forcella (1997). Doppler spectrum in mobile-to-mobile communications in the presence of three-dimensional multipath scattering, *IEEE Trans. Veh. Technol.*, Vol. 46, No. 1, Feb. 1997, 213-219.

J. M. G. Linnartz & R. F. Fiesta (1996). Evaluation of radio links and networks. [Online].Available: http://www.path.berkeley.edu/PATH/ Publications/PDF/PRR/96/ PRR-96-16.pdf.

I. Z. Kovacs, P. C. F. Eggers, K. Olesen & L. G. Petersen (2002). Investigations of outdoor-to-indoor mobile-to-mobile radio communication channels, in *Proc. IEEE Veh. Technol. Conf.*, Vancouver, BC, Canada, pp. 430-434, Sep. 2002.

J. Maurer, T. Fugen, K. Olesen & W. Wiesbeck (2002). Narrowband measurement and analysis of the intervehicle transmission channel at 5.2 GHz, in *Proc. IEEE Veh. Technol. Conf.*, Birmingham, AL, pp. 1274-1278, May 2002.

L. Cheng, B. E. Henty, D. D. Stancil, F. Bai & P. Mudalige (2007). Mobile Vehicle-to-Vehicle Narrow-Band Channel Measurement and Characterization of the 5.9 GHz Dedicated

Short Range Communication (DSRC) Frequency Band, *IEEE J. Sel. Areas in Commun.*, Vol. 25, No. 8, Oct. 2007, 1501-1516.

R. Wang & D. Cox (2002). Channel modeling for ad hoc mobile wireless networks, in *Proc. IEEE Veh. Technol. Conf.*, pp. 21-25, May 2002.

D. J. Young & N. C. Beaulieu (2000). The generation of correlated Rayleigh random variates by inverse discrete Fourier transform, *IEEE Trans. Commun.*, Vol. 48, No. 7, Jul. 2000, 1114-1127.

C. S. Patel, S.L. Stuber & T.G. Pratt (2003). Simulation of Rayleigh faded mobile-to-mobile communication channels, *IEEE Veh. Technol. Conf.*, pp. 163-167, Oct. 2003.

S. Gowrishankar, T. G. Basavaraju and S. K. Sarkar (2007), Effect of Random Mobility Models Pattern in Mobile Ad hoc Networks, *Int. J. Computer Science and Network*, Vol. 7, No. 6, Jun. 2007, 160-164.

J. Yoon, B. Noble (2006), A General Framework to Construct Stationary Mobility Models for the Simulation of Mobile Networks, *IEEE Trans. Mobile Computing*, Vol. 5, No. 7, Jul. 2006, 1-12.

K. Pawlikowski, H. D. J. Jeong, and J. S. R. Lee (2002), On credibility of simulation studies of telecommunication networks, *IEEE Commun. Mag.*, Vol. 40, No. 1, Jan. 2002, 132-139.

D. G. Michelson, J. Chuang (2006), Requirements for Standard Radiowave Progpagation Models for Vehicular Environments, in *IEEE 63rd Vehicular Technology Conference*, Vol. 6, pp. 2777-2781, Sep. 2006.

Network Simulator - ns-2, http://www.isi.edu/nsnam/ns/.

OPNET, http://www.opnet.com.

A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri (2006), Towards realistic mobility models for mobile ad hoc networks, in *Proceedings of ACM MobiCom, San Diego, CA*, pp. 217ÍC229, September 2006.

A. P. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri (2005), Real-world environment models for mobile network evaluation, *IEEE J. Sel. Areas in Commun.*, Vol. 23, No. 3, Mar. 2005, 622-632.

A. Mahajan, N. Potnis, K. Gopalan and A. Wang (2007), Modeling Vanet deployment in urban settings, in *International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*, Crete Island, Greece, pp. 151-158, Oct. 2007.

I. Stepanoy, K. Rothermel (2008), On the impact of a more realistic physical layer on MANET simulations results, *Ad Hoc Networks*, Vol. 6, No. 1, Jun. 2008, 61-78.

G. D. Durgin (2003). *Space-Time Wireless Channels*, 1st ed.Pearson Education, Inc, 2003.