VOIP AND BEST EFFORT SERVICE ENHANCEMENT ON FIXED WIMAX

A thesis submitted in partial fulfilment of the

requirements for the Degree

of Doctor of Philosophy in Electrical and Computer Engineering

in the University of Canterbury

by B. E. Shehan Perera

University of Canterbury

2008

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Professor Harsha Sirisena, of the Department of Electrical & Computer Engineering, University of Canterbury. His wide knowledge and logical way of thinking have been of great value for me. His understanding, encouragement and personal guidance have provided a good basis for the present thesis. I am also deeply grateful to my co-supervisor, Professor Krzysztof Pawlikowski, of the Department of Computer Science, University of Canterbury, for his support whenever it was needed. Sincere thanks also go out to Dr. Kishore Mehrotra of Tait Electronics for his readiness to provide assistance.

During this work I have collaborated with many colleagues for whom I have great regard, and I wish to extend my warmest thanks to all those who have inspired me in my work in the Department of Electrical & Computer Engineering, at the University of Canterbury. This includes Aiyathurai Jayanandan, Aun Haider, Malik Peiris, Priyan de Alwis and Thilan Gunawardena.

I owe my loving thanks to my wife Lan Nguyen, my father Earl Perera and my sister Menik Delpachitra. Without their encouragement and understanding it would have been impossible for me to finish this work.

The financial support of the University of Canterbury is gratefully acknowledged. I also wish to thank Medialab for the support provided through the NGN project, and the Royal Society of New Zealand for their travel grant.

Abstract

Fixed Broadband Wireless Access (BWA) for the last mile is a promising technology which can offer high speed voice, video and data service and fill the technology gap between Wireless LANs and wide area networks. This is seen as a challenging competitor to conventional wired last mile access systems like DSL and cable, even in areas where those technologies are already available. More importantly the technology can provide a cost-effective broadband access solution in rural areas beyond the reach of DSL or cable and in developing countries with little or no wired last mile infrastructure. Earlier BWA systems were based on proprietary technologies which made them costly and impossible to interoperate. The IEEE 802.16 set of standards was developed to level the playing field. An industry group the WiMAX Forum, was established to promote interoperability and compliance to this standard. This thesis gives an overview of the IEEE 802.16 WirelessMAN OFDM standard which is the basis for Fixed WiMAX. An in depth description of the medium access control (MAC) layer is provided and functionality of its components explained.

We have concentrated our effort on enhancing the performance of Fixed WiMAX for VoIP services, and best effort traffic which includes e-mail, web browsing, peer-to-peer traffic etc. The MAC layer defines four native service classes for differentiated QoS levels from the onset. The unsolicited grant service (UGS) class is designed to support real-time data streams consisting of fixed-size data packets issued at periodic intervals, such as T1/E1 and Voice over IP without silence suppression, while the non-real-time polling service (nrtPS) and best effort (BE) are meant for lower priority traffic. QoS and efficiency are at opposite ends of the scale in most cases, which makes it important to identify the trade-off between these two performance measures of a system. We have analyzed the effect the packetization interval of a UGS based VoIP stream has on system performance. The UGS service class has been modified so that the optimal packetization interval for VoIP can be dynamically selected based on PHY OFDM characteristics. This involves cross layer communication between the PHY, MAC and the Application Layer and selection of packetization intervals which keep the flow within packet loss and latency bounds

while increasing efficiency. A low latency retransmission scheme and a new ARQ feedback scheme for UGS have also been introduced. The goal being to guarantee QoS while increasing system efficiency. BE traffic when serviced by contention based access is variable in speed and latency, and low in efficiency. A detailed analysis of the contention based access scheme is done using Markov chains. This leads to optimization of system parameters to increase utilization and reduce overheads, while taking into account TCP as the most common transport layer protocol. nrtPS is considered as a replacement for contention based access. Several enhancements have been proposed to increase efficiency and facilitate better connection management. The effects of proposed changes are validated using analytical models in Matlab and verified using simulations. A simulation model was specifically created for IEEE 802.16 WirelessMAN OFDM in the QualNet simulation package. In essence the aim of this work was, to develop means to support a maximum number of users, with the required level of service, using the limited wireless resource.

Glossary

3GPP : 3rd Generation Partnership Project
3GPP2 : 3rd Generation Partnership Project 2
AMC : Adaptive modulation and coding
AMPS : Advanced Mobile Phone System
ARQ : Automatic Retransmission reQuest
BE : best effort
BS : Base Station
BSN : Block Sequence Number
BWA : Broadband Wireless Access
BW-REQ : Bandwidth Request
CDMA : Code Division Multiple Access
CPDP : Cellular Digital Packet Data
CRC : Cyclic Redundancy Check
CS : Convergence Sublayer
CSD : Circuit Switched Data
DCD : Downlink Channel Descriptor
DECT : Digital Enhanced Cordless Telecommunications
DIUC : Downlink Interface Usage Code
DL : Downlink
DL-MAP : Downlink Map
DSL : Digital Subscriber Loop
DVB : Digital Video Broadcasting
E-DCH : Enhanced Dedicated Channel
EDGE : Enhanced Data Rules for Global Evolution
ETSI : European Telecommunications Standards Institute
EV-DO : Evolution Data Optimized
FFT : Fast Fourier Transform
FMC : Fixed/Mobile Convergence
GMSK : Gaussian Minimum Shift Keying
GPRS : General Packet Radio Service
GSM : Global System for Mobile Communications
HSCSD : High Speed CSD
HSDPA : High Speed Downlink Packet Access
HS-DSCH : High-Speed Downlink Shared Channel
HSOPA : High Speed OFDM Packet Access
HSPA+ : High Speed Packet Access Plus
HSUPA : High Speed Uplink Packet Access
IE : Information Element
LMDS : local multipoint distribution systems
LOS : Line of sight
LTE : Long-Term Evolution
MIMO : Multiple Input Multiple Output
MMDS : multichannel multipoint distribution services

MS : Mobile Station
NLOS : Non Line of Sight
nrtPS : non-real-time polling service
OFDM : Orthogonal Frequency Division Multiplexing
OH : Overhead
OSI : Open Systems Interconnect
PDU : Protocol Data Unit
PHS : Personal Handy-phone System
PMP : point-to-multipoint
POTS : Plain Old Telephone Service
PSTN : Public Switched Telephone Network
QoS : Quality of Service
QPSK : Quadrature Phased Shift Keying
RLP : Radio Link Protocol
rtPS : real-time polling service
RTT : Radio Transmission Technology
SDU : Service Data Unit
SS : Subscriber Station
TDMA : Time Division Multiple Access
TTI : Transmission Time Interval
UCD : Uplink Channel Descriptor
UGS : Unsolicited Grant Service
UIUC : Uplink Interface Usage Code
UL : Uplink
UL-MAP : Uplink Map
UMTS : Universal Mobile Telecommunications System
UTRAN : UMTS Terrestrial Radio Access Network
VCC : Voice Call Continuity
WCDMA : Wideband CDMA
Wi-Fi : Wireless Fidelity
WiMAX : Worldwide Interoperability for Microwave Access
WLAN : Wireless Local Area Networks
WLL : Wireless Local Loop

Chapter 1

Introduction

Wireless communications have become increasingly popular in today's fast paced world. Instant access to virtually unlimited information has become the mantra of businesses and individuals alike. The evolution of wireless communications has been incredibly quick and the future of this technology is hard to predict. The impact of this technology on our lives will be tremendous and allow us to do things we never imagined a decade ago. Having access to information and being able to communicate easily and securely in any medium such as image, data, voice, video and multimedia in a cost effective manner is a requirement of modern technology savvy society. The conventional telephone network better known as the Public Switched Telephone Network (PSTN) has been in existence for a considerable time. Most consumer data communication is based on utilizing the fixed, voice centric PSTN as the communication link. While this strategy has proved to be very advantageous when considering access methods such as Digital Subscriber Loop (DSL) the prerequisites of having copper wire to the user's doorstep and being within a minimum distance from the local exchange cannot always be satisfied. Some communities even in developed countries cannot meet these demands and need to rely on data communication solutions provided by cellular network operators instead.

As such the need for wireless wide area coverage is on the increase now and will be in the foreseeable future. This demand for data based services has given rise to

many data centric wireless access technologies which can also transport voice as opposed to previous systems which were voice centric and later adapted to packet data. The development of wireless access technology through many generations to its current state is discussed in this chapter. IEEE 802.16 – 2004 based Metropolitan Area Networks are the topic of this thesis, and as such we present an overview of the said technology, as well as other broadband access schemes currently in use. We also highlight key business drivers which push this technology forward, and challenges which need to be overcome for its continuing success.

1.1 Cellular Technologies

We present some historical details about the progression from the earliest cellular communications systems which could also be used to transfer data at low speeds, to the current state of the art. It is clear, that initially, data transfer seemed to be an after thought while voice was the main component. As data transfer requirement keep growing steadily, current systems are specifically geared to transport many classes of data where voice, is one of the many services provided. However the fact remains, that these cellular mobile systems are a voice based architecture modified to include data services.

1.1.1 First Generation Systems

First-generation wireless telephone technology, are the analogue cell phone standards that were introduced in the 1980s (Stallings 2005) and continued until being replaced by 2G digital cell phones. It is worth mentioning these systems as some form of wireless data communication was possible albeit at sub-broadband speeds by today’s standards.

Cellular Digital Packet Data (CDPD) uses unused bandwidth normally used by Advanced Mobile Phone System (AMPS) mobile phones between 800 and 900 MHz to transfer data. Speeds up to 19.2 kbit/s are possible. Developed in the early 1990's, CDPD was large on the horizon as a future technology. However, it had difficulty competing against existing slower but less expensive Mobitex and DataTac systems, and never quite gained widespread acceptance before newer, faster standards such as General Packet Radio Service (GPRS) became dominant.

1.1.2 Second Generation Systems

Commonly known as 2G, these use digital signals where as the previous systems retrospectively dubbed 1G, used analogue. However both systems use digital signalling to connect the radio towers (which listen to the handsets) to the rest of the telephone system. 2G systems can be broadly divided into Time Division Multiple Access (TDMA) based and Code Division Multiple Access (CDMA) based standards depending on the multiple access technology used.

1.1.2.1 Global System for Mobile Communications

The Global System for Mobile Communications, GSM (original acronym: Groupe Spécial Mobile) is the most popular standard for mobile phones in the world. GSM service is used by over 2.3 billion people across more than 220 countries and territories (GSM Association 2008). The ubiquity of the GSM standard makes international roaming very common between mobile phone operators, enabling subscribers to use their phones in many parts of the world. GSM differs significantly from its predecessors in that both signalling and speech channels are digital, which means that it is considered a second generation (2G) mobile phone system. The 3rd Generation Partnership Project (3GPP) was able to add advanced data transmission upgrades such as GPRS and EDGE using this “all digital” architecture.

GSM has gone through a few steps in its progression towards a true Third Generation access method. Namely Circuit Switched Data (CSD), High Speed CSD (HSCSD) and GPRS. These data centric “upgrades” are described in detail in section 1.1.3 - Third Generation Systems.

1.1.2.2 Digital – Advanced Mobile Phone System (D-AMPS)

IS-54 and IS-136 are 2G mobile phone systems, known as Digital AMPS (D-AMPS). It is used throughout the Americas, particularly in the United States and Canada. D-AMPS is considered end-of-life, and existing networks are in the process of being replaced by GSM/GPRS and CDMA2000 technologies.

Although this system is most often referred to as TDMA, a common multiple access technique which is used by multiple protocols, including GSM, IS-54 and IS-136. The two different uses of this term can be confusing. TDMA (the technique) is also used in the GSM standard. However, TDMA (the standard, i.e. IS-136) has been

competing against GSM and systems based on CDMA for adoption by the network carriers, although this standard is now being phased out in favour of GSM technology. This technique allows a bit rate of 48.6 kbps with 30 kHz channel spacing, to give a bandwidth efficiency of 1.62 b/s/Hz. This value is 20% better than GSM.

1.1.2.3 Interim Standard 95 (IS-95)

Interim Standard 95 (IS-95), is the first CDMA-based digital cellular standard pioneered by Qualcomm. The brand name for IS-95 is “cdmaOne”. IS-95 is also known as TIA-EIA-95.

Since voice and user data are intermittent, the traffic channels support variable-rate operation. Every 20 ms frame may be transmitted at a different rate, as determined by the service in use (voice or data). For voice calls, the traffic channel carries frames of vocoder data. The mobile receiving a variable-rate traffic frame does not know the rate at which the frame was transmitted. Typically, the frame is decoded at each possible rate, and using the quality metrics of the Viterbi decoder, the correct result is chosen.

Traffic channels may also carry circuit-switched data calls in IS-95. The variable-rate traffic frames are generated using the IS-95 Radio Link Protocol (RLP). RLP provides a mechanism to improve the performance of the wireless link for data. Where voice calls might tolerate the dropping of occasional 20 ms frames, a data call would have unacceptable performance without RLP.

Under IS-95B revision 5, it was possible for a user to use up to seven supplemental "code" (traffic) channels simultaneously to increase the throughput of a data call. Very few mobiles or networks ever provided this feature, which could in theory offer 115.2 Kbps to a user.

1.1.2.4 Personal Digital Cellular (PDC)

Personal Digital Cellular is a 2G mobile phone standard developed and used exclusively in Japan. Like D-AMPS and GSM, PDC uses TDMA. PDC uses a 25 kHz bandwidth, 3 time slots, pi/4-DQPSK modulation and low bit-rate 11.2 kbps and 5.6 kbps (half-rate) voice codecs. PDC is implemented in the 800 MHz and 1.5 GHz bands.

The services include voice (full and half-rate), supplementary services (call waiting, voice mail, three-way calling, call forwarding, and so on), data service (up to 9.6 kbps CSD), and packet-switched wireless data (up to 28.8 kbps PDC-P). Compared to GSM, PDC's weak broadcast strength allows small, portable phones with light batteries at the expense of substandard voice quality and problems maintaining the connection, particularly in enclosed spaces like elevators.

After a peak of nearly 80 million subscriber to PDC, and 45 million subscribers by the end of 2005, it is being phased out in favour of 3G technologies like W-CDMA and CDMA2000.

1.1.2.5 Personal Handy-phone System (PHS)

PHS is, essentially, a cordless telephone like Digital Enhanced Cordless Telecommunications (DECT) (Stallings 2005), with the capability to handover from one cell to another. PHS cells are small, with transmission power of base station a maximum of 500 mW and range typically measures in tens or at most hundreds of meters (some of brand-new base stations can range about 2 kilometres at line-of-sight), as opposed to the multi-kilometre ranges of GSM. This makes PHS suitable for dense urban areas, but impractical for rural areas, and the small cell size also makes it difficult if not impossible to make calls from rapidly moving vehicles

PHS uses TDMA/TDD for its radio channel access method. Modern PHS phones can support many value-added services such as high speed wireless data/Internet connection (64 kbps and higher), e-mailing, text messaging and even colour image transfer.

PHS technology is also a popular option for providing a wireless local loop, where it is used to bridge the "last mile" gap between the Plain Old Telephone Service (POTS) network and the subscriber's home. It was developed under the concept that it makes up a wireless front-end of ISDN network. So a base station of PHS has a compatibility with, and is often connected directly to ISDN telephone exchange equipment (digital switch).

PHS has many advantages over 3G cellular phone systems such as its low-price base station, micro-cellular system and 'Dynamic Cell Assign' system which can

afford more number-of-digits frequency use efficiency with lower cost compared with typical 3G cellular telephone systems. It makes possible the flat-rate wireless service such as AIR-EDGE all over Japan.

The speed of AIR-EDGE data connection is accelerated by combining lines, each of which basically is 32 kbps. AIR-EDGE 1x or first version introduced in 2001 provide only 32 kbps service. In 2002, 128 kbps service (AIR-EDGE 4x) started. In 2005, 256 kbps (AIR-EDGE 8x) service started. Furthermore, in 2006, the speed of each line was also upgraded to 1.6 times. Using the latest equipment AIR-EDGE 8x can achieve speeds up to 402 kbps which exceeds the speeds of popular W-CDMA based 3G.

1.1.3 Third Generation Systems

3G (or 3-G) is short for third-generation technology. It is used in the context of mobile phone standards. The services associated with 3G provide the ability to transfer simultaneously both voice data (a telephone call) and non-voice data (such as downloading information, exchanging email, and instant messaging). In marketing 3G services, video telephony has often been used as the killer application for 3G although uptake has been low due to practicality and cost.

1.1.3.1 Evolution of GSM and the 3GPP

The 3GPP collaborative agreement was formed in December 1998 to facilitate cooperation between the previously disparate standards groups in Europe, U.S., Japan and Korea. The 3GPP standards were based on the original European Telecommunications Standards Institute (ETSI) GSM specification, dominant in Asia Pacific and Europe. GSM accounts for nearly 65% of global mobile phone subscribers. GSM-based data services took their roots as CSD in the early GSM specifications dating back to 1991 and 1992, where a single Gaussian Minimum Shift Keying (GMSK) modulated GSM timeslot was entirely consumed for data transmission. This provided users with a 9.6 kbps data service. Higher-speed service could be delivered via HSCSD, which simply tied together four GSM timeslots, allowing users about 50 kbps and consuming four radio timeslots per connection MS. This is depicted in the GSM evolution shown in Figure 1-1.

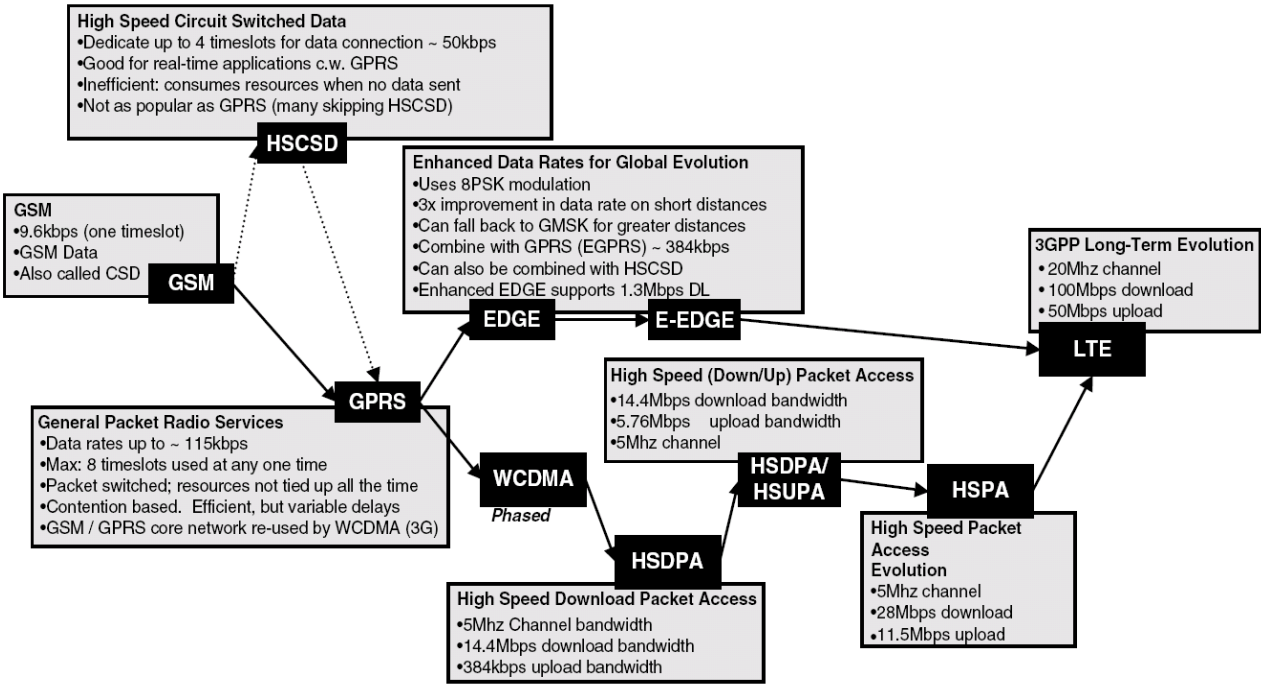

Figure 1-1 ETSI evolution of GSM towards WCDMA

Connection-oriented CSD services generally are not the most efficient way to use a scarce resource (such as the RF constrained radio interface) to transmit IP packet-based data. They occupy an entire GSM radio timeslot, even when no data or small amounts of data are being transferred. Many IP applications are “bursty” in nature and can gain large efficiencies through statistical multiplexing, which can be accomplished through packet switching. With packet-switched communications, the network delivers a data packet only when the need arises, so the radio resource can be multiplexed between many users.

GPRS gives GSM users direct IP network access while achieving significantly higher spectral efficiency than previously available circuit-switched data services. GPRS affords individual users data rates over 100 kbps and does so, using a statistically multiplexed packet-switched radio interface that consumes a maximum of eight timeslots shared by multiple end users. Although higher bandwidth per end user can be provided via GPRS, contention and congestion on the radio interface cause higher and sometimes variable delays. This means that although GPRS is useful for many burst-tolerant” applications (e-mail, file transfers, web browsing), “interactive” types of IP applications (VoIP, streaming video, Push to Talk) can be less forgiving on a GPRS connection.

The next step in the GSM-based evolution was Enhanced Data Rules for Global Evolution (EDGE). EDGE modified the radio link modulation scheme from GMSK to a phased-shift keying modulation scheme, 8 Quadrature Phased Shift Keying (8QPSK). This high-order PSK modulation allows multiple bits to be encoded per symbol transmitted while also minimizing bit error rates. EDGE in combination with GPRS, known as E-GPRS, can deliver single user data rates over 300 kbps.

The 3GPP standards are specified in "releases", which are used to describe both the baseline network architecture, and its evolution. This architecture and evolution include the wireless "air interfaces" as well as the terrestrial access network and support systems, providing a path for operators and vendors to evolve their networks toward Universal Mobile Telecommunications System (UMTS). One of the benefits of this approach is that, it is accomplished through incremental changes that minimize the disruptiveness and costs associated with the evolution. Releases 5, 6 and 7 introduced High Speed Downlink Packet Access (HSDPA), High Speed Uplink Packet Access (HSUPA), better Quality of Service (QoS) (lower latency and jitter), Voice Call Continuity (VCC), additional Fixed/Mobile Convergence (FMC) features, and High Speed Packet Access (HSPA+).

Many carriers are already deploying the next phase of 3GPP radio interfaces: HSDPA. HSDPA allows for download user data rates of up to 14.4 Mbps using a 5MHz-wide channel (Chris Hellberg, Dylan Greene et al. 2007).

1.1.3.2 HSDPA/HSUPA

HSDPA works by establishing a shared high-speed downlink channel, known as the High-Speed Downlink Shared Channel (HS-DSCH), which is shared between multiple users. Using fast scheduling techniques, the HS-DSCH is divided into up to 15 parallel channels, with a very short "duty cycle," or Transmission Time Interval (TTI), of 2 ms each. These channels can be simultaneously used by a single terminal for an entire TTI or split between multiple users as needed. Due to the fast scheduling techniques and short 2ms cycle, the system can quickly allocate bandwidth where it's needed. This scheduling technique is comparable to that used in WLAN networks.

HSDPA also supports a fast Automatic Retransmission reQuest (ARQ) mechanism, allowing corrupted datagrams to be retransmitted inside a 10ms window. This helps facilitate higher TCP throughput by creating a more reliable transport

layer. Using these techniques in conjunction with new enhanced modulation and code schemes, HSDPA can offer theoretical speeds up to 14.4Mbps. It may be possible to increase this throughput in the future through further technology enhancements such as the adoption of multiple antenna systems such as Multiple Input Multiple Output (MIMO), as found in Release 6.

In 3GPP Release 6, the HSUPA uplink air interface, also called Enhanced Dedicated Channel (E-DCH), is redefined. HSUPA can support speeds ranging from about 1Mbps to 14.4Mbps. HSUPA uses a dedicated channel, unlike the shared channel found in HSDPA. HSUPA achieves its higher bandwidth through fast radio (Node B) scheduling, ARQ, and new signalling and data channels.

Looking at Release 7, 3GPP is architecting a system to provide higher bandwidth and lower delay air interfaces. One goal is to be able to handle the same or greater voice capacity as previous circuit switched network. Some of the Release 7 goals include enhancing VoIP quality, reducing delay to the 20 to 40 ms range, increasing peak bandwidth rates to the 40 to 50 Mbps range, and other enhancements to enable more mobile broadband services. This is commonly referred to as HSPA+.

1.1.3.3 Evolution of CDMA and the 3GPP2

The 3GPP2 collaboration dates back to 1998 during early IMT-2000 discussions. It was formed to define third-generation (3G) specifications and standards for non-GSM-based mobile telecommunications systems that were based on CDMA.

1x Radio Transmission Technology (1xRTT, or 1x for short) was the first generation of the CDMA2000 air interface to reach wide deployment. It occupies a single pair of 1.25MHz radio channels. The first 1xRTT revision offered more traffic channels than IS-95, which resulted in more voice and data capacity. 1xRTT Revision 0 was typically deployed for 144kbps per user, although it could support higher rates. 1xRTT can also coexist with IS-95, allowing for easy deployment and migration because it uses the same core network as IS-95 and the air interfaces do not interfere. With the next release of 1xRTT, Revision A, more than 300 kbps could be delivered across a single 1.25MHz channel. These channels could also be combined, allowing mobile operators to offer 3xRTT services (three 1.25MHz radio channels).

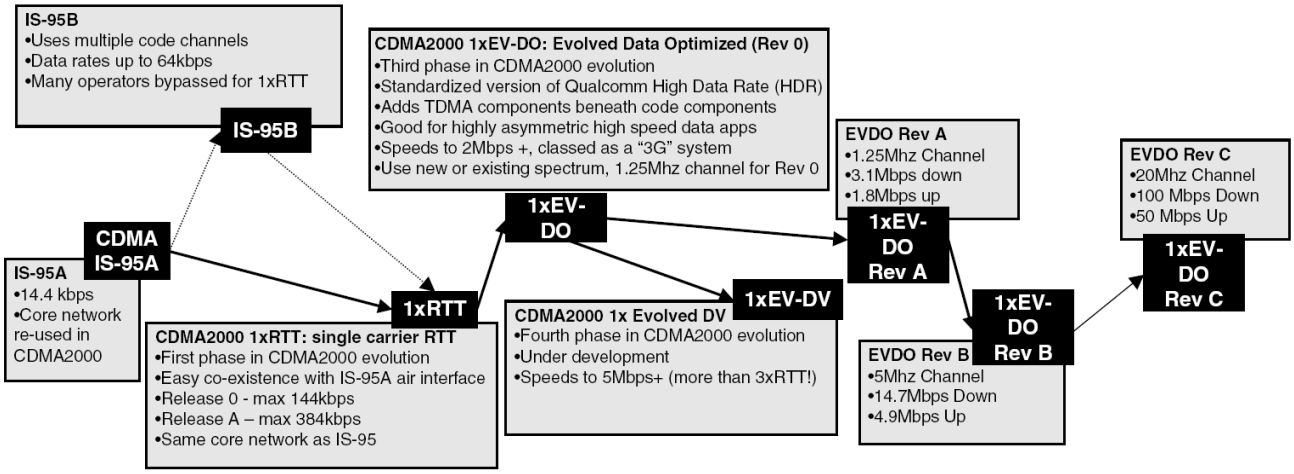

Figure 1-2 Evolution of IS-95A towards EV-DO and beyond.

The next step in the evolution of CDMA2000 to be deployed is known as Evolution Data Optimized CDMA, or 1xEVDO. EVDO is an evolution of CDMA2000 1xRTT, with high data rate (HDR) capabilities added and a TDMA component added below the code division layer. The EVDO air interface is documented in IS-856. Revision 0 EVDO supports data rates of up to 2.5Mbps toward the user and about 154kbps up. Revision 0 EVDO is generally classified as a 3G technology. Revision A, which was being deployed in North America in 2006, can support up to 3.1Mbps downlink and 1.8 Mbps in the opposite direction over the same 1.25 MHz channels. In the future, EVDO Revisions B and C will expand the radio channel bandwidth to 5MHz and increase user data rates to 14.7 Mbps down/4.7 Mbps up and 100 Mbps down/50 Mbps up, respectively. These channel bandwidth and user data rates are comparable to 3GPP's Long-Term Evolution (LTE). Figure 1-2 shows the CDMA evolution.

1.1.3.4 Beyond CDMA

With the widespread subscriber adoption of higher bandwidth services, packet based mobile voice, rich media, gaming, video, and other advanced services, operators are setting long-term capacity targets that far exceed today's air interface capabilities. To reach these long-term capacity goals, spectrum allocation will have to evolve from today's standards, which include CDMA's 1.25MHz channels and WCDMA's 5MHz channels. Future standards are aiming at wider band channels up to 20MHz to provide higher transmission rates to end users.

Global spectrum allocation plays a major role in how and which technologies are developed by operators, vendors, and standards groups. Operators will require large amounts of spectrum in the right bands to cost-effectively deliver higher speed mobile services. It's important to note a major trade-off that exists in all radio networks: higher frequencies allow for more capacity but have shorter ranges for a given power output. This translates directly into costs for the operator, because higher-frequency services typically require more cell sites, backhaul network, and equipment to deploy. Most wide-area mobile radio networks are designed for use in the 800MHz to 2600MHz range, which is generally considered to offer the optimal trade-off between speed and coverage.

As a result of this demand for higher bandwidth, optimized coverage, and optimal spectral efficiency, most radio interface standards are moving away from CDMA based carriers to Orthogonal Frequency Division Multiplexing (OFDM) based schemes. OFDM is generally perceived to be more spectrally efficient than CDMA, is less susceptible to interference, and offers very efficient granular bandwidth to terminals and advanced scheduling algorithms. This gives OFDM based modulation schemes more control over quality of service, as well as higher bandwidth. OFDM modulation schemes have already been adopted in mobile broadcasting systems such as Qualcomm's FLO and DVB-H. OFDM and OFDMA based modulation schemes are the selected schemes for WiMAX and other technologies which are considered 4G.

This can be seen in 3GPP's 3G LTE UTRAN work, which is the planned successor to WCDMA and HSPA, targeted for 3GPP Release 8. Referred to as, High Speed OFDM Packet Access (HSOPA), is an entirely new air interface system, unrelated to and incompatible with W-CDMA. On the 3GPP2 side, 20MHz channel OFDM is being evaluated for CDMA2000 EVDO Revision C. The two systems' services become married at this point; 3GPP LTE and EVDO RevC offer equivalent data rates and a converging service delivery platform. OFDM is already a very mature and widely deployed technology, partially thanks to its use in the 802.11 WLAN protocol.

1.1.3.5 Beyond 3G

Technologies which support high data rates in the upstream as well as downstream directions, provide several classes of QoS, and have an Internet Protocol (all IP) based architecture are generally defined as Beyond 3G or Fourth Generation (4G). IEEE 802.16, which is introduced in section 1.4, is one such technology. Some may consider HSDPA/HSUPA as being 4G technology but they lack a defining characteristic of having an all-IP architecture.

1.2 Wireless LANs and Wi-Fi

Synonymous with IEEE 802.11 – Wireless LAN standards, Wi-Fi is a trademark of the Wi-Fi Alliance which promotes conformance to the IEEE standards. Wi-Fi has dominated as the most popular access technology in the last 30 meters within the home or office since 2005 as prices have plummeted due to the economies of scale brought about by standardization. In homes and offices, Wi-Fi allows untethered connectivity to the network at moderate distances with LAN like speeds. While the residential market rapidly embraced the use of Wi-Fi technology, it has been slower to move into enterprise networks due to concerns over security.

During recent times, the hotspot market has emerged in public locations world-wide. Wi-Fi adapters being built in to practically all current laptops and PDAs give users connectivity to the internet even when away from home or the workplace. In short, Wi-Fi has been an overwhelming success because it is interoperable, easy to use, and cheap. When Wi-Fi technologies are used in a MAN, three factors must be kept in mind: Range, QoS and Security. Standard Wi-Fi technology is limited to a 100 m range in a LOS environment. The range delivered is significantly reduced if there are obstacles which impede radio propagation. To overcome this when building an urban canopy coverage model, service providers need to build a significant number of wireless Points of Presence (POPs) with a transport network (backhaul) using either wireline or wireless technologies delivering the bandwidth to each access point. Because of limited range with standards-based solutions many wireless ISPs use proprietary outdoor wireless solutions that allow for a greater coverage losing the desirable benefit of Wi-Fi's low cost and interoperability (Suitor 2004).

Until recently Wi-Fi did not support any sort of QoS. With the amendment IEEE 802.11e (IEEE Computer Society 2005) changes to the MAC layer introduced some much needed QoS functionality. This is considered critical for delay sensitive services such as voice over wireless IP and streaming multimedia. In pre 802.11e Wi-Fi there were two coordination functions defined, Distributed Coordination Function (DCF) and Point Coordination Function (PCF). PCF included a contention free polling period (CFP) where the AP polled the SS in a round robin fashion which guaranteed access. However PCF was hardly used. The new standard introduces Enhanced Distributed Channel Access (EDCA) and Hybrid Coordination Function (HCF) Controlled Channel Access (HCCA) which may be seen as enhanced versions of DCF and PCF (Ramos, Panigrahi et al. 2005). HCCA is not mandatory for access points (APs) and few manufacturers implement it in their products. Legacy 802.11a/b/g hardware is not compatible with the amendment which impacts a service provider's ability to deliver different grades of service to individual business or residential customers.

The issue of limited range and data rate in Wi-Fi is being looked into by Task Group n (TGn) and WWiSE (Griffith 2006). There have been draft and pre-release versions of 802.11n which promise a maximum data rate of 274 Mbps using MIMO at both the AP and the subscriber station (SS) with a range of under 250 m. Many vendors have already marketed products based on the draft versions 1.0 and 2.0 in anticipation of ratification of the standard. The aim of these endeavours is to obtain actual throughput around 100 Mpbs. Wi-Fi has been proven to be inherently inefficient in terms throughput (H. S. Chhaya 1997; Bianchi 2000; Bianchi and Tinnirello 2005; Ching-Ling and Wanjiun 2007) and have scalability issues as shown by (Chuan Heng and Zukerman 2001; Zhen-ning, Tsang et al. 2004; Maaroufi, Ajib et al. 2007) and a host of work in the last few years.

Wireless broadcast networks require strong security. This is paramount not only in the enterprise market but also in the home environment where several APs may be within range to a given residence. The Wi-Fi Alliance reacted to shortcomings in WEP (Wired Equivalency Privacy) by developing WPA (Wi-Fi Protected Access) and more recently WPA2. WPA2, a subset of IEEE 802.11i (IEEE Computer Society 2004), is considered completely secure and is mandatory in order to obtain Wi-Fi certification since 2006.

Through much hard work by the IEEE, its Task Groups and the Wi-Fi Alliance, wireless LANs have cemented their place in today's wireless communication environment. However Wi-Fi remains limited to the LAN space in terms of range, with a dependence on WAN technology to bridge the last mile for access to the internet and external connectivity.

1.3 Wireless Local Loop

In traditional communications the provision of voice and other data services to the end user over the local loop or subscriber loop has been done through wired systems. For residential subscribers, twisted pair has been and still is the primary means of delivery given that landline telephone services exist. As demands increased ways and means of squeezing more out of the copper were developed in the form of ISDN, DSL and xDSL (Stallings 2005). As an alternative to this, narrowband (voice only) and broadband (voice and data) wireless local loop (WLL) or fixed wireless access technologies have become popular. One advantages of WLL is the reduced installation cost. The switching equipment may be more complex and expensive but this is offset by the savings in cable costs and maintenance. Another advantage is the rapid installation. After suitable sites for base stations (BSs) are found and the necessary licenses are obtained adding users to existing infrastructure is relatively simple.

1.3.1 First-Generation Broadband Systems

Systems which were competitive in terms of data rate, were developed for higher frequencies, such as the 2.5GHz and 3.5GHz bands. Very high speed systems, called local multipoint distribution systems (LMDS), supporting up to several hundreds of megabits per second, were also developed in millimetre wave frequency bands, such as the 24GHz and 39GHz bands. LMDS based services were targeted at business users and in the late 1990s enjoyed rapid but short-lived success. Problems obtaining access to rooftops for installing antennas, coupled with its shorter range capabilities, squashed its growth.

In the late 1990s, one of the more important deployments of wireless broadband happened in the multichannel multipoint distribution services (MMDS) band at 2.5GHz. The MMDS band was historically used to provide wireless cable

broadcast video services, especially in rural areas where cable TV services were not available. The advent of satellite TV ruined the wireless cable business, and operators were looking for alternative ways to use this spectrum. A few operators began to offer one-way wireless Internet-access service, using telephone line as the return path. Later on due to relaxation of rules, companies started developing high-speed, bi-directional, fixed wireless solutions for this band.

The first generation of these fixed broadband wireless solutions were deployed using the same towers that served wireless cable subscribers. These towers were typically a hundred meters tall and enabled LOS coverage to distances up to 55 kilometers, using high-power transmitters. First-generation MMDS systems required that subscribers install at their premises outdoor antennas high enough and pointed toward the tower for a clear LOS transmission path.

1.3.2 Second-Generation Broadband Systems

Second-generation broadband wireless systems were able to overcome the LOS issue and to provide more capacity. This was done through the use of a cellular architecture and implementation of advanced-signal processing techniques to improve the link and system performance under multipath conditions. Several start-up companies developed advanced proprietary solutions that provided significant performance gains over first-generation systems. Most of these new systems could perform well under non-line-of-sight conditions, with customer-premise antennas typically mounted under the eaves or lower. Many partially solved the NLOS problem by using such techniques as OFDM, CDMA, and multiantenna processing. Some systems, such as those developed by SOMA Networks and Navini Networks, demonstrated satisfactory link performance over a few miles to desktop subscriber terminals without the need for an antenna mounted outside. A few megabits per second throughput over cell ranges of a few miles had become possible with second generation fixed wireless broadband systems.

1.3.3 Future of Wireless Local Loop

Current cellular technology is too expensive and does not provide enough always-on bandwidth to act as a realistic alternative to WLL. Cellular data rates are very dependant on network loading and tariffs for data are much higher than in WLL. A

major advantage over cellular is the ability to use directional antenna which provides improved signal strength in both directions. In response to the need and interest in LMDS WLL, the IEEE through their 802.16 working group standardized the air interface and related functions associated with LMDS as IEEE 802.16. Since then, an industry group, the WiMAX Forum (WiMAX Forum), has been formed to promote interoperability between manufacturers and vendors and, conformance to the standard. An in-depth look at the IEEE 802.16 standard is given in the following chapter.

1.4 Introduction to IEEE 802.16 and WiMAX

The IEEE 802.16 group was formed in 1998 to develop an air-interface standard for wireless broadband. The group's initial focus was the development of a LOS-based point-to-multipoint wireless broadband system for operation in the 10GHz–66GHz millimetre wave band. The resulting standard—the original 802.16 standard, completed in December 2001—was based on a single-carrier physical (PHY) layer with a burst time division multiplexed (TDM) MAC layer. Many of the concepts related to the MAC layer were adapted for wireless from the popular cable modem DOCSIS (data over cable service interface specification) standard.

The IEEE 802.16 group subsequently produced 802.16a, an amendment to the standard, to include NLOS applications in the 2GHz–11GHz band, using an OFDM based physical layer. Additions to the MAC layer, such as support for OFDMA, were also included. Further revisions resulted in a new standard in 2004, called IEEE 802.16-2004, which replaced all prior versions and formed the basis for the first WiMAX solution. The WiMAX solutions based on IEEE 802.16-2004 targeted fixed applications, and are referred to as fixed WiMAX. In December 2005, the IEEE group completed and approved IFEEE 802.16e-2005, an amendment to the IEEE 802.16-2004 standard that added mobility support. The IEEE 802.16e-2005 forms the basis for the WiMAX solution for nomadic and mobile applications and is often referred to as mobile WiMAX.

A system profile defines the subset of mandatory and optional physical- and MAC-layer features selected by the WiMAX Forum from the IEEE 802.16-2004 or IEEE 802.16e-2005 standard. Currently, the WiMAX Forum has two different system

profiles: one based on IEEE 802.16-2004, OFDM PHY, called the fixed system profile; the other one based on IEEE 802.16e-2005 scalable OFDMA PHY, called the mobility system profile. The WiMAX Forum has thus far defined five fixed certification profiles and fourteen mobility certification profiles (Jeffrey G. Andrews, Arunabha Ghosh et al. 2007). To date, there are two fixed WiMAX profiles against which equipment have been certified. These are 3.5GHz systems operating over a 3.5MHz channel, using the fixed system profile based on the IEEE 802.16-2004 OFDM physical layer with a point-to-multipoint MAC. One of the profiles uses frequency division duplexing (FDD), and the other uses time division duplexing (TDD).

With the completion of the IEEE 802.16e-2005 standard, interest within the WiMAX group has shifted sharply toward developing and certifying mobile WiMAX system profiles based on this newer standard. This however was beyond the timeline of this thesis and is out of scope in this work. All mobile WiMAX profiles use scalable OFDMA as the physical layer. At least initially, all mobility profiles will use a point-to-multipoint MAC. It should also be noted that all the current candidate mobility certification profiles are TDD based (Carl Eklund, Roger B. Marks et al. 2007). Although TDD is often preferred, FDD profiles may be needed for in the future to comply with regulatory pairing requirements in certain bands.

1.4.1 Business Drivers

The business drivers or market forces which have an impact on the future of Fixed WiMAX are considered here. Applications using a fixed wireless solution can be classified as point-to-point or point-to-multipoint. Point-to-point applications include interbuilding connectivity and microwave backhaul. Point-to-multipoint applications include (1) broadband for residential, small office/home office (SOHO), and small- to medium-enterprise (SME) markets, (2) T1 or fractional T1-like services to businesses, and (3) wireless backhaul for Wi-Fi hotspots.

1.4.1.1 Consumer and small-business broadband

Clearly, one of the largest applications of WiMAX in the near future is likely to be broadband access for residential, SOHO, and SME markets. Broadband services provided using fixed WiMAX could include high-speed Internet access, telephony

services using voice over IP, and a host of other Internet-based applications. Fixed wireless offers several advantages over traditional wired solutions. These advantages include lower entry and deployment costs; faster and easier deployment and revenue realization; ability to build out the network as needed; lower operational costs for network maintenance, management, and operation; and independence from the incumbent carriers.

From a customer premise equipment (CPE) or subscriber station (SS) perspective, two types of deployment models can be used for fixed broadband services to the residential, SOHO, and SME markets. One model requires the installation of an outdoor antenna at the customer premise; the other uses an all-in-one integrated radio modem that the customer can install indoors like traditional DSL or cable modems. Using outdoor antennas improves the radio link and hence the performance of the system. This model allows for greater coverage area per base station, which reduces the density of base stations required to provide broadband coverage, thereby reducing capital expenditure. Requiring an outdoor antenna, however, means t hat installation will require a truck-roll with a trained professional and also implies a higher SS cost. Clearly, the two deployment scenarios show a trade-off between capital expenses and operating expense: between base station capital infrastructure costs and SS and installation costs. In developed countries, the high labour cost of truck-roll, coupled with consumer dislike for outdoor antennas, will likely favour an indoor SS deployment, at least for the residential application. Further, an indoor self-install SS will also allow a business model that can exploit the retail distribution channel and offer consumers a variety of SS choices. In developing countries, however, where labour is cheaper and aesthetic and zoning considerations are not so powerful, an outdoor-SS deployment model may make more economic sense.

In developed countries with good wired infrastructure, fixed wireless broadband is more likely to be used in rural or underserved areas, where traditional means of serving them is more expensive. Services to these areas may be provided by incumbent telephone companies or by smaller players, such as WISPs, or local communities and utilities. It is also possible that competitive service providers could use WiMAX to compete directly with DSL and cable modem providers in urban and suburban markets. In the United States, the FCC’s August 2005 decision to rollback

cable plant sharing needs is likely to increase the appeal of fixed wireless solutions to competitive providers as they look for alternative means to reach subscribers. The competitive landscape in the United States is such that traditional cable TV companies and telephone companies are competing to offer a full bundle of telecommunications and entertainment services to customers. In this environment, satellite TV companies may be pushed to offering broadband services including voice and data in order to stay competitive with the telephone and cable companies, and may look to WiMAX as a potential solution to achieve this.

1.4.1.2 T1 emulation for business

The other major opportunity for fixed WiMAX in developed markets is as a solution for competitive T1/E1, fractional T1/E1, or higher-speed services for the business market. Given that only a small fraction of commercial buildings worldwide have access to fibre, there is a clear need for alternative high-bandwidth solutions for enterprise customers. In the business market, there is demand for symmetrical T1/E1 services that cable and DSL have so far not met the technical requirements for. Traditional telco services continue to serve this demand with relatively little competition. Fixed broadband solutions using WiMAX could potentially compete in this market and trump landline solutions in terms of time to market, pricing, and dynamic provisioning of bandwidth.

1.4.1.3 Backhaul for Wi-Fi hotspots

Another interesting opportunity for WiMAX in the developed world is the potential to serve as the backhaul connection to the burgeoning Wi-Fi hotspots market. In the United States and other developed markets, a growing number of Wi-Fi hotspots are being deployed in public areas such as convention centres, hotels, airports, and coffee shops. The Wi-Fi hotspot deployments are expected to continue to grow in the coming years. Most Wi-Fi hotspot operators currently use wired broadband connections to connect the hotspots back to a network point of presence. WiMAX could serve as a faster and cheaper alternative to wired backhaul for these hotspots. Using the point-to-multipoint transmission capabilities of WiMAX to serve as backhaul links to hotspots could substantially improve the business case for Wi-Fi hotspots and provide further momentum for hotspot deployment. Similarly, WiMAX could serve as 3G (third-generation) cellular backhaul.

A potentially larger market for fixed broadband WiMAX exists outside the United States, particularly in urban and suburban locales in developing economies—China, India, Russia, Indonesia, Brazil and several other countries in Latin America, Eastern Europe, Asia, and Africa—that lack an installed base of wireline broadband networks. National governments that are eager to quickly catch up with developed countries without massive, expensive, and slow network rollouts could use WiMAX to leapfrog ahead. A number of these countries have seen sizable deployments of legacy WLL systems for voice and narrowband data. Vendors and carriers of these networks will find it easy to promote the value of WiMAX to support broadband data and voice in a fixed environment.

1.4.2 Market Challenges

Despite the marketing hype and the broad industry support for the development of WiMAX, its success is not a forgone conclusion. In fact, broadband wireless in general and WiMAX in particular face a number of challenges that could impede their adoption in the marketplace.

1.4.2.1 The rising bar of traditional broadband

In the fixed broadband application space, WiMAX will have to compete effectively with traditional wired alternatives, such as DSL and cable, to achieve widespread adoption in mature markets, such as the United States. DSL and cable modem technologies continue to evolve at a rapid pace, providing increasing data rate capabilities. For example, DSL services in the United States already offer 3Mbps–6Mbps of downstream throughput to the end user, and solutions based on the newer VDSL2 standard will soon deliver up to 50Mbps–100Mbps, depending on the loop length. With incumbent carriers pushing fibre deeper into the networks, the copper loop lengths are getting shorter, allowing for significantly improved data rates. Cable modem technologies offer even higher speeds than DSL. Even on the upstream, where bandwidth had been traditionally limited, data rates on the order of several megabits per second per user are becoming a reality in both DSL and cable. The extremely high data rates supported by these wired broadband solutions allow providers to offer not only data, voice, and multimedia applications but also entertainment TV, including HDTV. It will be extremely difficult for broadband wireless systems to match the

rising throughput performance of traditional broadband. WiMAX will have to rely on portability and mobility as differentiators as opposed to data rate. WiMAX may have an advantage in terms of network infrastructure cost, but DSL and cable benefit from the declining cost curves on their CPE, due to their mature-market state. Given these impediments, fixed WiMAX is more likely to be deployed in rural or underserved areas in countries with a mature broadband access market. In developing countries, where existing broadband infrastructure is weak, the business challenges for fixed WiMAX are less daunting, and hence it is much more likely to succeed.

1.4.2.2 Differences in global spectrum availability

There are considerable differences in the allocation and regulations of broadband spectrum worldwide. Although 2.5GHz, 3.5GHz, and 5.8GHz bands are allotted in many regions of the world, many growth markets require new allocations. Given the diverse requirements and regulatory philosophy of various national governments, it will be a challenge for the industry to achieve global harmonization. For WiMAX to be a global success like Wi-Fi, regulatory bodies need to allow full flexibility in terms of the services that can be offered in the various spectrum bands.

1.4.2.3 Competition from 3G

For mobile WiMAX, the most significant challenge comes from 3G technologies that are being deployed worldwide by mobile operators. Incumbent mobile operators are more likely to seek performance improvements through 3G evolution than to adopt WiMAX. New entrants and innovative challengers entering the mobile broadband market using WiMAX will have to face stiff competition from 3G operators and will have to find a way to differentiate themselves from 3G in a manner that is attractive to the users. They may have to develop innovative applications and business models to effectively compete against 3G.

1.4.2.4 Device development

For mobile WiMAX to be successful, it is important to have a wide variety of terminal devices. Embedding WiMAX chips into computers could be a good first step but may not be sufficient. Perhaps WiMAX can differentiate from 3G by approaching the market with innovative devices. Some examples could include WiMAX embedded into MP3 players, video players, or handheld PCs. Device-development efforts should also include multimode devices. A variety of broadband systems will

likely be deployed, and it is critical that diverse networks interoperate to make ubiquitous personal broadband services a reality. Ensuring that device development happens in parallel with network deployment will be a challenge.

1.4.3 Technical Challenges

So far, we have discussed the history, applications, and business challenges of broadband wireless. We now address the technical challenges of developing and deploying a successful broadband wireless system. To gain widespread success, broadband wireless systems must deliver multimegabit per second throughput to end users, with robust QoS to support a variety of services, such as voice, data, and multimedia. Given the remarkable success of the Internet and the large variety of emerging IP-based applications, it is critical that broadband wireless systems be built to support these IP-based applications and services efficiently. Fixed broadband systems must, ideally, deliver these services to indoor locations, using subscriber stations that can be easily self installed by the enduser. Mobile broadband systems must deliver broadband applications to laptops and handheld devices while moving at high speeds. Customers now demand that all this be done without sacrificing quality, reliability, or security. For WiMAX to be successful, it must deliver significantly better performance than current alternatives, such as 3G and Wi-Fi. This is indeed a high bar. Meeting these stringent service requirements while being saddled with a number of constraints imposed by wireless make the system design of broadband wireless a formidable technical challenge. Some of the key technical design challenges are

- Developing reliable transmission and reception schemes to push broadband data through a hostile wireless channel
- Achieving high spectral efficiency and coverage in order to deliver broadband services to a large number of users, using limited available spectrum
- Supporting and efficiently multiplexing services with a variety of QoS (throughput, delay, etc.) requirements
- Supporting mobility through seamless handover and roaming

- Achieving low power consumption to support handheld battery-operated devices
- Providing robust security
- Adapting IP-based protocols and architecture for the wireless environment to achieve lower cost and convergence with wired networks

As is often the case in engineering, solutions that effectively overcome one challenge may aggravate another. Design trade-offs have to be made to find the right balance among competing requirements—for example, coverage and capacity. Advances in computing power, hardware miniaturization, and signal-processing algorithms, however, enable increasingly favourable tradeoffs, albeit within the fundamental bounds imposed by laws of physics and information theory. Despite these advances, researchers continue to be challenged as wireless consumers demand even greater performance. From the above list of challenges we elaborate the following as these are felt to be best overcome by WiMAX as compared to the competition.

1.4.3.1 Quality of Service

QoS is a broad and loose term that refers to the "collective effect of service," as perceived by the user. For the purposes of this discussion, QoS more narrowly refers to meeting certain requirements— typically, throughput, packet error rate, delay, and jitter—associated with a given application. Broadband wireless networks must support a variety of applications, such as voice, data, video, and multimedia, and each of these has different traffic patterns and QoS requirements, as shown in Table 1.4. In addition to the application-specific QoS requirements, networks often need to also enforce policy-based QoS, such as giving differentiated services to users based on their subscribed service plans. The variability in the QoS requirements across applications, services, and users makes it a challenge to accommodate all these on a single-access network, particularly wireless networks, where bandwidth is at a premium.

The problem of providing QoS in broadband wireless systems is one of managing radio resources effectively. Effective scheduling algorithms that balance the QoS requirements of each application and user with the available radio resources need to be developed. In other words, capacity needs to be allocated in the right proportions

among users and applications at the right time. This is the challenge that the MAC-layer protocol must meet: simultaneously handling multiple types of traffic flows—bursty and continuous—of varying throughputs and latency requirements. Also needed are (1) an effective signalling mechanism for users and applications to indicate their QoS requirements and (2) a mechanism for the network to differentiate among various flows towards the users.

Delivering QoS is more challenging for mobile broadband than for fixed. The time variability and unpredictability of the channel become more acute, and complication arises from the need to hand over sessions from one cell to another as the user moves across their coverage boundaries. Handovers cause packets to be lost and introduce additional latency. Reducing handover latency and packet loss is also an important aspect of delivering QoS. Handover also necessitates coordination of radio resources across multiple cells.

So far, our discussion of QoS has been limited to delivering it across the wireless link. From a user perspective, however, the perceived quality is based on the end-to-end performance of the network. To be effective, therefore, QoS has to be delivered end-to-end across the network, which may include, besides the wireless link, a variety of aggregation, switching, and routing elements between the communication end points. IP-based networks are expected to form the bulk of the core network; hence, IP-layer QoS is critical to providing end-to-end service quality.

1.4.3.2 Supporting IP in Wireless

The Internet Protocol (IP) has become the networking protocol of choice for modern communication systems. Internet-based protocols are now beginning to be used to support not only data but also voice, video, and multimedia. Voice over IP is quickly emerging as a formidable competitor to traditional circuit-switched voice and appears likely to displace it over time. Video over IP and IPTV are also emerging as potential rivals to traditional cable TV. Because more and more applications will migrate to IP, IP-based protocols and architecture must be considered for broadband wireless systems.

A number of arguments favour the use of IP-based protocols and architecture for broadband wireless. First, IP-based systems tend to be cheaper because of the economies of scale they enjoy from widespread adoption in wired communication

systems. Adopting an IP architecture can make it easier to develop new services and applications rapidly. The large IP application development community can be leveraged. An IP-based architecture for broadband wireless will enable easier support for such applications as IP multicast and anycast. An IP-based architecture makes it easy to integrate broadband wireless systems with other access technologies and thereby enable converged services.

IP-based protocols are simple and flexible but not very efficient or robust. These deficiencies were not such a huge concern as IP evolved largely in the wired communications space, where transmission media, such as fibre-optic channels, offered abundant bandwidth and very high reliability. In wireless systems, however, introducing IP poses several challenges: (1) making IP-based protocols more bandwidth efficient, (2) adapting them to deliver the required QoS (delay, jitter, throughput, etc.) when operating in bandwidth-limited and unreliable media, and (3) adapting them to handle terminals that move and change their point of attachment to the network.

1.4.4 WiMAX versus 3G and Wi-Fi

The throughput capabilities of WiMAX depend on the channel bandwidth used. Unlike 3G systems, which have a fixed channel bandwidth, WiMAX defines a selectable channel bandwidth from 1.25 MHz to 20 MHz, which allows for a very flexible deployment. When deployed using the more likely 10 MHz TDD channel, assuming a 3:1 downlink-to-uplink split and 2x2 MIMO, WiMAX offers 46 Mbps peak downlink throughput and 7 Mbps uplink. The reliance of Wi-Fi and WiMAX on OFDM modulation, as opposed to CDMA as in 3G, allows them to support very high peak rates. The need for spreading makes very high data rates more difficult in CDMA systems (Jeffrey G. Andrews, Arunabha Ghosh et al. 2007).

More important than peak data rate offered over an individual link, is the average throughput and overall system capacity, when deployed in a multicellular environment. From a capacity standpoint, the more pertinent measure of system performance is spectral efficiency. WiMAX can achieve spectral efficiencies higher than what is typically achieved in 3G systems. The fact that WiMAX specifications accommodated multiple antennas right from the start gives it a boost in spectral efficiency. In 3G systems, on the other hand, multiple-antenna support is being added

in the form of revisions. Further, the OFDM physical layer used by WiMAX is more amenable to MIMO implementations than are CDMA systems from the standpoint of the required complexity for comparable gain. OFDM also makes it easier to exploit frequency diversity and multiuser diversity to improve capacity. Therefore, when compared to 3G, WiMAX offers higher peak data rates, greater flexibility, and higher average throughput and system capacity.

Another advantage of WiMAX is its ability to efficiently support more symmetric links— useful for fixed applications, such as T1 replacement—and support for flexible and dynamic adjustment of the downlink-to-uplink data rate ratios. Typically, 3G systems have a fixed asymmetric data rate ratio between downlink and uplink.

The WiMAX media access control layer is built from the ground up to support a variety of traffic mixes, including real-time and non-real-time constant bit rate and variable bit rate traffic, fundamental bounds on achievable data rates and coverage range. From a global perspective, the 2.3 GHz, 2.5 GHz, 3.5 GHz, and 5.7 GHz bands are most likely to see WiMAX deployments. The WiMAX Forum has identified these bands for initial interoperability certifications. A brief description of these bands follows.

In summary, WiMAX occupies a somewhat middle ground between Wi-Fi and 3G technologies when compared in the key dimensions of data rate, coverage, QoS, mobility, and price.

1.5 Overview of Thesis

This work concentrates on enhancement of the IEEE 802.16 Wireless MAN OFDM MAC layer in terms of QoS and efficiency. We use the terms “Fixed WiMAX”, “WiMAX” and “802.16d” interchangeably and imply the above stated standard. Where ever the term “MAC layer” is used it refers to the MAC layer of the above standard. Providing QoS when bandwidth is unlimited is not an issue. Our task is to allow for QoS requirements to be met while making maximum utilization of the limited and constantly changing wireless resource. WiMAX has put a important step in the right direction with its all-IP infrastructure. The contribution of this work is

enhancements to WiMAX to improve utilization. Following is an overview of the contents of the chapters in this thesis

The relevant sections of the standard are covered in detail in Chapter 2. The standard is vast as there is coverage of all variants including the point-to-point specifications. The scope is limited to the MAC layer specified above.

A simulation model was developed using the QualNet simulation package. The relevant sections of the MAC and PHY were coded into the package using the C programming language. A description of all approximations, configurable parameter and modifications to the standard is presented in Chapter 3.

Chapter 4 deals with optimising the MAC layer service class, UGS, designed to transport VoIP traffic. We have analyzed the effect the packetization interval of the voice encoder has on the system resource usage, taking into account the characteristics, of the OFDM PHY used in Fixed WiMAX. A subscriber's perception of the quality of the service depends on the packet loss rate and the latency. From a service provider's point of view, the requirement would be to honour the service level agreement using the least amount of system resources. A method to facilitate the usage of an optimal packetization interval is proposed, and verified through analysis and simulation.

Wireless channels are prone to errors and this holds true for the OFDM physical layer used in WiMAX. Much research has been done to improve the reliability of wireless links by upper layer techniques. Local retransmission is one of the most commonly used in networks ranging from Wi-Fi to 3G. WiMAX also has included an optional retransmission method based on Automatic Repeat Request (ARQ). The transmitter decides whether to retransmit lost packets, based on feedback messages from the receiver. This is a bandwidth consuming function in broadcast networks as well as incurring extra delay for real-time flows. In Chapter 5 a novel ARQ feedback mechanism for downlink real-time flows based on contention is presented along with a detailed analysis. It is shown that our scheme is flexible and able to improve QoS for real-time traffic with minimal overhead.

Modern user traffic can be considered to be a small proportion of high priority traffic and a large proportion of low priority traffic. As such we have analysed, in depth, the mechanism for serving best effort (BE) traffic using Fixed WiMAX. Our approach is based on Markov chains. Similar approaches have been used successfully

to analyse Wi-Fi, DOCSIS and other contention based access technologies. The effects on the data streams, and utilization of system resources is investigated. A method for controlling throughput by allocating an optimal amount of resources for the contention process has been proposed. This is the subject of Chapter 6.

As stated above servicing BE traffic in the most efficient way is extremely important due to its sheer volume. WiMAX includes a non-real-time polling service (nrtPS) for low priority traffic. In Chapter 7, we investigate its ability to serve bulk transfers, as well as bursty traffic. Several enhancements are looked at which greatly increase the efficiency of nrtPS. A new scheme, enhanced-nrtPS (e-nrtPS) is introduced. This is compared against the contention based access method for throughput, and efficiency. It is shown through simulations that the e-nrtPS is a far better option.

1.6 Published Papers

Shehan Perera, Harsha Sirisena, Krzysztof Pawlikowski. Optimal Packetization Interval for VoIP Applications over IEEE 802.16 Networks, 17th International Teletraffic Congress (ITC) Specialist Seminar, May 2006, Melbourne Australia.

Shehan Perera, Harsha R. Sirisena. Generic Handover Loss Model for Micro-Cellular Wireless Networks, 17th International Teletraffic Congress (ITC) Specialist Seminar, May 2006, Melbourne Australia.

Shehan Perera, Harsha Sirisena, Krzysztof Pawlikowski. Optimal Packetization Interval for VoIP Applications Over IEEE 802.16 Networks, African Journal of Information & Communication Technology (AJICT), UTS Press, Vol 2, No 4 (2006) - Wireless and Mobile Communications.

Shehan Perera, Harsha Sirisena. Contention Based Negative Feedback ARQ for VoIP Services in IEEE 802.16 Networks, 14th International Conference on Networks (ICON 2006) September 2006, Singapore.

Shehan Perera, Harsha R. Sirisena - Analysis of Contention Based Access for Best Effort Traffic on Fixed WiMAX, 4th International Conference on Broadband Communications, Networks and Systems, IEEE Broadnets 2007, North Carolina, USA

Enoch C. Kao, Krzysztof Pawlikowski, Shehan Perera, Harsha R. Sirisena. A Dynamic Random Channel Reservation for Mac Protocols In Multimedia Networks - 11th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2004), in conjunction with 18th European Simulation Multiconference (ESM 2004) June 2004, Magdeburg, Germany.

Chapter 2

IEEE 802.16 Standard

In this chapter we provide an overview of the IEEE 802.16 standard and the roadmap of its past, present and future. Often this standard is referred to as WiMAX, although WiMAX is actually the name of the non-profit consortium which was established for the promotion of interoperability and conformance to the standard. Initially designed as a backhaul solution (Shepard 2006) and then adapted to the point-to-multipoint role, WiMAX has a niche to fill in the broadband market. Many of the heavyweights in the broadband equipment manufacture area have backed WiMAX through the WiMAX Forum and trials have been underway for a few years in several countries. If a cellular service provider could deliver voice, video and data seamlessly to a mobile device, customers would be unwilling to change providers. However amidst many promises from 3G technology this has still not been realized. This sets the stage for an emergent technology to enter the market, which is the motivation behind the IEEE 802.16 suit of standards.

2.1 Standardization Roadmap

The IEEE 802.16, the ‘Air Interface for Fixed Broadband Wireless Access Systems’, also known as the IEEE WirelessMAN air interface comprises of a suit of standards for fixed, portable and mobile broadband wireless access (BWA) in metropolitan area networks (MANs). Originally the standard covered WLL technologies in the 10-66

GHz radio spectrum, which was later extended through many amendments to include both licensed and unlicensed spectra from 2 to 11 GHz (Pareek 2006).

A non-profit organization called the WiMAX Forum was started in 2001, with the sole purpose of promoting interoperability among vendors, testing and certifying interoperability of equipment. This standardized approach is expected to drive down costs through mass production. It is often thought that WiMAX is a technology when in fact it a trade name for a group of IEEE wireless access standards. The WiMAX umbrella presently includes 802.16-2004, 802.16e and 802.16-2005 (collection of all amendments). 802.16 has been around since the late 1990's, initially with the adoption of 802.16 standard and then with 802.16a. Although the work on IEEE 802.16 started in 1999, it was only during 2003 that it received wide attention when the IEEE 802.16a standard was ratified.

2.1.1 IEEE 802.16

The first version of the standard addressed spectrum ranges above 10 GHz (specifically 10 GHz to 66 GHz). Since line-of-sight (LOS) is a primary issue in this range, multipath was addressed in this first version with orthogonal frequency division multiplexing (OFDM) techniques (Baines 2006). Thus it supports wide channels, defined as being greater than 10 MHz in size. This first standard basically addressed licensed-only service delivery (although there is license-free spectrum in this range).

2.1.2 IEEE 802.16a

The 802.16a update added support for spectrum ranges of 2 GHz to 11 GHz. It addressed both licensed and unlicensed ranges. It also incorporated non-line-of-sight (NLOS) capability. This version enhanced the medium access control (MAC) layer capabilities. It also improved QoS features. The European HiperMAN (Stallings 2005; ETSI 2007) standard was supported and a total of three supported physical layers (PHY) were defined. Support for both time division duplexing (TDD) and frequency division duplexing (FDD) was incorporated, providing for both half duplex and full duplex data transmission in cases where FDD is used. Transmission protocols

such as Ethernet, ATM or IP are supported. ETSI BRAN together with the WiMAX forum promotes interoperability of the two standards (WiMAX Forum 2005).

2.1.3 IEEE 802.16c

This standard update dealt mostly with updates in the 10 GHz to 66 GHz range. However, it also addressed issues such as performance evaluation, testing and detailed system profiling. This was a crucial element of the WiMAX toolkit. As there are a great deal of options available with 802.16 in general, the system profile methodology evolved to define what would be mandatory features and what would be optional features. The intent was to guide vendors on mandatory elements that must be met to ensure interoperability. Optional elements such as different levels of security protocols incorporated allow vendors to differentiate their products by price, functionality and market niche.

2.1.4 IEEE 802.16d (IEEE 802.16 – 2004)

All of the Fixed WiMAX standards mentioned above have been rolled into 802.16-2004: it incorporates the original 802.16, 802.16a and 802.16c updates. This final standard supports numerous mandatory and optional elements. Vendors have been shipping their 802.16-2004 products to the Cetecom labs in Spain for interoperability testing since 2005 (Goldman 2005; Hardasmal and Sanders 2007).

The technology supports both TDD and FDD. Its theoretical effective data rate is around 70 Mbps, although real world performance will probably top out around 40 Mbps. It should be noted that while the technology supports at least three PHY layer Modulation schemes, the system profile chosen is OFDM 256-FFT. This is different from the OFDMA flexible FFT system used in 802.16e. Both standards, however, support the former PHY. This distinction is really a market choice. The Forum could have chosen to use OFDM 256-FFT instead of OFDMA. Market forces and in particular the WiBro standard (TTA 2005) may have precluded that.

Enhancements in this version include, support for concatenation of both protocol data units (PDU), and service data units (SDU), which reduces the MAC overhead. The technology improves QoS, particularly with very large SDUs. One clear improvement is support for multiple polling methodologies. The MAC

facilitates polling individually or in groups. It can access allocated bandwidth to make requests, or signal that it needs polling. It can even piggyback polling requests over other traffic. The upshot being that constant cross-talk is obviated with this system, reducing packet collisions and system overhead.

This is a fixed wireless access technology, which is meant to serve as a competitor for incumbent DSL providers and provide VoIP services plus data. It is also capable of serving poorly serviced areas such as in developing countries where laying of Copper can be considered uneconomical or impractical. 802.16-2004 is also a viable solution for backhaul for Wi-Fi hotspots or potentially for cellular networks.

The customer premises equipment (CPE) consists of an outdoor antenna and an indoor modem. Self installable indoor units are also available for customers with good signal reception. Even though designed to be a fixed access technology a certain degree of nomadic behaviour is possible with the user travelling with the CPE to other locations. It has a flexible structure that allows it to be configured for various performance levels, depending on the application.

Carrier Class: Uses licensed spectrum. Typical application is backhaul to connect a cellular base station into the Internet backbone. These applications typically need guaranteed performance and reliability.

Business Class: Uses mostly licensed spectrum, but also uses unlicensed spectrum. Typical application is backhaul to connect a Wi-Fi (802.11a/b/g) hotspot or small business into the Internet backbone. This complements Wi-Fi by enabling less expensive access costs and allows hotspots to be installed almost anywhere.

Consumer Class: Often deployed in unlicensed spectrum. Typical application is wireless digital subscriber line (DSL) for residential or very low-end/non-critical commercial applications.

2.1.5 IEEE 802.16e (IEEE 802.16 – 2005)

The last amendment to be released, IEEE 802.16e conserves the technical updates of Fixed WiMAX while adding robust support for mobile broadband. The technology is based upon the OFDMA technology developed by Runcom (Wreless Design & Development Asia 2007). This OFDMA technique supports 2K-FFT, 1K-FFT, 512-FFT, 256-FFT and 128-FFT. Interestingly, both standards do support the 256-FFT

chosen for 802.16-2004. Many of the mandatory elements for this standard have been agreed upon, and a lot of the remaining work centres around the optional elements.

The OFDMA system allows signals to be divided into many lower-speed sub-channels to increase resistance to multi-path interference. For example, if a 20 MHz channel is subdivided into 1000 sub-channels, each individual user would be allowed a dynamic number of sub-channels based on their distance and needs from the cell (i.e. 4, 64, 298, 312, 346, 610 and 944). If close in, a higher modulation scheme such as 64 quadrature amplitude modulation (64QAM) can be used for higher bandwidth across more channels. If the user is farther away, the number of channels can be reduced with a resultant power increase per channel. Distant users are not dropped but maintained at a lower throughput level.

The specifics of the 802.16e standard are beyond the scope of this work, and not factored in the analysis, or the simulation model.

2.2 Network Topologies

There are two network topologies defined in the standard, which are (1) two-way point-to-multipoint (PMP) topology and (2) mesh topology. PMP is a very common form of network seen in all cellular networks, cable networks, infrastructure based Wi-Fi networks etc. Ad-hoc Wi-Fi networks are a popular example of mesh networks.

2.2.1 Point-to-Multipoint Topology

The downlink, from the BS to the user, operates on a PMP basis. The IEEE Std 802.16 wireless link operates with a central BS and a sectorized antenna that is capable of handling multiple independent sectors simultaneously. Within a given frequency channel and antenna sector, all stations receive the same transmission, or parts thereof. The BS is the only transmitter operating in this direction, so it transmits without having to coordinate with other stations, except for the overall time division duplexing (TDD) that may divide time into uplink and downlink transmission periods. The downlink is generally broadcast. In cases where the downlink map (DL-MAP) does not explicitly indicate that a portion of the downlink subframe is for a specific SS, all SSs capable of listening to that portion of the downlink subframe shall listen.

The SSs check the CIDs in the received PDUs and retain only those PDUs addressed to them.

Subscriber stations share the uplink to the BS on a demand basis. Depending on the class of service utilized, the SS may be issued continuing rights to transmit, or the right to transmit may be granted by the BS after receipt of a request from the user. In addition to individually addressed messages, messages may also be sent on multicast connections (control messages and video distribution are examples of multicast applications) as well as broadcast to all stations. Within each sector, users adhere to a transmission protocol that controls contention between users and enables the service to be tailored to the delay and bandwidth requirements of each user application. This is accomplished through four different types of uplink scheduling mechanisms. These are implemented using unsolicited bandwidth grants, polling, and contention procedures.

2.2.2 Mesh Topology

The main difference between the PMP and optional Mesh modes is that in the PMP mode, traffic only occurs between the BS and SSs, while in the Mesh mode traffic can be routed through other SSs and can occur directly between SSs. Depending on the transmission protocol algorithm used, this can be done on the basis of equality using distributed scheduling, or on the basis of superiority of the Mesh BS, which effectively results in centralized scheduling, or on a combination of both.

Within a Mesh network, a system that has a direct connection to backhaul services outside the Mesh network is termed a Mesh BS. All the other systems of a Mesh network are termed Mesh SS. In general, the systems of a Mesh network are termed nodes. Within Mesh context, uplink and downlink are defined as traffic in the direction of the Mesh BS and traffic away from the Mesh BS, respectively.

The other three important terms of Mesh systems are neighbour, neighbourhood and extended neighbourhood. The stations with which a node has direct links are called neighbours. Neighbours of a node shall form a neighbourhood. A node’s neighbours are considered to be “one hop” away from the node. An extended neighbourhood contains, additionally, all the neighbours of the neighbourhood.

In a Mesh system not even the Mesh BS can transmit without having to coordinate with other nodes. Using distributed scheduling; all the nodes including the Mesh BS shall coordinate their transmissions in their two-hop neighbourhood and shall broadcast their schedules (available resources, requests and grants) to all their neighbours. Optionally the schedule may also be established by directed uncoordinated requests and grants between two nodes. Nodes shall ensure that the resulting transmissions do not cause collisions with the data and control traffic scheduled by any other node in the two-hop neighbourhood. There is no difference in the mechanism used in determining the schedule for downlink and uplink.

Using centralized scheduling, resources are granted in a more centralized manner. The Mesh BS shall gather resource requests from all the Mesh SSs within a certain hop range. It shall determine the amount of granted resources for each link in the network both in downlink and uplink, and communicates these grants to all the Mesh SSs within the hop range. The grant messages do not contain the actual schedule, but each node shall compute it by using the predetermined algorithm with given parameters. All the communications are in the context of a link, which is established between two nodes. One link shall be used for all the data transmissions between the two nodes.

QoS is provisioned over links on a message-by-message basis. No service or QoS parameters are associated with a link, but each unicast message has service parameters in the header. Traffic classification and flow regulation are performed at the ingress node by upper-layer classification/regulation protocol.

2.3 Physical Layer

The Physical Layer (PHY) used in WiMAX is not completely new but rather a combination of many proven technologies (Ohrtman 2005) such as OFDM, TDD, FDD, QPSK, and QAM to name a few. All the work done as part of this thesis is based on a TDD system using an OFDM PHY, hence this variant is described in detail.

2.3.1 Orthogonal Frequency Division Multiplexing

Orthogonal frequency division multiplexing (OFDM) is a multicarrier modulation technique that has recently found wide adoption in a variety of high-data-rate communication systems. This includes digital subscriber lines, wireless LANs (802.11a/g/n), digital video broadcasting (DVB), WiMAX, other emerging wireless broadband systems such as the proprietary Flash-OFDM developed by Flarion (now QUALCOMM), and 3G LTE. OFDM's popularity for high-data-rate applications stems primarily from its efficient and flexible management of intersymbol interference (ISI) in highly dispersive channels.

As the channel delay spread, τ, becomes an increasingly large multiple of the symbol duration, T_s, the ISI becomes very severe. By definition, a high-data-rate system will generally have τ very much greater than T_s, since the number of symbols sent per second is high. In a non-line of sight (NLOS) system, such as WiMAX, which must transmit over moderate to long distances, the delay spread will also be large. In short, wireless broadband systems of all types will suffer from severe ISI and hence will require transmitter and/or receiver techniques that overcome the ISI. Although the 802.16 standards include single-carrier modulation techniques, the vast majority of, if not all, 802.16-compliant systems will use the OFDM modes, which have also been selected as the preferred modes by the WiMAX Forum.

In its simplest form, multicarrier modulation divides the wideband incoming data stream into L narrowband substreams, each of which is then transmitted over a different orthogonal-frequency subchannel. It is based on the Fast Fourier Transform (FFT), which enables channels to partially overlap without degrading the performance of the adjacent channels.

In order to overcome the daunting requirement for multiple RF radios in both the transmitter and the receiver, OFDM uses an efficient computational technique, discrete Fourier transform (DFT), which lends itself to a highly efficient implementation commonly known as the fast Fourier transform (FFT). The FFT and its inverse, the IFFT, can create a multitude of orthogonal subcarriers using a single radio. Fixed WiMAX uses a 256 subcarrier OFDM system. Not all subcarriers are

used for data transmission. Some are used as pilot frequencies while some are used as guard bands to isolate from adjacent spectrum allocations, as shown in Figure 2-1.

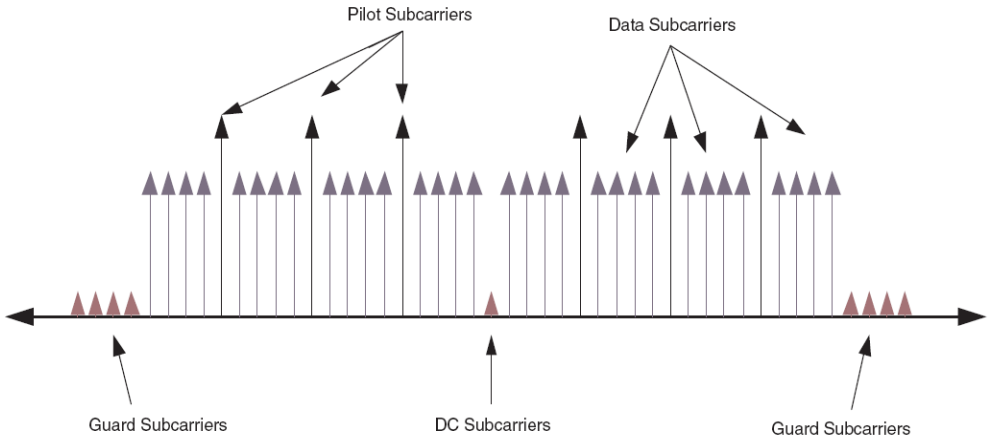

Figure 2-1 Frequency domain view of an OFDM system. All subcarriers other than data subcarriers are used for synchronization and frequency isolation.

Table 2-1 Basic OFDM parameters

OFDM parameters	**Value**	**Scenario**
Bandwidth BW		20 MHz
Sampling rate $F_s = 1/T$	Depends on BW	23.04 MHz
Useful time T_B	256.T	12 μs
T_G/T_B	1/4, 1/8, 1/16, 1/32 1/4	1/4
CP time T_G		3 μs
Symbol time T_{sym}	$T_G + T_B$	15 μs
Carriers NFFT	256	
Data carriers	192	

2.3.2 Overview of Burst Profiles

A Downlink Interface Usage Code (DIUC) is a code used in every IE of the DL-MAP message to inform the SS of the usage of the particular entry. As an example, a DIUC value between 1 and 11 implies a particular burst profile. A value of 14 signifies the end of map. This leads us to a very important term, “burst profile”. A burst profile is a collection of parameters, which gives a complete description of the PHY attributes a

transmitter or receiver should use in communication. Among these attributes is the Forward Error Correction Code Type (FEC Code Type). This field (0-255, of which 20-255 is reserved) specifies which modulation scheme and code rate to use. There are 4 modulation schemes defined, 64 QAM, 16 QAM, QPSK and BPSK and different code rates. Seven representative burst profiles have been used for all PHY specifications in the rest of this thesis. When referring to a modulation scheme, the code rate is considered part of it, e.g. 16 QAM 3/4.

2.3.3 Adaptive Antenna Systems (AAS)

The standard and WiMAX specify beam forming techniques where an array of antennas can be used to increase the gain in the direction of an intended subscriber while nulling out interference to other users. AAS can enable the use of Spatial Division Multiple Access (SDMA) which allows frequency reuse for multiple users who are spatially separated. This can increase system gain in the Downlink (DL) direction.

2.3.4 Adaptive Modulation and Coding

Adaptive modulation and coding (AMC) is used to allow users with a wide range reception conditions to communicate with the BS. A transmitter may pick a suitable “burst profile” depending on channel conditions.

A randomizer adds a pseudo-random binary sequence to the DL and UL bit stream to avoid long rows of zeros or ones for better coding performance. It appends a tail byte to bring the convolutional coder in the zero state after each burst.

The forward error correction (FEC) scheme consists of the concatenation of a Reed–Solomon (RS) outer code and a convolutional inner code (CC). The RS coder corrects burst errors at the byte level. It is particularly useful for OFDM links in the presence of multipath propagation. The CC corrects independent bit errors. A CC decoder can benefit from softbit input generated from de-modulation and de-puncturing. The concatenation of both codes is made rate-compatible by the following puncturing functionality. Based on four puncturing patterns bits are removed to realize different code rates. The support of block turbo coding and convolutional turbo coding is an optional mode.

The interleaving is composed of a block and a bit interleaver. The block interleaver maps adjacent coded bits onto non-adjacent subcarriers to overcome burst errors. The bit interleaver maps adjacent coded bits alternately onto less and more significant bits of the constellation to avoid long runs of unreliable bits.

BPSK, QPSK, 16-QAM and 64-QAM are the modulation schemes to modulate bits to the complex constellation points. The FEC options are paired with the modulation schemes to form burst profiles, i.e., PHY modes of varying robustness and efficiency. The possible PHY modes are listed in Table 2-2. The basic IEEE 802.16 OFDM parameters are outlined in the second column of Table 2-1 (Hoymann 2005).

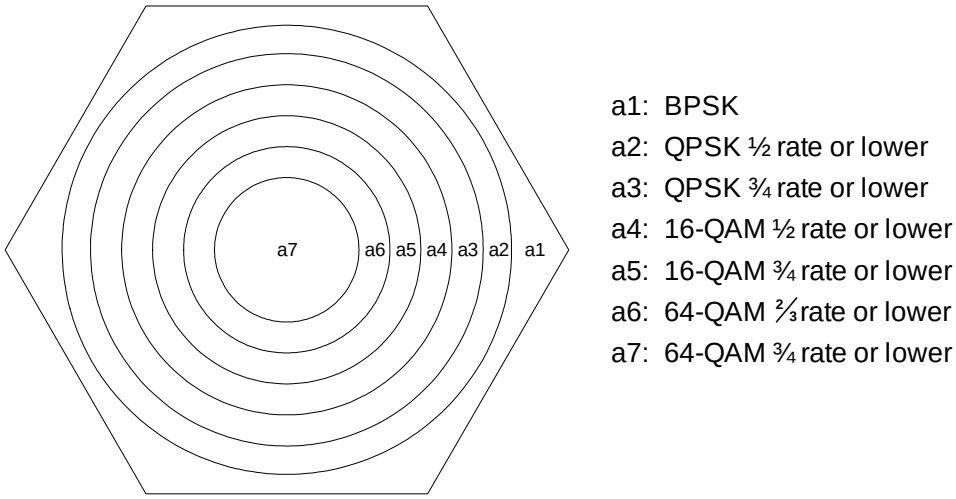

Figure 2-2 Annulus area which can be served by the different modulation schemes (not to scale)

Table 2-2 PHY mode modulations schemes, SNRs and coverage

Modulation	**Coding rate**	**Receiver SNR (dB)**	**Surface [%]**
BPSK	1/2	6.4	39.40
QPSK	1/2	9.4	20.56
QPSK	3/4	11.2	27.95
16 QAM	1/2	16.4	4.10
16 QAM	3/4	18.2	5.15
64 QAM	1/2	22.7	0.92
64 QAM	3/4	24.4	1.92

Each switching point between two different PHY modes results in a certain radius. The radius of the last switching point, i.e., BPSK 1/2 marks the cell boundary. In this illustration the cell area has a maximum radius of approximately 7.4 km. The parts of the surface area of the cell which are covered by specific PHY modes are regions lying between two concentric circles, see Figure 2-2. The area of the annulus ($F_{Annulus}$) formed by two circles of radii R_1 and R_2 is,

$$F_{Annulus} = \pi.\left(R_1^2 - R_2^2\right). \tag{2.1}$$

The cell boundary in an ideal cellular deployment is a hexagonal cell so that the area belonging to the last mode, i.e., BPSK 1/2 is not a whole annulus but certain parts of it are cut away. The area covered by BPSK 1/2 can be calculated as,

$$F_{BPSK\,1/2} = \frac{3}{2}\sqrt{3}.R_{BPSK\,1/2}^2 - \pi.R_{QPSK\,1/2}^2 . \tag{2.2}$$

The area per PHY mode is a certain fraction of the whole cell area. The proportion of each surface area per PHY mode is listed in Table 2-2. Note that the distribution of PHY modes in a hexagonal cell neither depends on the frequency band in use nor on the transmission power. One can easily see that the annulus where the most robust PHY mode BPSK 1/2 is in use is represented over proportionately. The sensitive and powerful modes in the inner circles of the cell cannot be utilized very often because their range is limited to a small area. In an area with high user density smaller cells can be used so that only the few highest burst profiles are utilised. This is discussed further in a preceding section.

2.3.5 Physical Layer Variants

There are five variants based on PHY specifications. Some of these use single carrier (SC) and the rest use multicarrier OFDM. The variants along with frequency ranges, duplexing methods etc are listed in Table 2-3. Residential applications dictate the need for NLOS propagation which the drives the 2-11 GHz variants. Often residential rooftops are not high enough for the antenna to see the BS, or outdoor equipment could be expensive due to hardware and installation costs.

Table 2-3 Variants of WiMAX

Dsignation	Function	LOS/NLOS	Frequency	Duplexing Methods
WirelessMAC SC	Point-to-point	LOS	10-66 GHz	TDD, FDD
WirelessMAC SCa	Point-to-point	NLOS	2-11 GHz	TDD, FDD
WirelessMAC OFDM	Point-to-multipoint	NLOS	2-11 GHz	TDD, FDD
WirelessMAC OFDMA	Point-to-multipoint	NLOS	2-11 GHz	TDD, FDD
WirelessMAC HUMAN	Point-to-multipoint	NLOS	2-11 GHz	TDD, FDD

2.3.5.1 WirelessMAN SC

This PHY specification, targeted for operation in the 10–66 GHz frequency band, is designed with a high degree of flexibility in order to allow service providers the ability to optimize system deployments with respect to cell planning, cost, radio capabilities, services, and capacity. Both TDD and FDD configurations are supported. Both cases support adaptive burst profiling in which transmission parameters, including the modulation and coding schemes, may be adjusted individually to each SS on a frame-by-frame basis. The FDD case supports full-duplex SSs as well as halfduplex SSs, which do not transmit and receive simultaneously.

The uplink PHY is based on a combination of TDMA and DAMA. In particular, the uplink channel is divided into a number of time slots. The number of slots assigned for various uses (registration, contention, guard, or user traffic) is controlled by the MAC in the BS and may vary over time for optimal performance. The downlink channel is TDM, with the information for each SS multiplexed onto a single stream of data and received by all SSs within the same sector. To support half-duplex FDD SSs, provision is also made for a TDMA portion of the downlink.

The downlink PHY includes a Transmission Convergence sublayer that inserts a pointer byte at the beginning of the payload to help the receiver identify the beginning of a MAC PDU. Data bits coming from the Transmission Convergence sublayer are randomized, FEC encoded, and mapped to a QPSK, 16 QAM, or 64-QAM (optional) signal constellation.

The uplink PHY is based upon TDMA burst transmission. Each burst is designed to carry variable-length MAC PDUs. The transmitter randomizes the incoming data, FEC encodes it, and maps the coded bits to a QPSK, 16-QAM (optional), or 64-QAM (optional) constellation (IEEE 802.16 WG 2004).

2.3.5.2 WirelessMAN SCa

The WirelessMAN-SCa PHY is based on single-carrier technology and designed for NLOS operation in frequency bands below 11 GHz. For licensed bands, channel bandwidths allowed shall be limited to the regulatory provisioned bandwidth divided by any power of 2 no less than 1.25 MHz. Elements within this PHY include, TDD and FDD definitions, one of which must be supported, TDMA uplink, TDM or TDMA downlink. block adaptive modulation and FEC coding for both uplink and downlink, framing structures that enable improved equalization and channel estimation performance over NLOS and extended delay spread environments, physical slot unit granularity in burst sizes, concatenated FEC using Reed–Solomon and pragmatic trellis coded modulation (TCM) with optional interleaving, additional BTC and CTC FEC options, no-FEC option using ARQ for error control, space time coding (STC) transmit diversity option, robust modes for low CINR operation, and parameter settings and MAC/PHY messages that facilitate optional AAS implementations.

2.3.5.3 WirelessMAN OFDM

This air interface uses an OFDM 256 (256 point FFT) with TDMA access. It is intended mainly for point-to-multipoint fixed access deployments where SSs are residential gateways deployed within homes, small office home office (SOHO) or even businesses. Even though both TDD and FDD modes are supported, TDD is the simpler option in terms of hardware.

In TDD the frame structure is shown in Figure 2-3. The frame is divided into the DL and UL subframes which are separated by the receive-transmit-gap (RTG). The DL subframe is made up of a preamble, a Frame Control Header (FCH) which contains the downlink map (DL-MAP), uplink map (UL-MAP), downlink channel descriptor (DCD) and uplink channel descriptor (UCD) and one of more DL bursts. These could be broadcast data or for specific SSs. The two MAP messages mentioned

above complete define the entire frame. They specify the SSs that are transmitting/receiving in each burst and which burst profile will be used.

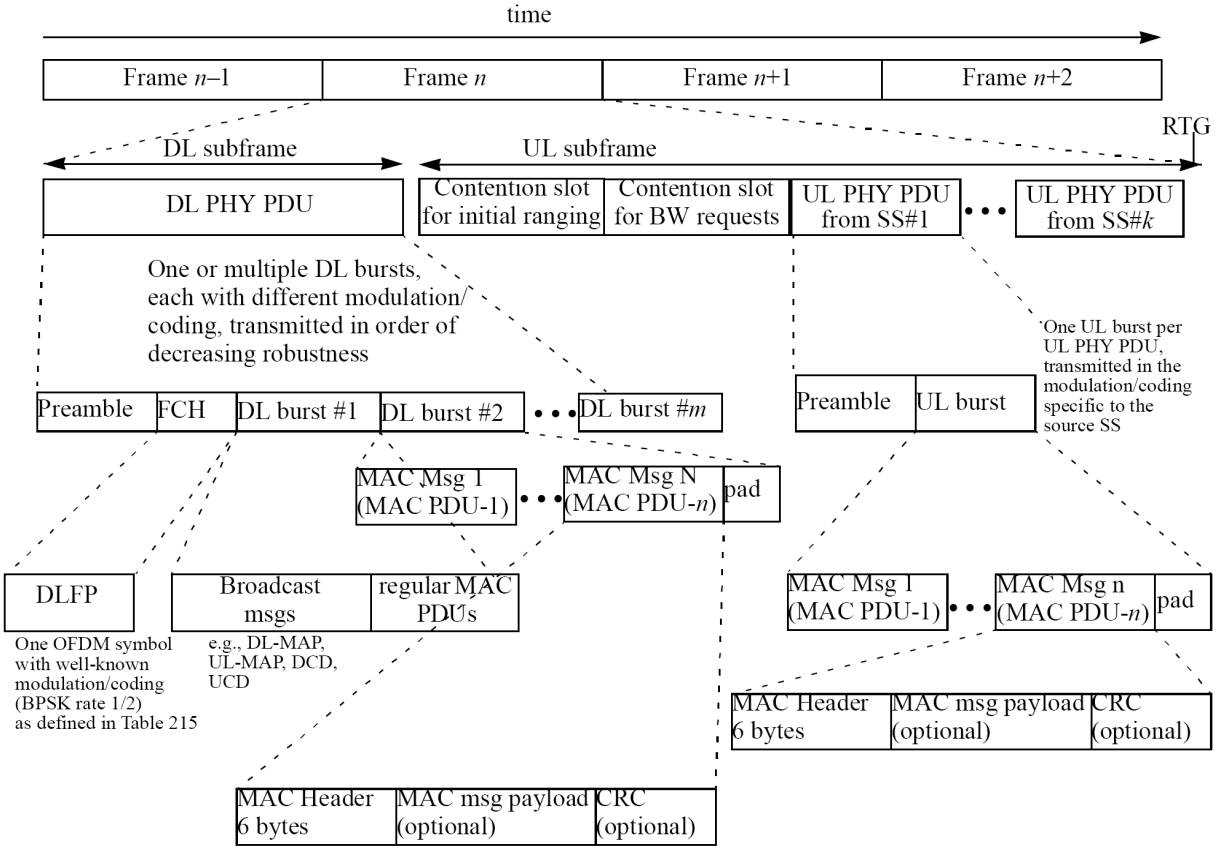

Figure 2-3 Example of OFDM frame structure with TDD

Each burst consists of an integer number of OFDM symbols and is assigned a burst profile relevant to the individual SS. The UL subframe consists of ranging slots, contention based request opportunities, unicast BW polls and UL bursts by SSs. The smallest allocation unit is one OFDM symbol duration.

2.3.5.4 WirelessMAN OFDMA

This variant uses an OFDM 2048-FFT function, and supports subchannelization in both directions, i.e., UL and DL. The standard supports five subchannelization schemes. The OFDMA PHY supports both TDD and FDD operation. The modulations schemes are the same as explained in the previous section. STC, AAS and MIMO are also supported. MIMO involves using multiple antennas at the BS as well as the SS.

In the most common TDD variant the frame structure is very similar to that described before except for the subchannel allocation in both directions which could have multiple stations receiving or transmitting different data simultaneously. The frame is divided into a number of zones to accommodate the subchannelization schemes. The MAC layer is responsible for time and frequency resources to SSs as necessary.

2.3.5.5 WirelessMAN HUMAN

This stands for Wireless High Speed Metropolitan Area Network. It is similar to the afore mentioned OFDM based schemes and is focused on Unlicensed National Information Infrastructure (UNII) devices and other unlicensed bands (Ohrtman 2005). The MAC layer is based on 802.16 with a few modifications with a primary focus on the 5-6 GHz band (Marks, Satapathy et al. 2001).

2.4 Medium Access Control Layer

The IEEE 802.16 MAC layer performs the standard Medium Access Control (MAC) layer function of providing a medium-independent interface to the 802.16 PHY. Because the 802.16 PHY is a wireless PHY layer, the main focus of the MAC layer is to manage the resources of the airlink in an efficient manner. Our focus is on the PMP network model. Following are the basic functions of the Mac layer.

- Segment or concatenate the service data units (SDUs) received from higher layers into the MAC PDU (protocol data units), the basic building block of MAC-layer payload
- Select the appropriate burst profile and power level to be used for the transmission of MAC PDUs
- Retransmission of MAC PDUs that were received erroneously by the receiver when automated repeat request (ARQ) is used
- Provide QoS control and priority handling of MAC PDUs belonging to different data and signalling bearers
- Schedule MAC PDUs over the PHY resources
- Provide support to the higher layers for mobility management
- Provide security and key management
- Provide power-saving mode and idle-mode operation

The MAC layer of WiMAX, as shown in Figure 2-4, is divided into three distinct components: (1) the service-specific convergence sublayer (CS), (2) the common-part sublayer, and (3) the security sublayer.

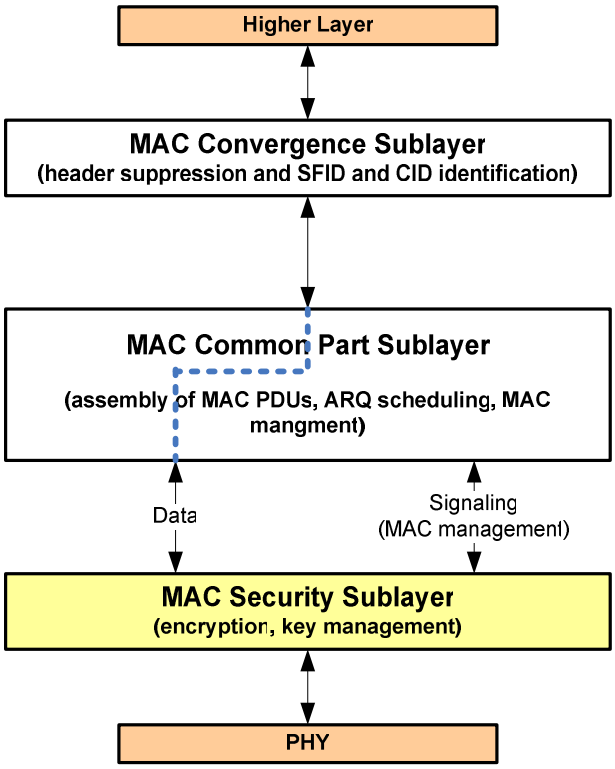

Figure 2-4 The three main components of the MAC layer

In this chapter, we first describe the CS and its various functions. Next, we describe the MAC common-part sublayer, the construction of MAC PDUs, bandwidth allocation process, QoS control, and network-entry procedures. We then turn to the mobility-management and power-saving features of the WiMAX MAC layer. The security sublayer is responsible for encryption, authorization, and proper exchange of encryption keys between the BS and the SS. We do not describe the security sublayer functionality in the chapter.

2.4.1 Convergence Sublayer

The CS, which is the interface between the MAC layer and layer 3 of the network, receives data packets from the higher layer. These higher-layer packets are also known as MAC service data units (SDU). The CS is responsible for performing all operations that are dependent on the nature of the higher-layer protocol, such as

header compression and address mapping. The CS can be viewed as an adaptation layer that masks the higher-layer protocol and its requirements from the rest of the MAC and PHY layers of a WiMAX network.

IPv4, IPv6, Ethernet, 802.1q are some of the higher layer protocols that are supported in WiMAX. Apart from header compression, the CS is also responsible for mapping higher-layer addresses, such as IP addresses, of the SDUs onto the identity of the PHY and MAC connections to be used for its transmission. This functionality is required because there is no visibility of higher-layer addresses at the MAC and PHY layers.

2.4.1.1 Flow Identification and CID allocation

The WiMAX MAC layer is connection oriented and identifies a logical connection between the BS and the MS by a unidirectional connection identifier (CID). The CIDs for UL and DL connections are different. The CID can be viewed as a temporary and dynamic layer 2 address assigned by the BS to identify a unidirectional connection between the peer MAC/PHY entities and is used for carrying data and control plane traffic. In order to map the higher-layer address to the CID, the CS needs to keep track of the mapping between the destination address and the respective CID. It is quite likely that SDUs belonging to a specific destination address might be carried over different connections, depending on their QoS requirements, in which case the CS determines the appropriate CID, based on not only the destination address but also various other factors, such as, service flow ID (SFID) and source address. The IEEE 802.16 suite of standards defines a CS for ATM (asynchronous transfer mode) services and packet service (Nair, Chou et al. 2004). However, the WiMAX Forum has decided to implement only IP and Ethernet (802.3) CS.

2.4.1.2 Packet Header Suppression

One of the key tasks of the CS is to perform packet header suppression (PHS). At the transmitter, this involves removing the repetitive part of the header of each SDU. For example, if the SDUs delivered to the CS are IP packets, the source and destination IP addresses contained in the header of each IP packet do not change from one packet to the next and thus can be removed before being transmitted over the air. Similarly at the receiver: The repetitive part of the header can be reinserted into the SDU before

being delivered to the higher layers. The PHS protocol establishes and maintains the required degree of synchronization between the CSs at the transmitter and the receiver.

In WiMAX, PHS implementation is optional; however, most systems are likely to implement this feature, since it improves the efficiency of the network to deliver such services as VoIP. The PHS operation is based on the "PHS rule", which provides all the parameters related to header suppression of the SDU.

2.4.2 MAC Common Part Sublayer

The common-part sublayer of the MAC layer performs all the packet operations that are independent of the higher layers, such as fragmentation and concatenation of SDUs into MAC PDUs, transmission of MAC PDUs, QoS control, and ARQ.

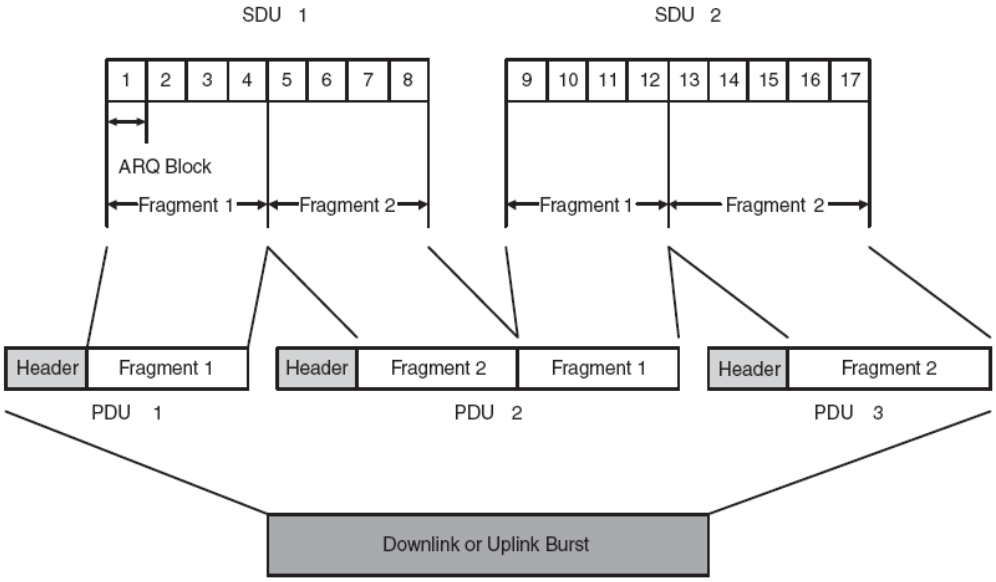

Figure 2-5 Segmentation and concatenation of SDUs into MAC PDUs

2.4.2.1 MAC PDU Construction and Transmission

As the name suggests, the MAC common-part sublayer is independent of the higher-layer protocol and performs such operations as scheduling, ARQ, bandwidth allocations, modulation, and code rate selection. The SDUs arriving at the MAC common-part sublayer from the higher layer are assembled to create the MAC PDU, the basic payload unit handled by the MAC and PHY layers. Based on the size of the

payload, multiple SDUs can be carried on a single MAC PDU, or a single SDU can be fragmented to be carried over multiple MAC PDUs, see Figure 2-5.
When an SDU is fragmented, the position of each fragment within the SDU is tagged by a sequence number. The sequence number enables the MAC layer at the receiver to assemble the SDU from its fragments in the correct order.

In order to efficiently use the PHY resources, multiple MAC PDUs destined to the same receiver can be concatenated and carried over a single transmission opportunity or data region. In the UL and DL data regions of a SS is a contiguous set of slots reserved for its transmission opportunities. For non-ARQ-enabled connections, each fragment of the SDU is transmitted in sequence. For ARQ-enabled connections, the mechanics of the ARQ scheme used are detailed in Chapter 5.

2.4.3 MAC Header Types and Management Messages

Each MAC PDU consists of a header followed by a payload and a cyclic redundancy check (CRC). The CRC is based on IEEE 802.3 and is calculated on the entire MAC PDU; the header and the payload. WiMAX has two types of PDUs, each with a very different header structures:

1) Generic MAC header: A generic MAC PDU starts with a generic MAC header. Generic MAC PDUs are used for carrying data and MAC-layer signalling messages.

2) Bandwidth Request (BR) MAC header: The bandwidth request PDU is used by the MS to indicate to the BS that more bandwidth is required in the UL, due to pending data transmission. A bandwidth request PDU consists only of a bandwidth-request header, with no payload or CRC

The maximum length of the MAC PDU is 2048 bytes, including header, payload, and CRC. For PMP, the MAC defines the following subheaders.

1) Mesh subheader: Follows generic header when mesh networking is used.

2) Fragmentation subheader: Follows the generic MAC header and indicates that the SDU is fragmented over multiple MAC PDUs.

3) Packing subheader: Indicates that multiple SDUs or SDU fragments are packed into a single MAC PDU and are placed at the beginning of each SDU or SDU fragment.

4) Fast-feedback allocation subheader: Indicates that the PDU contains feedback from the MS about the DL channel state information. This subheader provides the functionality for channel state information feedback for MIMO and non-MIMO implementations.

5) Grant-management subheader: Used by the MS, conveys various messages related to bandwidth management, such as polling request and additional-bandwidth request. Using this subheader is more efficient than the bandwidth-request PDU for additional bandwidth during an ongoing session, since it is more compact and does not require the transmission of a new PDU. The bandwidth-request PDU is generally used for the initial bandwidth request.

The standard defines a number of MAC management messages that are used to pass control information between the SS and BS. These messages are divided into broadcast messages, initial ranging messages, basic messages, and primary management messages.

2.4.4 Network Entry

In order to communicate on the network, an SS needs to successfully complete the network entry process with the desired BS. The network entry process is divided into the following stages:

1) DL channel synchronization
2) initial ranging
3) capabilities negotiation
4) authentication message exchange
5) registration
6) IP connectivity

The network entry state machine moves to reset if it fails to succeed from a state. Upon completion of the network entry process, the SS creates one or more service flows to send data to the BS. The following subsections describe each of these stages in more detail.

2.4.4.1 Downlink Channel Synchronization

When an SS wishes to enter the network, it scans for a channel in the defined frequency list. Normally an SS is configured to use a specific BS with a given set of operational parameters, when operating in a licensed band. If the SS finds a DL

channel and is able to synchronize at the PHY level (it detects the periodic frame preamble), then the MAC layer looks for DCD and UCD to get information on modulation and other DL and UL parameters.

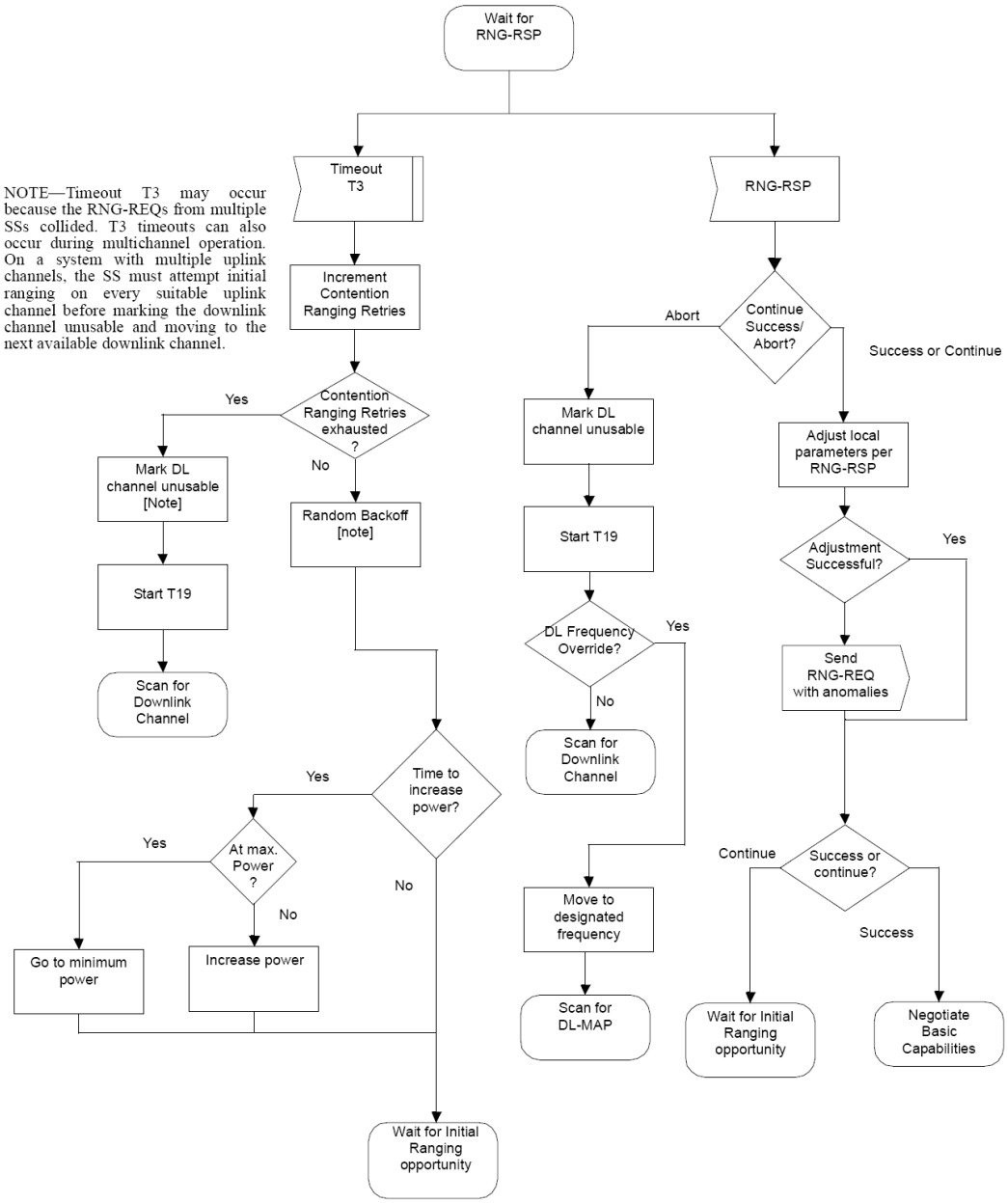

Figure 2-6 Detailed view of Initial Ranging of a SS

2.4.4.2 Initial Ranging

When an SS has synchronized with the DL channel and received the DL and UL MAP for a frame, it begins the initial ranging process by sending a ranging request

MAC message on the initial ranging interval using the minimum transmission power. If it does not receive a response, the SS sends the ranging request again in a subsequent frame, using higher transmission power. Eventually the SS receives a ranging response. The response either indicates power and timing corrections that the SS must make or indicates success. If the response indicates corrections, the SS makes these corrections and sends another ranging request. If the response indicates success, the SS is ready to send data on the UL. This process is illustrated in Figure 2-6.

2.4.4.3 Negotiation of Capabilities

After successful completion of initial ranging, the SS sends a capability request message to the BS describing its capabilities in terms of the supported modulation levels, coding schemes and rates, and duplexing methods. The BS accepts or denies the SS, based on its capabilities.

2.4.4.4 Authentication

After capability negotiation, the BS authenticates the SS and provides key material to enable the ciphering of data. The SS sends the X.509 certificate of the SS manufacturer and a description of the supported cryptographic algorithms to its BS. The BS validates the identity of the SS, determines the cipher algorithm and protocol that should be used, and sends an authentication response to the SS. The response contains the key material to be used by the SS. The SS is required to periodically perform the authentication and key exchange procedures to refresh its key material.

2.4.4.5 Registration

After successful completion of authentication the SS registers with the network. The SS sends a registration request message to the BS, and the BS sends a registration response to the SS. The registration exchange includes IP version support, SS managed or non-managed support, ARQ parameters support, classification option support, CRC support, and flow control.

2.4.4.6 IP Connectivity

The SS then starts Dynamic Host Configuration Protocol (DHCP) (IETF RFC 2131) to get the IP address and other parameters to establish IP connectivity. The BS and SS maintain the current date and time using the time of the day protocol (IETF RFC 868).

The SS then downloads operational parameters using Trivial File Transfer Protocol (TFTP) (IETF RFC 1350).

2.4.4.7 Transport Connection Creation

After completion of registration and the transfer of operational parameters, transport connections are created. For preprovisioned service flows, the connection creation process is initiated by the BS. The BS sends a dynamic service flow addition request message to the SS and the SS sends a response to confirm the creation of the connection. Connection creation for non-preprovisioned service flows is initiated by the SS, by sending a dynamic service flow addition request message to the BS. The BS responds with a confirmation. This procedure will be discussed in detail in Chapter 4.

2.4.5 Scheduling and Link Adaptation

The goal of scheduling and link adaptation is to provide the desired QoS treatment to the traffic traversing the airlink, while optimally utilizing the resources of the airlink. Scheduling in the 802.16 MAC is divided into two related scheduling tasks: (1) scheduling the usage of the airlink among the SSs, and (2) scheduling individual packets at the BSs and SSs.

The airlink scheduler runs on the BS and is generally considered to be part of the BS MAC layer. However it is not defined by the standard and is implementation dependant. This scheduler determines the contents of the DL and UL portions of each frame. When optional modes such as transmit diversity, AAS, and MIMO are used, the MAC layer must divide the UL and DL subframes into normal, transmit diversity, AAS, and MIMO zones, to accommodate SSs that are to be serviced using one of these modes. Having divided the subframes into zones, the scheduler allocates transmission opportunities to individual SSs within the zone in which they operate. In the OFDM, DL transmission opportunities are time slots, while in the OFDM UL and OFDMA UL and DL, transmission opportunities are time slots within individual subchannels. When AAS with SDMA is employed within the BS, a given time slot on a given subchannel can be allocated to multiple SSs. This means that the two-dimensional scheduling problem (with time slots along one axis and subchannels along the other) becomes a three-dimensional problem, with the third axis being the

spatial axis. The MAC must determine which SSs have orthogonal spatial signatures, making them good candidates for sharing the same subchannel/time slot combinations.

The airlink scheduler must also determine the appropriate burst profile for communication with each SS. The BS monitors the SNR and increases or decreases the coding rate and modulation level, accordingly for traffic for an SS. This achieves the highest possible throughput, while maintaining a given BER level.

The airlink scheduler determines the bandwidth requirements of the individual SSs, based on the service classes of the connections, and on the status of the traffic queues at the BS and SS. The BS monitors its own queues to determine the bandwidth requirements of the DL, and utilizes a number of different communication mechanisms (such as polling and unsolicited bandwidth requests) to keep informed of the bandwidth requirements of the SSs for the UL.

Finally, there is a packet scheduler in the BS and SS. This scheduler schedules packets from the connection queues into the transmission opportunities allocated to the SS within each frame.

2.4.6 Quality of Service

The principal mechanism for providing QoS is to associate packets traversing the MAC interface into a service flow as identified by the CID. A service flow is a unidirectional flow of packets that is provided a particular QoS, according to the QoS Parameter Set defined for the service flow. The primary purpose of the QoS features is to define transmission ordering and scheduling on the air interface. However, these features often need to work in conjunction with mechanisms beyond the air interface in order to provide end-to-end QoS or, to police the behaviour of SSs. Service flows exist in both the uplink and downlink direction, and may exist without actually being activated to carry traffic.

2.4.6.1 Scheduling Services

Scheduling services represent the data handling mechanisms supported by the MAC scheduler for data transport on a connection. Each connection is associated with a single data service. Each data service is associated with a set of QoS parameters that

quantify aspects of its behaviour. These parameters are managed using the DSA and DSC message dialogs. Four services are supported:

Unsolicited Grant Service (UGS): The UGS is designed to support real-time data streams consisting of fixed-size data packets issued at periodic intervals, such as T1/E1 and Voice over IP without silence suppression. The mandatory QoS service flow parameters for this scheduling service are Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter, and Request/Transmission Policy. If present, the Minimum Reserved Traffic Rate parameter shall have the same value as the Maximum Sustained Traffic Rate parameter.

Real-time Polling Service (rtPS): The rtPS is designed to support real-time data streams consisting of variable-sized data packets that are issued at periodic intervals, such as moving pictures experts group (MPEG) video. The mandatory QoS service flow parameters for this scheduling service are Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Maximum Latency, and Request/Transmission Policy.

Non-real-time Polling Service (nrtPS): The nrtPS is designed to support delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required, such as FTP. The mandatory QoS service flow parameters for this scheduling service are Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Traffic Priority, and Request/Transmission Policy.

Best Effort (BE): The BE service is designed to support data streams for which no minimum service level is required and therefore may be handled on a space-available basis. The mandatory QoS service flow parameters for this scheduling service are Maximum Sustained Traffic Rate, Traffic Priority, and Request/ Transmission Policy.

2.5 Conclusion

IEEE 802.16 - 2004 specifies the WirelessMAN air interface for wireless MANs. This standard defines the Media Access Control (MAC) layer and the physical layer

specifications of a fixed point-to-multipoint broadband wireless access system. In actual fact, this is a collection of standards which cover, point-to-point, point-to-multipoint and mesh architectures, along with complete PHY and MAC layer specifications for each of them. The portion of the standard which is considered in scope of this work is, the WirelessMAN OFDM specification. OFDM provides excellent reception in multipath environments, and the ability of broadband data rates. AAS and MIMO have been included from the onset, as have security consideration in the form a separate Security CS.

QoS has been given its due place, unlike Wi-Fi, which has evolved towards an architecture that supports QoS after many amendments. In other words this standard was in a relatively mature state from the start. There are four native QoS classes to support, real-time VoIP and video traffic, interactive traffic, bulk data transfer and bursty low priority traffic. Procedures are in place to classify and associate traffic flows to the appropriate service classes. Upper layer scheduling is not within scope. This is vendor and implementation specific, but is a vital link in the chain of QoS provision.

Chapter 3

IEEE 802.16 Simulation Model

This chapter describes the simulation model used for all the simulations conducted as part of this thesis. At the time when this project was started, no simulator existed which could accurately model the IEEE WirelessMAN standard. QualNet by Scalable Network Technologies was chosen from among a few other popular commercial and open source simulators, due to its ease of use, flexibility of modifications and excellent debugging capabilities using Microsoft Visual C++. Its modular structure allows for simple swapping and integration of new layer models. All coding is done using the C programming language, which is robust and efficient. This chapter gives a brief introduction to the different components of QualNet, and introduces the protocol stack that forms the basis of QualNet architecture. We give an overview of the organisation of a protocol within the simulator, and the common functions performed. We then provide an overview of the design aspects of the Fixed WiMAX standard which has been coded with all features important to this project. As with any simulation package approximations have been made to keep the task to a reasonable level of complexity. These approximations and exclusions are also stated in the following sections.

3.1 QualNet

QualNet has several core components, as well as various add-on components. This section provides a brief description of the core components of QualNet. Detailed descriptions, functions, and usage instructions for each of the QualNet components are available in the IDE User's Guide and Programmer's Guide (Scalable Network Technologies Inc. 2005, 2006).

QualNet Simulator

QualNet Simulator is a state-of-the-art simulator for large, heterogeneous networks and the distributed applications that execute on those networks. QualNet Simulator is an extremely scalable simulation engine, accommodating high-fidelity models of networks of tens of thousands of nodes. QualNet makes good use of computational resources, and models large-scale networks with heavy traffic and mobility, in reasonable simulation times. QualNet Simulator has the following attractive features:

- Fast model set up with a powerful Graphical User Interface (GUI) for custom code development and reporting options
- Instant playback of simulation results to minimize unnecessary model executions
- Fast simulation results for thorough exploration of model parameters
- Scalable up to tens of thousands of nodes
- Real-time simulation for man-in-the-loop and hardware-in-the-loop models
- Multi-platform support

QualNet Scenario Designer

QualNet Scenario Designer is a graphical tool that provides an intuitive model set up capability and is used to create and design experiments in QualNet. The Scenario Designer enables a user to define the geographical distribution, physical connections and the functional parameters of the network nodes, all using intuitive click and drag tools, and to define network layer protocols and traffic characteristics for each node.

QualNet Animator

QualNet Animator is used to execute and animate experiments created in the Scenario Designer. Using the Animator a user can watch traffic flow through the network and create dynamic graphs of critical performance metrics as a simulation is running.

QualNet Packet Tracer

QualNet Packet Tracer is a packet-level visualization tool for viewing the contents of packets as they travel up and down the protocol stack. Packet tracing can be enabled so that selected headers are displayed and packets matching filtering criteria are captured. Tracing can be done as ascii data so that the traces may be analyzed by external applications such as tcptrace.

QualNet Analyzer

QualNet Analyzer statistical graphing tool that displays network statistics generated from a QualNet experiment. Using the Analyzer, a user can view statistics as they are being generated, as well as compare results from different experiments.

QualNet Protocol Designer

QualNet Protocol Designer has the following two primary functions: Design of new network protocols, and simplified mechanism for incorporating protocols into the QualNet Simulator. The Protocol Designer provides an intuitive state-based visual tool to define the events and processes of a protocol model. A user can modify ready-made protocol models or generate code from scratch for custom protocols and for special statistical reporting.

3.1.1 QualNet Protocol Stack

QualNet uses a layered architecture similar to that of the TCP/IP network protocol stack. Within that architecture, data moves between adjacent layers. QualNet's protocol stack consists of, from top to bottom, the Application, Transport, Network, Link (MAC) and Physical Layers.

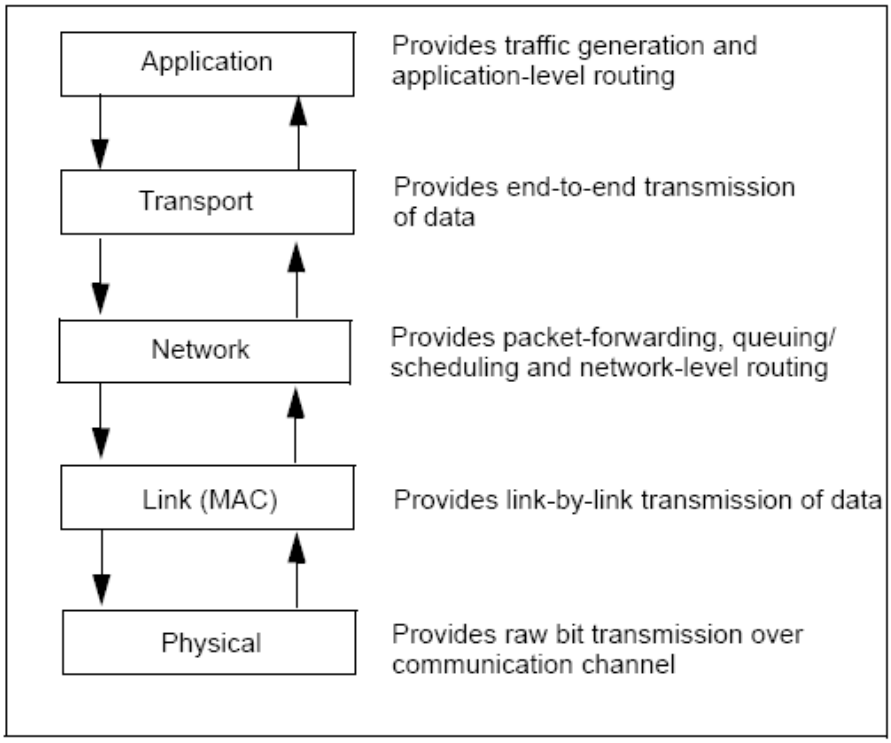

Figure 3-1 QualNet protocol stack

Adjacent layers in the protocol stack communicate via well-defined APIs, and generally, layer communication occurs only between adjacent layers. For example, Transport Layer protocols can get and pass data to and from the Application and Network Layer protocols, but cannot do so with the Link (MAC) Layer protocols or the Physical Layer protocols. This rule concerning communication only between adjacent layers may be circumvented when innovative cross layer communication may be required. Figure 3-1 depicts the QualNet protocol stack and the general functionality of each layer. Each node in QualNet runs a protocol stack. Each layer provides a service to the layer above it, by using the services of the layers below it.

3.1.2 Modelling Protocols in QualNet

Each protocol operates at one of the layers of the stack. Protocols in QualNet essentially operate as a finite state machine. The occurrence of an event corresponds to a transition in the finite state machine. The interface between the layers is also event based. Each protocol can either create events that make it change its own state (or perform some event handling), or create events that are processed by another

protocol. To pass data to, or request a service from, an adjacent layer, a protocol creates an event for that layer. Figure 3-2 shows the finite state machine representation of a protocol in QualNet.

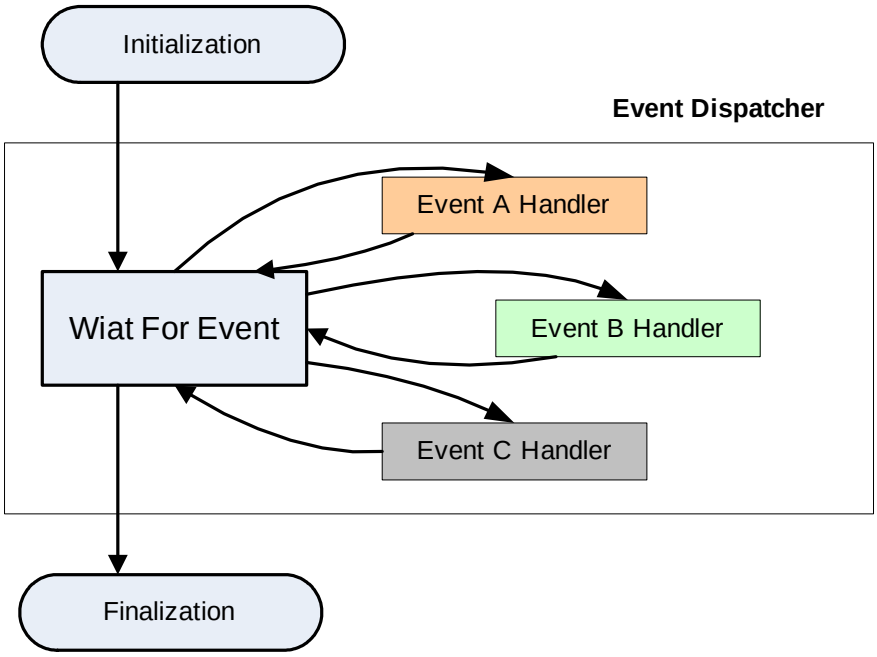

Figure 3-2 QualNet protocol model

At the heart of a protocol model is an Event Dispatcher, which consists of a 'Wait For Event' state and one or more 'Event Handler' states. In the 'Wait For Event' state, the protocol waits for an event to occur. When an event for the protocol occurs, the protocol transitions to the 'Event Handler' state corresponding to that event (e.g., when Event 1 occurs, the protocol transitions to the Event 1 Handler state). In this 'Event Handler' state, the protocol performs the actions corresponding to the event, and then returns to the 'Wait For Event' state. Actions performed in the 'Event Handler' state may include updating the protocol state, or scheduling other events, or both.

Besides the 'Event Dispatcher', the protocol finite state machine has two other states: the 'Initialization' state and the 'Finalization' state. In the 'Initialization' state, the protocol reads external input to configure its initial state. From a development perspective this is where all variable and data structure are created and initialized. The protocol then transitions to the 'Wait For Event' state.

The transition to the ‘Finalization’ state occurs automatically at the end of simulation. In the ‘Finalization’ state, protocol statistics collected during the simulation are saved or displayed.

3.1.3 Discrete-event Simulation in QualNet

QualNet is a discrete-event simulator. In discrete-event simulation, a system is modelled as it evolves over time by a representation in which the system state changes instantaneously when an event occurs, where an event is defined as an instantaneous occurrence that causes the system to change its state or to perform a specific action. Examples of events are: arrival of a packet, a periodic alarm informing a routing protocol to send out routing update to neighbours, etc. Examples of actions to take when an event occurs are: sending a packet to an adjacent layer, updating state variables, starting or restarting a timer, etc.

In discrete-event simulation, the simulator maintains an event queue. Associated with each event is its event time, i.e., the time at which the event is set to occur. Events in the event queue are sorted by the event time. The simulator also maintains a simulation clock which is used to simulate time. The simulation clock is advanced in discrete steps, as explained below. The simulator operates by continually repeating the following series of steps until the end of simulation.

- The simulator removes the first event from the event queue, i.e., the event scheduled for the earliest time.
- The simulator sets the simulation clock to the event time of the event. This may result in advancing the simulation clock.
- The simulator handles the event, i.e., it executes the actions associated with the event. This may result in changing the system state, scheduling other events, or both. If other events are scheduled, they may be scheduled to occur at the current time or in the future.

3.1.3.1 Events and Messages

In QualNet, the data structure used to represent an event is called a message. A message holds information about the event such as the type of event, and the associated data. In the context of QualNet, the terms event and message are often used interchangeably. There are two types of events: packet events and timer events.

Packet events are used to simulate exchange of data packets between layers or between nodes. Packet events are also used for modelling communication between different entities at the same layer. Timer events are used to simulate time-outs and are internal to a protocol.

3.1.3.2 Packet Events

Packet events are used to simulate transmission of packets across the network. A packet is defined as a unit of virtual or real data at any layer of the protocol stack. When a node needs to send a packet to an adjacent layer in the protocol stack, it schedules a packet event at the adjacent layer. The occurrence of the packet event at the adjacent layer simulates the arrival of the packet.

When a protocol residing at a particular layer at one node sends packets to the corresponding protocol at the same layer at another node, the packet is passed down through the protocol stack at the sending node, across the network, and then up through the protocol stack at the receiving node. At each level of the protocol stack at the sending node, header information is added to the packet as it is sent to the layer below. Each layer is responsible for sending the packet to its adjacent layer. At the receiving node, each layer strips off its header and sends the packet to the layer above, until the original packet is finally available to the receiving protocol. Figure 3-3 shows an example of this process for the case when the originating protocol resides at the Application Layer. The steps in this process are listed below.

- The originating protocol creates a new message by using the API MESSAGE_Alloc. The protocol creates the packet field of this message by using the API MESSAGE_PacketAlloc.
- The protocol puts the data to be sent to the receiving node in the packet field of the message, sets the other fields of the message appropriately, and sends the message to the next layer (Transport Layer in this case) by using the API MESSAGE_Send. Function MESSAGE_Send schedules a packet event for the next layer to occur after a delay that is specified as a parameter.
- When the packet is received by the Transport Layer protocol, the Transport Layer protocol appends its header to the packet by using the API MESSAGE_AddHeader and sets the header fields appropriately. The Transport Layer protocol then sends the resulting packet to the next layer in the stack by using the API MESSAGE_Send.

- The previous step is repeated at each layer in the protocol stack: Each layer adds its header to the packet and sends the resulting packet to the next layer.
- When the packet arrives at the Physical Layer of the source node, it schedules a packet receive event for the Physical Layer at the destination node.
- When a layer at the destination node receives a packet, it removes the corresponding header using the API MESSAGE_RemoveHeader, and sends the resulting packet to the next higher layer in the protocol stack using the API MESSAGE_Send.
- The previous step is repeated at each layer in the protocol stack: Each layer removes its header and sends the resulting packet to the next higher layer.
- When the packet arrives at the Application Layer at the destination node, the receiving protocol processes the packet and frees the message using the API MESSAGE_Free.

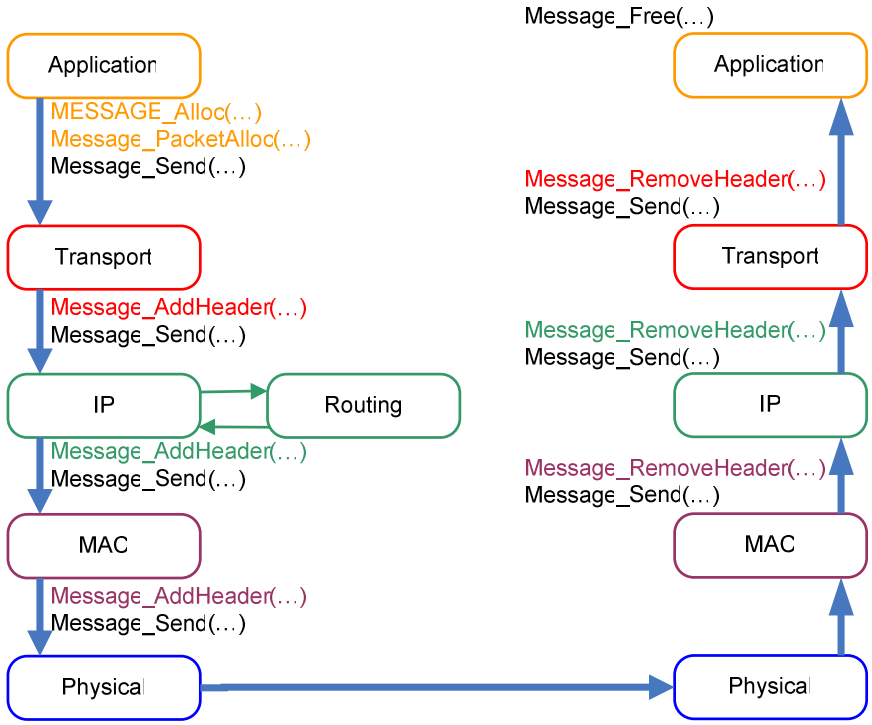

Figure 3-3 Life cycle of a packet. Packet originates on the left hand side and travels to the receiver on the right hand side of the diagram.

3.1.3.3 Timer Events

Timer events are used to perform the function of alarms. They essentially allow an application to schedule events for itself at a future time. Periodic alarms are implemented by re-setting the timer event after it has occurred. Timer events are set and received within a protocol and they do not travel through the protocol stack.

3.1.4 QualNet Simulator Architecture

As discussed in Section 3.1.2, a protocol model in QualNet has three components: Initialization, Event Handling, and Finalization. Each of these functions is performed hierarchically: first at the node level, then at the layer level, and finally at the protocol level. The following sections describe the hierarchy of these three functions.

3.1.4.1 Initialization Hierarchy

At the start of simulation, each node in the network is initialized. Function PARTITION_InitializeNodes is the function which initializes nodes. Function PARTITION_InitializeNodes initializes the layers of the protocol stack running at every node by calling the initialization function for each layer. The layers are initialized in a bottom-up order, starting from the bottom-most layer. Some layers, such as the MAC Layer, are initialized globally, while the other layers are initialized one node at a time. For example, function MAC_Initialize initializes the MAC Layer for all nodes, while function TRANSPORT_Initialize initializes the Transport Layer at a given node. There are two initialization functions for the Application Layer: one for traffic-generating protocols and the other for routing protocols running at the Application Layer. Function APP_Initialize initializes the Application Layer routing protocols for a given node, and function APP_InitializeApplications initializes the Application Layer traffic-generating protocols at all nodes.

Each layer initialization function, in turn, calls an initialization function for each protocol running at that layer. For example, function TRANSPORT_Initialize, calls the initialization functions for the TCP and UDP protocols, TransportTcpInit and TransportUdpInit, respectively.

The initialization function of a protocol creates and initializes the protocol state variables, as well as the protocol statistics variables. For example, the TransportUdpInit function creates the UDP state variable udp, which is a data structure of type TransportDataUdp. If UDP statistics collection is enabled, TransportUdpIn it also creates and initializes the UDP statistics variable, which is a data structure of type TransportUdpStat.

3.1.4.2 Event Handling Hierarchy

When an event occurs, the QualNet kernel gets a handle to the node for which the event is scheduled. It then calls a dispatcher function, NODE_ProcessEvent, This function determines the layer for which the event has occurred and calls the event dispatcher function for the appropriate layer, e.g., if the event is for the Application Layer, NODE_ProcessEvent calls the Application Layer event dispatcher function, APP_ProcessEvent.

The event dispatcher function for a layer determines the protocol for which the event has occurred, and calls the event handler for that protocol. For example, when an event for the Bellman-Ford protocol occurs, the Application Layer dispatcher function, APP_ProcessEvent, calls function RoutingBellmanfordLayer, which is the event handler for the Bellman-Ford protocol.

The protocol event dispatcher, like the other dispatcher functions, consists of a switch statement. It calls the event handler function for the event that has occurred. An event handler is specific to an event and performs the required actions on the occurrence of that event. For example, the Bellman-Ford dispatcher function, RoutingBellmanfordLayer, calls function HandleFromTransport when an event of type MSG_APP_FromTransport occurs. MSG_APP_FromTransport indicates that a packet has been received from the Transport Layer, and function HandleFromTransport performs the actions required to handle the received packet.

3.1.4.3 Finalization Hierarchy

At the end of simulation, the finalization function for each protocol is called to print the protocol statistics. Like the initialization and event handling functions, the finalization function is called hierarchically. The node finalization function, PARTITION_Finalize, calls the finalization function for each layer in the protocol stack running at each node. For example, MAC_Finalize is the finalization function for the MAC Layer.

The finalization function for a layer calls the finalization function for each protocol running at that layer. For example, consider the MAC Layer finalization function, MAC_Finalize, which calls the finalization function for the MAC protocol running at that interface, e.g., if the 802.11a protocol is running at an interface, MAC_Finalize calls the 802.11a finalization function Mac802_11aFinalize.

The finalization function for a protocol prints the statistics for the protocol if statistics collection is enabled for the layer in which the protocol resides. For example, function Mac802_11aFinalize, calls the function to print 802.11a statistics, Mac802_11aPrintStats, if statistics collection is enabled for the MAC Layer.

3.2 Fixed WiMAX MAC Layer Model

The simulation model used in the work described in the following chapters was developed in accordance with the infrastructure provided by QualNet 3.8, QualNet 3.9.5 and the IEEE 802.16 – 2004 standard. The existing Application, Network and Transport layers were used, with additions due to the new protocol. The mesh mode of operation defined in the standard is not considered within the scope of this work and hence only the PMP mode is coded in to the model. In the following sub sections the detailed organisation and operation of the model is provided.

3.2.1 Basic Functions

Initialization is the first basic function performed by a SS. The procedure can be divided into the following phases:

a) Obtain transmit parameters (from UCD message)
b) Perform ranging and registration
c) Perform co-location with BS
d) Establish IP connectivity
e) Transfer operational parameters
f) Set up connections

The SS in the model do not scan for DL channel parameters as these values are used to configure the SS in the configurations script. Prior knowledge is assumed. Initial transmit parameters are also considered known. All SS begin communication with the BS using the lowest UIUC, DIUC and move to higher burst profile after ranging. Once the SS has received the DCD, UCD, DL-MAP and UL-MAP messages it begins initial ranging. The process flow diagram for the ranging process is given in Figure 3-4.

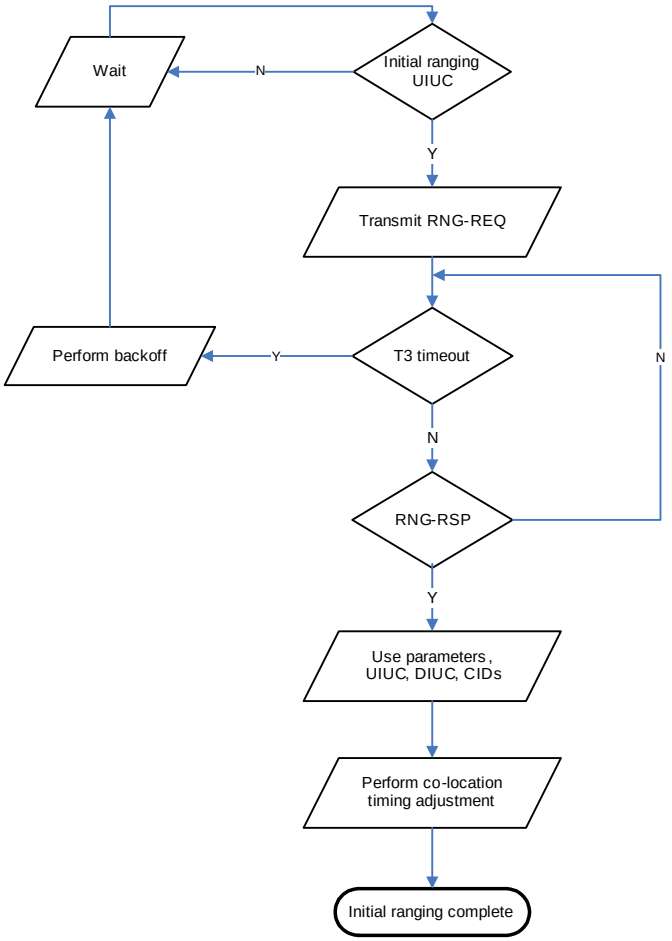

Figure 3-4 Initial ranging process for a simulated SS

Co-location of all SSs with the BS is required to maintain synchronization of UL transmission. This is implemented as part of the ranging process. We allocate two OFDM symbols for the initial ranging slots to accommodate delay variations of up to 10 µs. The BS calculates the delay of the received RNG-REQ from the expected RNG-REQ slot start time and informs the SS of the required timing compensation. We have modified the RNG-RSP MAC management message to include a field called TIME_LAG which is used to facilitate co-location adjustments. Once the RNG-RSP is received by the SS the recommended compensation is used on all UL transmissions without exception.

Registration is the process of obtaining a Secondary Management CID, and this is combined with the ranging process for simplicity. The RNG-RSP from the BS contains the following CIDs.

- Basic CID
- Primary Management CID
- Secondary Management CID
- Transmit CIDs

The number of Transmit CIDs allocated to a SS is determined by the configuration script and is user configurable.

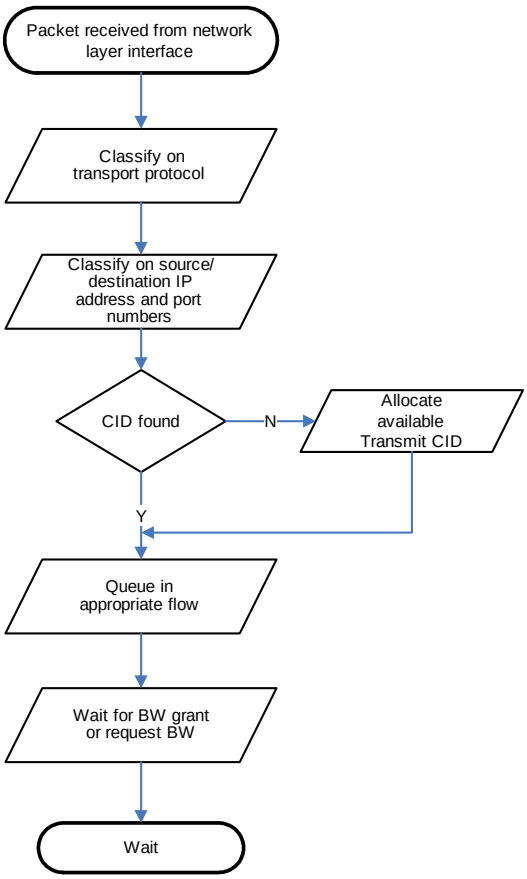

Figure 3-5 UL packet classification and allocation of CIDs by the SS

3.2.2 Connections

The MAC is connection-oriented, as discussed in chapter 2. For the purposes of mapping to services on SSs, and associating varying levels of QoS, all data communications are in the context of a connection. Service flows may be provisioned when an SS is installed in the system. Shortly after SS registration, connections are associated with these service flows (one connection per service flow) to provide a reference against which to request bandwidth. Additionally, new connections may be

established when a customer's service needs change. A connection defines both the mapping between peer convergence processes that utilize the MAC and a service flow. The service flow defines the QoS parameters for the PDUs that are exchanged on the connection. The concept of a service flow on a connection is central to the operation of the MAC protocol. Service flows provide a mechanism for uplink and downlink QoS management. In particular, they are integral to the bandwidth allocation process. An SS requests uplink bandwidth on a per connection basis (implicitly identifying the service flow). Bandwidth is granted by the BS to an SS as an aggregate of grants in response to per connection requests from the SS.

3.2.3 Service Class Modelling

The standard defines four service classes which can be grouped into three basic types. These are UGS, which supports real-time VoIP, rtPS/nrtPS which are polled services supporting anything from real-time video to file transfer, and BE, which is the lowest priority class. UGS, nrtPS and BE classes are represented in the model.

3.2.3.1 UGS

The UGS service class is modelled as per the standard with a modification for better latency control. The grant management subheader for the UGS contains 14 reserved bits, of which 5 are used to define two new fields as shown in Table 3-1

Table 3-1 Modifications to UGS Grant Management Subheader are shown shaded

Syntax	Size (bits)
SI (slip indicator)	1
PM (poll me bit)	1
FLI (flow lag indicator)	1
FL (flow lag compensation)	4
Reserved	9

Once the FLI flag is detected, the BS will attempt to reduce the lag for the next UGS grant, as per the FL value. UGS grants do not require any other active maintenance, since they have a constant bandwidth allocated every time.

Retransmission strategies for UGS are not defined by the standard. A fast retransmission strategy is devised and coded into the model. When a UGS packet is not received on the UL, the BS will grant a retransmission slot in the very next frame.

A specific UIUC (U_ReTxLastPacket = 17) is defined to indicate to the SS that a retransmission is needed. If the packet cache has already been cleared, a dummy packet consisting of only a generic MAC header is transmitted, which prevents the BS from providing any more retransmission opportunities.

Traditional services such as Channelized T1 services require some maintenance due to the dynamic (but relatively slowly changing) bandwidth requirements if compressed, coupled with the requirement that full bandwidth be available on demand.

3.2.3.2 nrtPS

The nrtPS service class is modelled in detail in the simulation model. We use nrtPS to service BE traffic in Chapter 7. The process flow implemented is given in Figure 3-6 below.

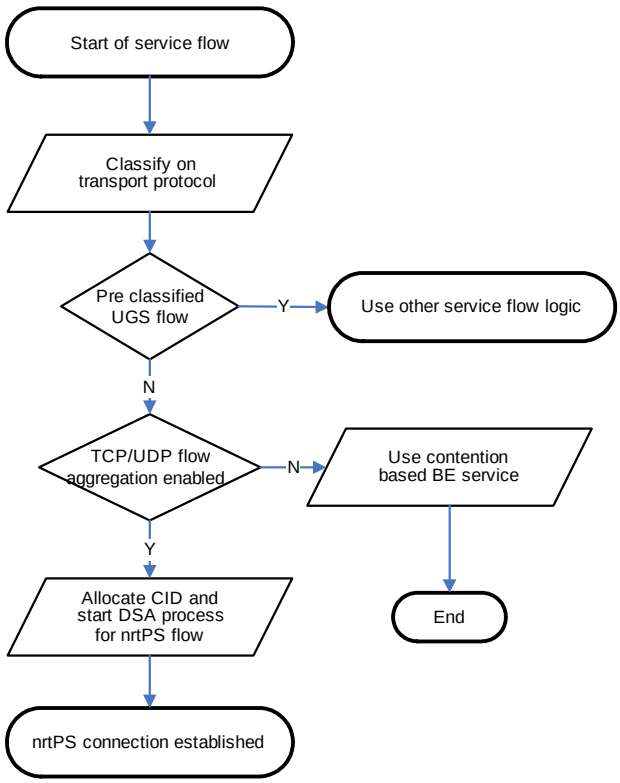

Figure 3-6 UL packet classification and allocation to nrtPS flow by SS

When a new flow classified as BE starts, it will be aggregated in to the existing nrtPS service by the SS in our implementation. Further enhancements have also been made to this service and are described in detail in Chapter 7.

3.2.3.3 BE

Traditional BE service is based on broadcast or multicast contention, which in principle are identical in terms of the contention resolution procedure. Hence we have implemented the broadcast contention scheme and omitted the optional multicast contention.

The parameters for the contention resolution phase are communicated to all SSs by the BS, through the UCD MAC management message. The simulation model adheres to the specification of the standard in every way. We have included additional retransmission capabilities at the MAC layer, for BE traffic similar to that provided for UGS and nrtPS services. The BS can provide explicit retransmission BW to the SS, using the UIUC of 17. This capability can be enabled or disabled using the configuration script which is input to the simulator.

As described in the previous section nrtPS may also be used to service BE traffic with a higher level of efficiency and at the same time, better scalability.

3.2.3.4 Policing, Shaping and Scheduling

These three aspects of the protocol are not defined by the standard and implementation specific. The IP layer of QualNet provides all the important queuing and scheduling methods. In addition to that, a class based scheme has been implemented at the MAC layer of the model.

UGS flows having the highest priority and strictest delay bounds are given strict priority and scheduled at the root level. The rates are policed tightly, with only a burst leeway of one UGS interval allowed.

nrtPS has the next highest priority and is queued into a separate queue. It is serviced using a weighted fair scheduler with configurable weights.

BE traffic gets the lowest priority with no rate guarantee, but high buffer length for burst tolerance. Under normal circumstances BE is given any remaining BW after higher priority flows have been serviced. When nrtPS is used for BE traffic the BE class is disabled in our implementation.

3.2.4 Cross Layer Communication

By default here is no cross layer communication defined in the standard for the MAC layer. However, in the simulation model, in order to facilitate the proposed

enhancements, we have enabled information exchange between the Physical layer (PHY) and the MAC layer and also, the Application layer and the MAC layer.

On the DL the SS polls the PHY for signal quality measurements, and steps through burst profiles, until the signal-to-noise (SNR) is above a predefined threshold. A certain amount of hysteresis can be included, by setting the thresholds at which burst profiles are stepped up and stepped down. These are set for the system as a whole or for each individual node in the configuration script.

The proposed scheme of 'Optimal packetization intervals for VoIP', Chapter 4, requires dynamic updating of packetization intervals. This needs the application layer to communicate with the MAC layer. We have coded this functionality so that it may be enabled when required by the user.

3.2.5 Approximations

Ranging is a 3-way hand shake process which the SB may recommend parameters which the SS may reject. In the simulation model the process is approximated by a 2-way handshake in which the SS always accepts the recommended parameters.

Any UL service with higher priority that BE needs to compete a Dynamic Service Addition/Change/Deletion process. These processes are defined as 3-way handshakes between the SS and BS, consisting of a request, a response and an acknowledgement. The implementation approximates this with 2-way handshakes to reduce the complexity of the SS and BS state machines. All connection management functions are supported through the use of static configuration and dynamic addition, modification, and deletion of connections.

The service-specific CS resides on top of the MAC CPS and utilizes, via the MAC SAP, the services provided by the MAC CPS. The CS performs the following functions:

- Accepting higher-layer protocol data units (PDUs) from the higher layer
- Performing classification of higher-layer PDUs
- Processing (if required) the higher-layer PDUs based on the classification
- Delivering CS PDUs to the appropriate MAC SAP
- Receiving CS PDUs from the peer entity

Currently, two CS specifications are provided: the asynchronous transfer mode (ATM) CS and the packet CS. The model does not include the ATM CS functionality. The packet CS is not defined as a separate layer, which uses an API to access the MAC. It is coded integrated into the MAC CPS itself. For example PDU classification is considered part of the MAC CPS itself.

3.2.6 Exclusions

As stated in the previous subsection there is no ATM support in the simulation model, and only the packet CS is included which is optimized for IP packets.

The model allows a multi-cell environment to be simulated, but no handover process is defined for a SS to perform intercell movements. SS have fixed frequency settings so cannot ‘scan’ the frequency band to find other active BSs in the vicinity.

The standard includes a Security Sublayer which provides subscribers with privacy across the fixed broadband wireless network. It does this by encrypting connections between SS and BS. The BS protects against unauthorized access to these data transport services by enforcing encryption of the associated service flows across the network. Privacy employs an authenticated client/server key management protocol in which the BS, the server, controls distribution of keying material to client SS. Additionally, the basic privacy mechanisms are strengthened by adding digital-certificate-based SS authentication to its key management protocol. The Security Sublayer is not modelled as part of the simulator, and the connections are not encrypted.

3.3 Fixed WiMAX OFDM Physical Layer Model

We have used a 256 subcarrier OFDM PHY with the following user configurable parameters.

- Frame duration
- OFDM symbol duration
- Bits per OFDM symbol for the available burst profiles and number of data subcarriers.
- SNR thresholds and error rates for available burst profiles
- Operating frequency and bandwidth

- Radio propagation method, two ray ground with shadowing and fading.

Bit Error Rates (BER) are read from waterfall curves provided by QualNet as part of the standard simulation package.

While the PHY is an important aspect the simulator it is not the object under test, so we maintain a consistent PHY for all simulations which is a general representation of the OFDM PHY.

3.3.1 Approximations

When a certain packet drop rate is needed to be simulated the packet drops are done at the MAC layer, at the interface between the PHY and the MAC. The random drop or bursty drop error state machines are introduced at this point which is not what would happen in reality.

At the PHY the entire packet is ‘received’, and then it is decided whether or not the packet is in error. The BER values are adjusted for the type of error protection and coding scheme used.

During packet reception the receiver spends the duration of the packet in a ‘receive’ state. If another reception event is triggered during this both packets are determined to be in error. This is the method for collision simulation.

We have used a zero processing delay at the PHY, which implies that packets do not spend any time at that layer. This delay is user configurable through the configuration script.

3.3.2 Exclusions

The standard includes five possible PHY configurations (depending on use, such as backhaul, point-to-multipoint etc) which are

- Wireless MAN Single Carrier (SC) PHY for line-of-sight
- Wireless MAN Single Carrier (SCa) PHY for non line-of-sight
- Wireless MAN OFDM
- Wireless MAN OFDMA
- Wireless HUMAN

Only the Wireless MAN OFDM PHY is modelled using a TDD frame structure. The relative performance of the PHY types is beyond the scope of this work. The PHY has been held constant for all MAC layer simulation scenarios.

3.4 Conclusion

A simulation model was created to represent the IEEE 802.16 - 2004 standard which encompasses the required features. This model integrates in to the infrastructure of the QualNet simulation package, so that pre-existing upper and lower layers of the protocol stack can be reused. All nodes go through the states of initialization, life time and finalization procedures.

The packet CS is coded to be suited for the all IP architecture we envision WiMAX to have. Basic functionality such as ranging is accurately modelled as well as the service classes UGS, nrtPS and BE.

Several enhancements, which are described in the following four chapters have been included in the simulation model. These include, dynamic packetization adjustment, latency aware retransmission schemes, scheduling and queuing for different services, and traffic class substitution for nrtPS.

The PHY is also modelled using good approximations for the Wireless MAN OFDM scheme. Where ever parameter values are provided they are extracted from the standard, a published source or libraries provided in the simulation package.

Chapter 4

Optimal Packetization Interval for VoIP

The emerging popularity of VoIP in the enterprise market coupled with the rise of WiMAX has allowed for an alternative to the PSTN for voice transport. In this chapter we discuss the mechanics of VoIP and how well equipped WiMAX is to handle it. Then we propose enhancements in terms of Quality of Service as well as efficiency so that the limited wireless resource can be better utilized. Firstly we consider the way voice calls are transported across the traditional fixed line telephone network. In the PSTN there are three main components to consider. They are: (1) transport, the transportation of conversations from one Central Office (CO) to another; (2) switching, the switching or routing of calls in the PSTN via a telephone switch contained in the CO; and (3) access, the connection between the switch in the CO and the subscribers telecommunications device, Figure 4-1.

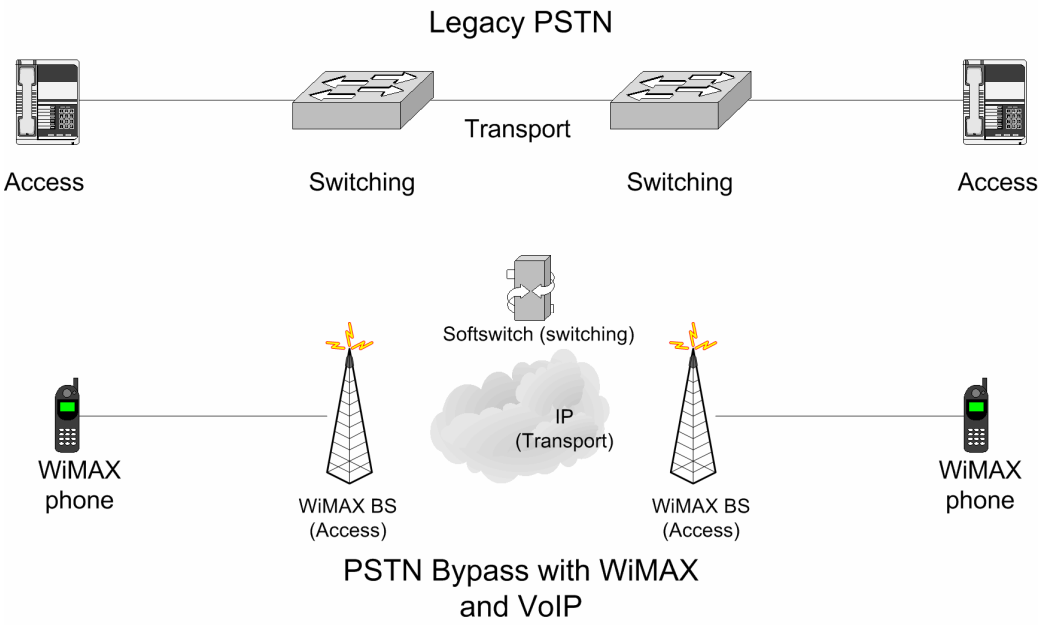

Figure 4-1 A broadband wireless alternative to PSTN based on WiMAX

As can be seen above, using WiMAX the relatively inexpensive IP backbone can be used with Softswitch technologies. So WiMAX presents a bypass technology to the telco's copper wire access. VoIP frees the voice stream from the confines of a voice-specific network and its associated platforms. VoIP can be received and transmitted via PC's, laptops, Wi-Fi and any IP capable handsets (Ohrtman 2005). In rural areas where no infrastructure currently exists or it is prohibitively expensive to establish wired infrastructure a WiMAX based backhaul and access system could prove to be a very viable solution. Providers can bypass the Internet when transporting the voice data from the access network to the softswitch or to the PSTN if it is the destination, so QoS can be guaranteed (Goldman 2005).

4.1 Analysis of Packetization Interval

In a VoIP application the stream of sampled voice data is broken into small segments each of a fixed length. These segments are then passed through a CODEC to convert them into a VoIP payload. The time duration of each segment is known as the packetization interval, t_{pkt}, Figure 4-2. This voice payload is carried using RTP which adds a 12 Byte header. This then becomes the payload of the transport layer, UDP, which adds an 8 Byte header. Finally the IP layer receives the UDP packet, and adds an IP header of 12 Bytes.

As the size of the voice payload is determined by the packetization interval used by the VoIP application, we analyze the effect the packetization interval has on packet loss rate, retransmit limit, packet latency and system resource usage.

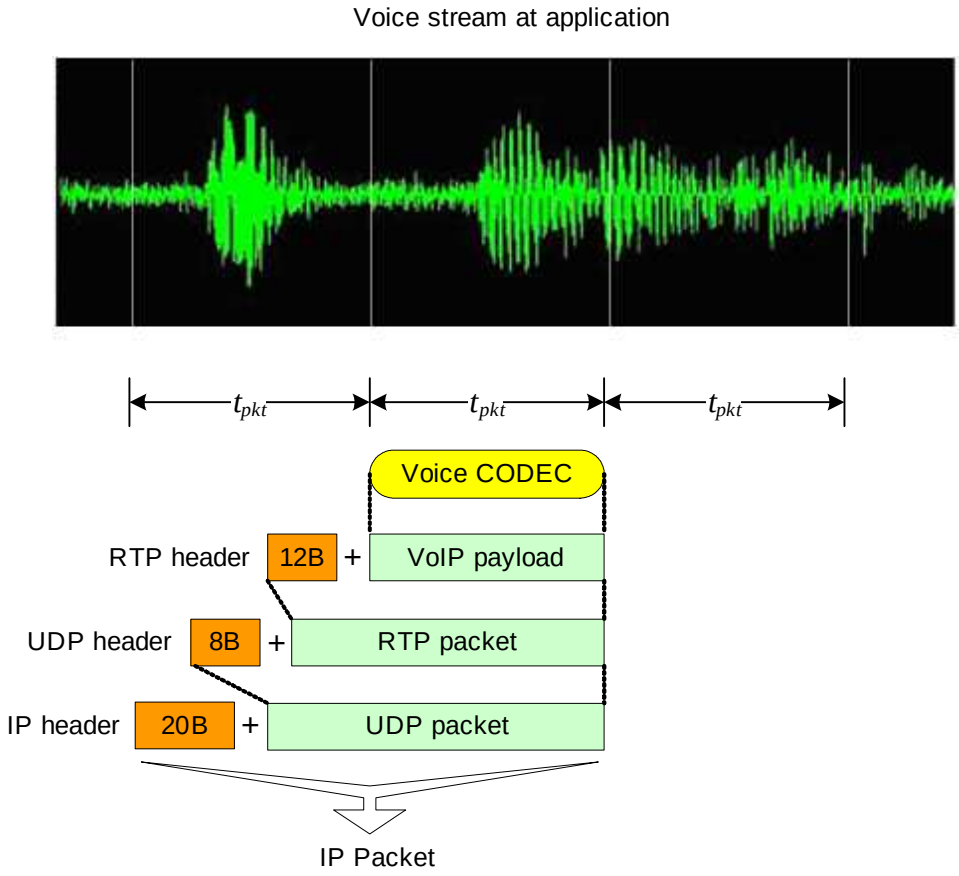

Figure 4-2 The process of converting a voice stream into IP packets. The header sizes are given in Bytes. The headers are appended to the previous layer payload and passed on.

4.1.1 Packet Loss Rate

Consider a VoIP application which produces a voice data stream of r bits-per-second (bps). The overhead due to headers, $OH_{headers}$ is the sum of the RTP, UDP, IP and MAC layers headers in bits. n_{pkt} is the packet size as seen at the MAC layer. t_{pkt} is the packetization interval of the VoIP application.

$$n_{pkt} = r \cdot t_{pkt} + OH_{headers} \tag{4.1}$$

$$n_{pdu} = \left\lceil \frac{n_{pkt}}{n_{bps}} \right\rceil \cdot n_{bps} \tag{4.2}$$

n_{pdu} and n_{bps} are the sizes in bits, of the PDU and, physical layer OFDM symbol respectively. $\lceil x \rceil$ represents the *ceil* function, which rounds any real number upwards towards the closest integer value. A PDU consists of an integral number of OFDM symbols. On the UL there can be no packet combining between multiple stations, hence the above constraint.

The symbol error rate (probability of an OFDM symbol being classified by the receiver's physical layer as erred), *SER*, is given by (4.3). Here *j* denotes the number of bit errors, and *m* denotes the maximum number of bit errors which can be tolerated or corrected. *BER* represents the bit error rate.

$$SER = 1 - \sum_{j=0}^{m} \binom{n_{bps}}{j} BER^{j} (1 - BER)^{n_{bps} - j} \tag{4.3}$$

PER is the packet error rate and n_{spp} is the number of symbols per packet.

$$PER = 1 - (1 - SER)^{n_{spp}} \tag{4.4}$$

The packet is considered lost when the retransmit limit, *n*, has been exceeded. This probability is P_{loss}. At this point there are no constraints placed on *n* other than being a non-negative integer value, thus.

$$P_{loss} = PER^{n+1} \tag{4.5}.$$

4.1.2 Bandwidth Usage

The total bandwidth used by the application at the MAC layer is considered here. This is the sum of the actual payload being delivered plus overheads. Overheads can be separated into the following components.

- ❖ Overheads due to lower layer headers which are pushed as the payload travels down the network stack.
- ❖ Overheads due to unsuccessful retransmissions of the same payload.
- ❖ Overhead due to the PDU size not being an integer multiple of OFDM symbol size. This means the unused part of the last symbol will be padded and wasted.

The total overhead due to retransmissions and upper layer headers is denoted by OH_{tot}. The mean value of OH_{tot} is give by (4.6), where *n* is the MAC layer retransmit

limit for this particular traffic class. The value of *n* has an upper bound which is defined in section 4.1.3.

$$\begin{aligned} OH_{tot} &= \sum_{i=1}^{n+1} \left\{ \left(i \cdot n_{pdu} - payload \right) \cdot PER^{i-1} \cdot \left(1 - PER\right) \right\} + n_{pdu} \cdot (n+1) \cdot PER^{n+1} \\ &= \sum_{i=1}^{n} \left\{ i \cdot n_{pdu} \cdot PER^{i-1} \cdot \left(1 - PER\right) \right\} + (n+1) \cdot n_{pdu} \cdot PER^{n} - payload \cdot \left(1 - PER^{n+1}\right) \end{aligned} \tag{4.6}$$

The BW used for the overheads can be obtained by averaging the total overhead over the interval:

$$OH_{bw} = \frac{OH_{tot}}{t_{pkt}} \ . \tag{4.7}$$

From the result in (4.6), by omitting the term including the payload we can obtain TB_{avg}, the mean total number of bits on the airlink during one transmission cycle (including all retransmissions), as:

$$TB_{avg} = \sum_{i=1}^{n} \left\{ i \cdot n_{pdu} PER^{i-1} (1 - PER) \right\} + (n+1) n_{pdu} PER^{n} \ . \tag{4.8}$$

The total BW requirement for the flow equals the total bits used during the interval TB_{avg}, averaged over t_{pkt}, as:

$$TB_{bw} = \frac{TB_{avg}}{t_{pkt}} \ . \tag{4.9}$$

If E_f is the efficiency of the system, then

$$E_f = \frac{payload \cdot \left(1 - PER^{n+1}\right)}{TB_{avg}} \ . \tag{4.10}$$

The payload has been scaled by the factor $\left(1 - PER^{n+1}\right)$ to account for PDUs dropped after all retransmit attempts have been unsuccessful.

4.1.3 MAC Retransmit Limit

If an UL packet is corrupted beyond repair, the BS will discard it and give the SS another BW allocation in the following frame. This method circumvents the need for the BS to provide explicit feedback for every packet. It also makes it possible to retransmit in consecutive frames and not alternating frames. By defining an Uplink Interface Usage Code (UIUC) specifically for retransmissions, the BS can implicitly request a retransmission of the last packet sent on a given Connection Identifier

(CID). An Information Element (IE) will be included in the UL-MAP with the relevant CID and UIUC. If the SS has already refreshed its packet cache, it will transmit a dummy packet with only a generic MAC header and no payload. This scheme is explained in detail in section 4.2, in Proposed Implementation Scheme.

If the maximum number of MAC retransmissions, based on BW or delay constraints, is defined as n_{max}, it is calculated as,

$$n_{retx} = \min\left(\left\lfloor \frac{t_{pkt}}{T_f} \right\rfloor - 1 \, , \, n_{\max} \right) . \tag{4.11}$$

$\lfloor x \rfloor$ represents the *floor* function, which rounds any real number down towards the closest integer value. If no such limitation exists, retransmit opportunities can be granted until the next UGS BW grant is scheduled. A UGS grant will be given periodically so the maximum time available for retransmission is t_{pkt} -T_f. If it cannot be sent successfully within this time period the packet needs to be dropped.

4.1.4 Latency in Packet Transmission and Delivery

The average latency of a ***successfully delivered*** packet, L_{avg} is given by (4.12). This includes an additional component t_{pkt} in the summation which accounts for the lag due to packetization.

$$\begin{aligned} L_{avg} &= \frac{\sum_{i=0}^{n_{retx}} \left[\left(t_{pkt} + \frac{t_{pkt}}{2} + i \cdot T_f \right) \cdot (1 - PER) \cdot PER^i \right]}{1 - PER^{n_{retx}+1}} \\ &= 1.5 \cdot t_{pkt} + \frac{\sum_{i=0}^{n_{retx}} \left[i \cdot T_f \cdot (1 - PER) \cdot PER^i \right]}{1 - PER^{n_{retx}+1}} \end{aligned} \tag{4.12}$$

T_f is the frame duration. For low BER values ($<10^{-4}$) latency is approximately equal to $1.5 \cdot t_{pkt}$. Even at higher BERs, since we are only considering the latency of successfully delivered packets, the average latency will only increase by at most n_{retx} frame durations ($n_{retx} . T_f$).

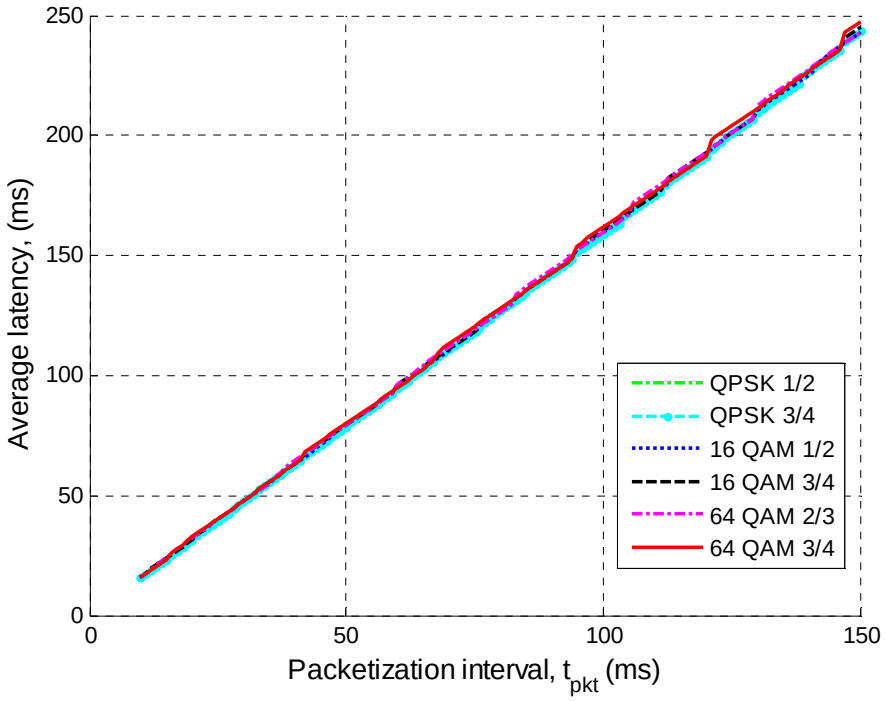

Figure 4-3 Average latency for all modulation schemes at a BER = $10^{-3.5}$. All the curves follow each other closely.

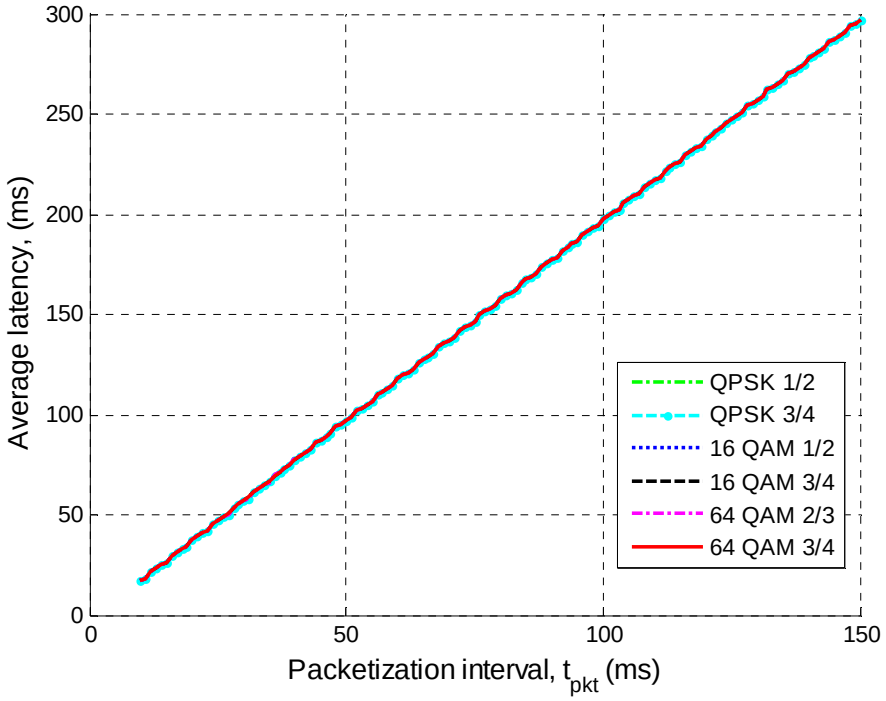

Figure 4-4 Average latency for all modulations schemes at a BER = $10^{-2.5}$. All the curves follow each other closely.

Since talkspurts can begin randomly, the BW grant from the BS may not be synchronized with the VoIP packet becoming available at the MAC layer. There may be a lag between these two events, which could be anywhere in the range (0, t_{pkt}]. Hence, an average latency of half the interval, $t_{pkt}/2$, is used. The packetization interval is added on to this latency, to give a total mean latency of 1.5 t_{pkt} , as given in (4.12).

Figure 4-3 and Figure 4-4 show the latency for the different modulation schemes. There is a one-to-one mapping between the modulation scheme and the burst profile as previously discussed. Compared to higher burst profiles there is a minute increase in latency for lower (more robust) burst profiles due to requiring multiple symbols. This is not apparent in the graphs as all the curves seem to coincide. For example, if using the highest burst profile, an entire PDU can be transported in one OFDM symbol. This would create a latency of a single OFDM symbol duration. On the other hand, if the lowest burst profile was used, multiple symbols would be required, which implies multiple OFDM symbol durations. The symbol duration is a value in the tens of micro seconds, typically 10~25 micro seconds. Hence the contribution to latency is relatively minimal, which causes the curves to apparently coincide.

4.2 Proposed Implementation Scheme

4.2.1 UGS Retransmission Strategy

The mechanics of the UGS scheduling class as defined by the standard (IEEE 2005) have been detailed in Chapter 2. There is no method to request the SSs to retransmit an erroneous packet, except for ARQ, which is not used for UGS connections. This implies UGS connections do not facilitate the recovery of lost packets on the UL. In the context of real-time traffic, often it is not cost effective to attempt recovery due to the introduced latency in doing so (Chia-Hui, Ray et al. 2003; Miki, Atarashi et al. 2003; Gurbuz and Ayanoglu 2004; Gyung-Ho and Dong-Ho 2004; Zhi and Jong-Moon 2004; Sik, Gyung-Ho et al. 2005). In order to introduce some probability of

packet recovery, while using a minimum overhead, we propose an implicit method for retransmit notification, shown in Figure 4-5.

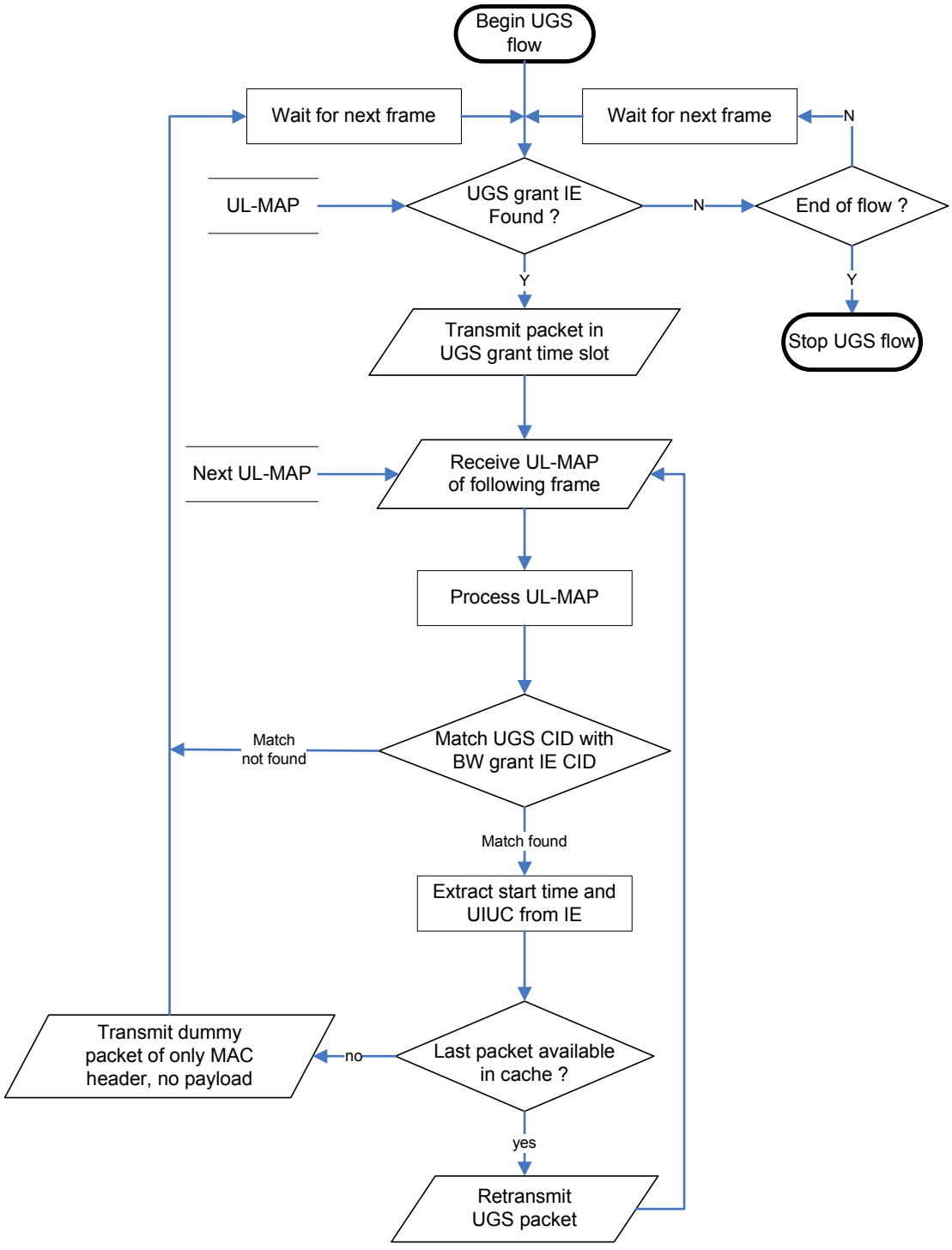

Figure 4-5 Mechanism used by the BS, to notify the SS when to retransmit erred UGS packets.

Once a SS transmits a packet to the BS on a UGS connection, the BS PHY checks for errors. If error free or recoverable the packet is passed up the stack. Next the MAC layer checks for errors in the PDU. If any errors are found, the packet will be discarded. All received packets are checked against a list of expected packets. Missing UGS packets will be recognized. These will be given the highest priority and be retransmitted in the immediately following frame. The BS includes in the UL-MAP, a BW grant IE directed at the CID of the UGS connection. This IE will have a flag, which tells the owner of the CID to retransmit the last sent packet on that connection. Once the UL-MAP is processed by the relevant SS, it will retransmit the packet in the given time slot. If no such IEs (retransmit grants to the UGS connection CID) are received before the next scheduled UGS grant, the SS assumes successful packet transmission. If for some reason, the packet in question is already flushed from the cache, when the retransmit opportunity is received, a dummy packet consisting of a MAC header and zero length payload is sent by the SS. On receipt of this dummy packet, the BS will not attempt any more retransmissions of that particular packet.

The OH associated with this method is 48 bits, for the IE in the UL-MAP, for every retransmit attempt (IEEE 802.16 WG 2004). We have also preserved the fundamental semantics of the UGS schedule type by not allowing the SS to actively request BW grants for retransmissions.

There is no explicit notification for successful reception from the BS to the SS, i.e., no positive feedback is given to the SS. It could be made mandatory to keep the transmitted packet until the next UGS grant is received for the next packet.

4.2.2 Usability Factor, K

In order to fulfill the QoS requirements of the flow and operate within the BW constraints of the network, three conditions need to be satisfied.

A. The packet loss rate P_{loss} of the flow must be less than the maximum allowable packet loss rate defined by the service level agreement between the provider and the subscriber. Excessive packet losses cause the conversation to be unintelligible to the listener.

B. The average latency L_{avg} must be within the upper bounds of the maximum latency. This could also be based on service level agreements or dynamically selected based on the one way delay to the other end point.

C. The BW utilized by the flow must be less than a predefined maximum allowable BW, as agreed upon by the service provider.

These requirements are summarized in binary function form as follows:

- A = (1,0 | $P_{loss} < P_{loss,max}$) , where $P_{loss,max}$ is the maximum P_{loss}
- B = (1,0 | $L_{avg} < L_{max}$) , where L_{max} is the maximum tolerated latency
- C = (1,0 | $BW < BW_{max}$) , where BW_{max} is the maximum allowed BW

Here, (x,y | z) implies a return value of x,when the condition z is true, and y,when the condition z is false. Losses cause a drop in VoIP quality and a drop rate greater than 1% is hard to conceal (Hattingh and Szigeti 2004). So $P_{loss,max}$ can be assumed to be 0.01.

Latency can cause voice quality degradation if it is excessive. The goal commonly used in designing networks to support VoIP is the target specified by ITU standard G.114. This states that 150 ms of one-way, end-to-end (from mouth to ear) delay ensures user satisfaction for telephony applications. This maximum delay should be apportioned to the various components of network delay (propagation delay through the backbone, scheduling delay because of congestion, and access link serialization delay) and service delay (because of VoIP gateway codec and de-jitter buffer). However, using a UGS flow, the maximum latency will be bounded as explained above. When communicating with another SS in the same cell or in a cell which is part of the WiMAX network, the latency will be lower than when using the public internet, for example, for international VoIP calls.

The BW_{max}, which can be allocated to a flow, very much depends on the SLA with the SS and also on the traffic mix and prevalent load dynamics.

We define a metric called '**Usability Factor**', K, which combines all three of the above requirements to give a value between 0 and 1. The closer it is to 1, the more suitable (or usable) the t_{pkt} is. The time scale used for K is the same as for t_{pkt}. A value of 0 implies that the interval cannot satisfy one or more of the three requirements, and should not be used.

First we define a logical function S, where A, B and C are the three requirements from above. Then,

$$S = (A)\,and\,(B)\,and\,(C)\ . \tag{4.13}$$

$$S = \begin{cases} 1 & , \text{if } (P_{loss,\max} > P_{loss}) \,\&\, (L_{\max} > L_{avg}) \,\&\, (BW_{\max} > BW) \\ 0 & , \text{if } (P_{loss,\max} < P_{loss})\ or\ (L_{\max} < L_{avg})\ or\ (BW_{\max} < BW) \end{cases} \tag{4.14}$$

The Usability Factor is defined as follows:

$$\begin{aligned} K &= S \times abs\left[\frac{\text{product of current conditions}}{\text{product of best possible conditions}}\right] \\ K &= S \times abs\left[\frac{(P_{loss,\max} - P_{loss})(L_{\max} - L_{avg})(BW_{\max} - BW)}{(P_{loss,\max} - 0)(L_{\max} - 1.5 \cdot t_{pkt})(BW_{\max} - r)}\right] \end{aligned} \tag{4.15}$$

As stated before, r is the nominal bitrate of the VoIP application. The denominator of (4.15) is the product of the three constraints, less the best attainable conditions. *abs* is the absolute value. S is given in (4.13) and (4.14) above. S has the effect of masking or selecting only the areas of the function where all three conditions are met. In any regions where at least one of the conditions is not true, $S = 0$. There is a possibility that L_{max} could be equal to $1.5 \cdot t_{pkt}$. If this condition is true, then $K = 0$.

4.2.3 Lookup Table Creation and Usage

The BS needs to have access to the K values so that it can select the most suitable value for the packetization interval. The K value can either be calculated in real-time when flow parameters are negotiated, or a set of pre-calculated values can be held in memory, at the BS to be used as a lookup table. The usability factor can be directly used in the lookup tables, or it can be quantized to a set of discrete levels to further simplify the process.

4.2.4 Dynamic Service Addition/Change Process for Setting/Updating t_{pkt}

Here we describe the proposed method for negotiating and selecting an optimal packetization interval, for a VoIP flow.

At the start of the service flow the initiating SS will send a Dynamic Service Addition Request (DSA_REQ) message to the BS. (If the other end point is also a part of an IEEE 802.16 cell it too should follow the same procedure.) To do this, the application layer of the SS must communicate with the MAC layer, and alert it of the beginning of the voice stream. We envisage native VoIP applications for WiMAX which would have cross layer communication between the application and MAC to

better perform this function. The DSA_REQ message will include parameters such as application/flow type, bandwidth requirements, delay and jitter requirements.

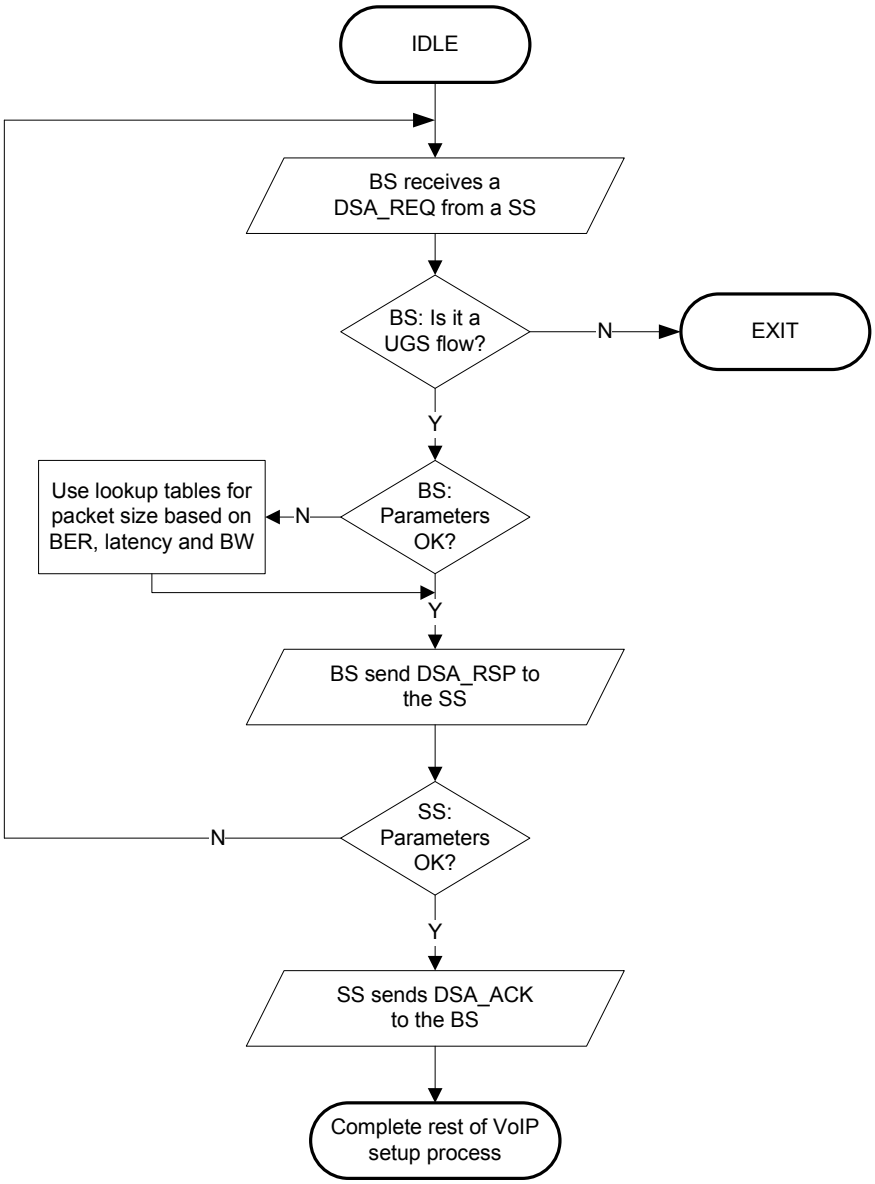

Figure 4-6 Procedure to determine an optimal parameter set at the start of a UGS service flow. In the first decision box, if the flow is not a UGS type, then the procedure will be different and is not shown here.

If the BS agrees to all the parameters requested by the SS it will echo these back in a Dynamic Service Addition Response (DSA_RSP) message. (It would seem logical that this step occurs before the SS has setup the session with the receiver using H.323, SIP or another setup protocol) The procedure for this is shown in Figure 4-6.

If however, the requested parameters are not optimal and can be substituted by more efficient ones, the BS will indicate these in the DSA_RSP message. Once a set of values is agreed upon, the SS will confirm the use of the parameters by sending a Dynamic Service Addition Acknowledgement message (DSA_ACK) to the BS.

For applications which cannot change t_{pkt} , the SS should indicate this to the BS. We propose using one of the unused Service Flow Parameters in the DSA_REQ as an indicator. The BS will not attempt to optimize such parameters.

In the mobile scenario, or when channel conditions vary with time, the SS will carry out periodic ranging to keep its modulation scheme at an optimal level by updating its burst profile. Every time an update takes place the BS will initiate a Dynamic Service Change (DSC) handshake process to renegotiate the parameters of the VoIP session. The MAC then informs the application of the required changes, which in turn, will inform or renegotiate packetization/encoder/compression settings with the other end-point of the conversation.

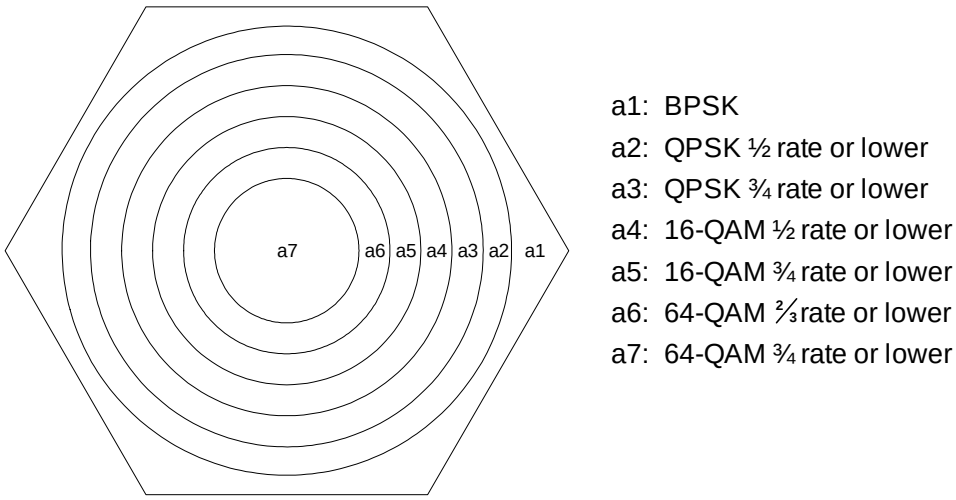

Figure 4-7 Annuluses in a cell area

4.2.5 The Number of Supported Users

Consider a hexagonal cell. Based on the SNR requirements of the different modulation schemes (or burst profiles), the cell area can be classified into annulus regions. If travelling from the centre of the cell in a radial direction, the boundaries of these annuluses mark the change to a lower burst profile. The area of these annuluses as a percentage of the total cell area, is given by a_i, where for example a_3 represents

the region which can use only the three lowest modulation schemes, Figure 4-7. We assume that all the SSs in the cell are uniformly distributed with a constant areal density (an equal number of SS in any given unit area of the cell). $b_{r,i}$ is the effective bit rate of a SS in the i^{th} annulus for a randomly chosen t_{pkt}. $b_{o,i}$ is the effective bit rate of a SS in the i^{th} annulus for an optimally chosen t_{pkt}. n_r and n_o are the numbers of users in the system for a random t_{pkt} and optimal t_{pkt}. Assume the same amount of resources, measured in OFDM symbols, is used in the random case as well as the optimal case, (4.16).

$$\sum_{i=m_l}^{m_h}\left[n_r \cdot a_i\left(\frac{b_{r,i}}{n_{bps,i}}\right)\right] = \sum_{i=m_l}^{m_h}\left[n_o \cdot a_i\left(\frac{b_{o,i}}{n_{bps,i}}\right)\right] \tag{4.16}$$

Then, the ratio n_o:n_r is obtained, (4.17). This gives the proportional increase in the number of users due to optimal selection of t_{pkt}. Here m_l and m_h represent the index of the lowest and highest modulation schemes respectively.

$$\frac{n_o}{n_r} = \sum_{i=m_l}^{m_h}\left[a_i \cdot \left(\frac{b_{r,i}}{n_{bps,i}}\right)\right] \Bigg/ \sum_{i=m_l}^{m_h}\left[a_i \cdot \left(\frac{b_{o,i}}{n_{bps,i}}\right)\right] \tag{4.17}$$

The cell boundary of an ideal cell in a cellular infrastructure is a hexagon so the boundary of the lowest modulation scheme is considered to be hexagonal. We also assume the cell is large enough so that the cell boundary is in effect the transmission range of the lowest modulation scheme.

Table 4-1 Different modulation schemes used and OFDM symbol parameters

Modulation Scheme	**Bits per OFDM symbol (n_{bps})**	**Percentage of total cell area (a_i)**
16 QAM 1/2	384	20.1
16 QAM 3/4	576	51.5
64 QAM 2/3	768	9.2
64 QAM 3/4	864	19.2

4.3 Sample Scenario

A sample scenario using common values for 802.16 is now considered to demonstrate the impact of the packetization interval. The analysis was carried out using Matlab. The number of bits per symbol (n_{bps}) for different modulation schemes is given in

Table 4-1. We are considering the UL phase of the flow of a 256 sub carrier OFDM system with a TDD frame structure and a 4ms frame duration. The burst profile used depends on the signal-to-noise ratio which depends on the distance from the BS. The VoIP application is assumed to produce a data stream at a rate of 32 kbps. This is an implementation dependant value. We have used 32 kbps as it will produce a carrier grade voice stream (Hattingh and Szigeti 2004).The retransmission limit is set as per (4.11). The maximum number of tolerated residual bit errors in a packet reaching the MAC layer is zero.

Table 4-2 SNR requirements for different modulation schemes used and cell coverage percentages

Modulation Scheme	**Minimum receiver SNR (dB)**	**Percentage of total cell area covered**
BPSK 1/2	6.4	39.4
QPSK 1/2	9.4	20.6
QPSK 3/4	11.2	27.9
16 QAM 1/2	16.4.	4.1
16 QAM 3/4	18.2	5.2
64 QAM 2/3	22.7	0.9
64 QAM 3/4	24.4	1.9

We consider a cell with a radius of 2 km. Using the SNR thresholds in Table 4-2, it can be shown that the entire cell can be covered using the three highest modulation schemes when a BS transmit power of 1W is used. Nevertheless we include the next lower modulation scheme to cope with shadowing variations. A two-ray ground propagation model (Balanis 1977), and the free space model (Haykin 1994) were used to calculate receiver SNRs, see section 2.3.4 in Chapter 2. At distances closer than the crossover distance, the free space model is more accurate as it does not show fluctuations due to alternating constructive and destructive interference (Fall and Varadhan 2006). The revised coverage percentages are given in Table 4-2 above.

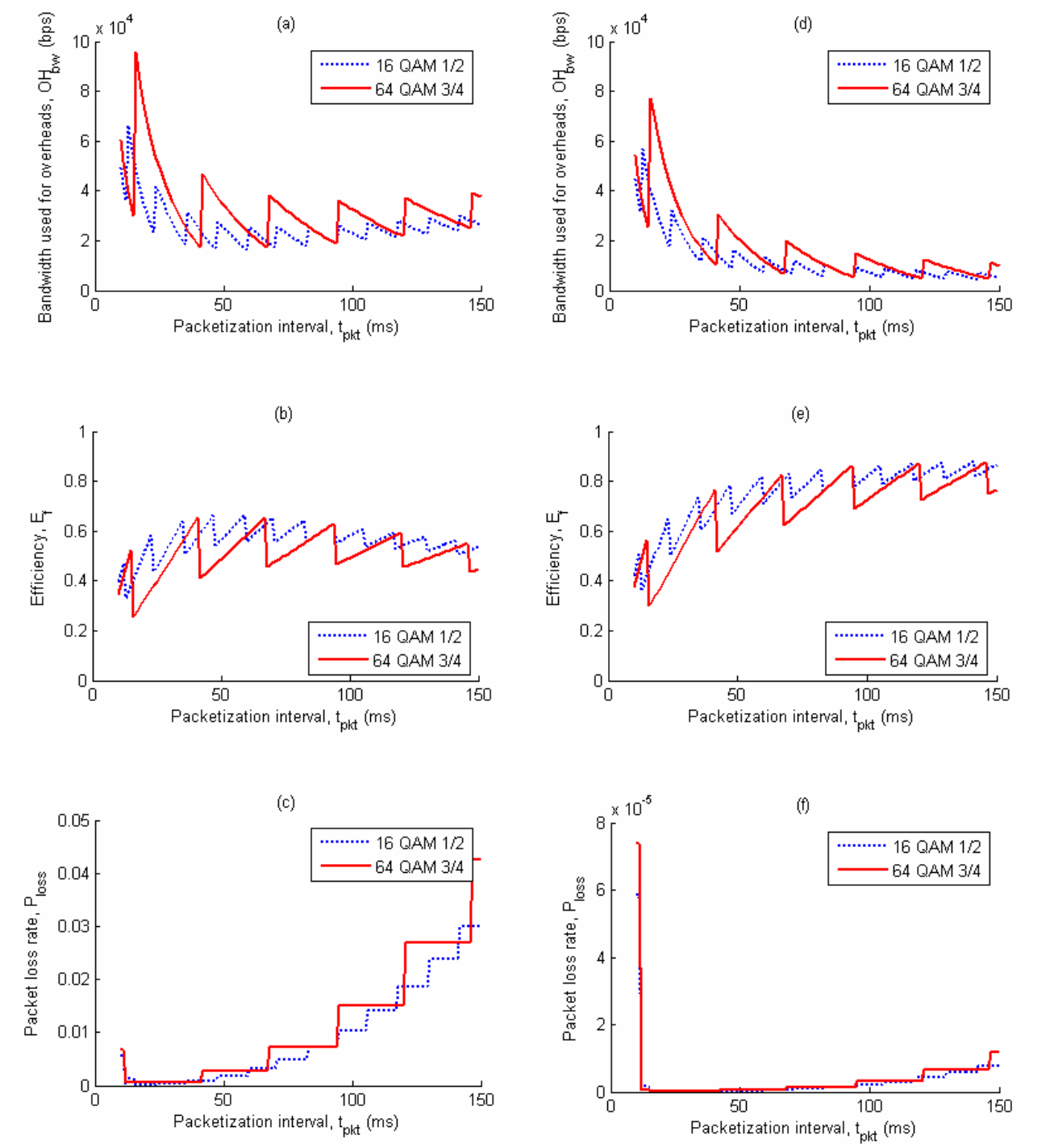

Figure 4-8 (a)-(c) are plots of overhead bandwidth, efficiency and PER for BER = 10^{-4} . (d)-(f) are for a BER = 10^{-5}. The two modulation schemes shown here are 16QAM ½ and 64QAM ¾. The spiky nature of the overhead bandwidth and efficiency plots, as well as the staircase shape of the PER plots is due to the packet size being an integer multiple of OFDM symbols

4.3.1 Overheads, Efficiency and Packet Loss Rates

Figure 4-8 (a), (b) and (c) give the efficiency, overhead bandwidth and packet loss rate for packetization intervals ranging from 10ms to 150ms with BERs of 10^{-4} and 10^{-5}. The saw-tooth effect is due to the transmission units being integer multiples of OFDM symbols.

When the BER is lower, as in Figure 4-8 (d), (e) and (f), the optimal packet size is larger, which is intuitive. In the context of VoIP it is not possible to select the largest possible packetization interval even if it satisfies the QoS packet loss limit. We also need to stay within the latency bounds of the flow which is the reason why the Usability Factor includes the maximum latency.

It is also clear from both Figure 4-8 (a) and (d) that a difference of a few milliseconds can increase the overhead bandwidth up to tens of kbps, which can be a few times as much as the bandwidth of the actual voice application.

The higher P_{loss} values at small intervals, clearly visible in Figure 4-8 (c) and (f), can be explained by considering (4.11). At small intervals, the maximum retransmit limit is also low which reduces the probability of recovering erroneously received packets.

4.3.2 Usability Factor, *K*

Using the results of section 4.2.2, the usability factor has been calculated and plotted for BERs of $10^{-3.5}$, 10^{-4} and 10^{-5}; see Figures 4-9, 4-10 and 4-11 respectively. The constrains used are:

A) $P_{loss,max} = 0.01$

B) $L_{max} = 100$ ms

C) $BW_{max} = 80$ kbps

These values except for the packet loss rate have been arbitrarily picked for the purpose of illustrating our scheme.

Note that at BER = $10^{-3.5}$ only a few intervals are usable. (For BER = 10^{-3} none of intervals can be used under the given conditions, so this has not been plotted.) These intervals only barely satisfy the three constraints. Constraints A and C have the biggest effect at high BERs.

At lower BERs the possibilities increase and more intervals will satisfy the requirements, as seen in Figure 4-10 and Figure 4-11. Constraint B (L_{max}) limits the maximum value of the packetization interval, while constraint C (BW_{max}) limits the minimum value of it. So constraints B and C dominate. The closer K is to 1, the better the chosen interval.

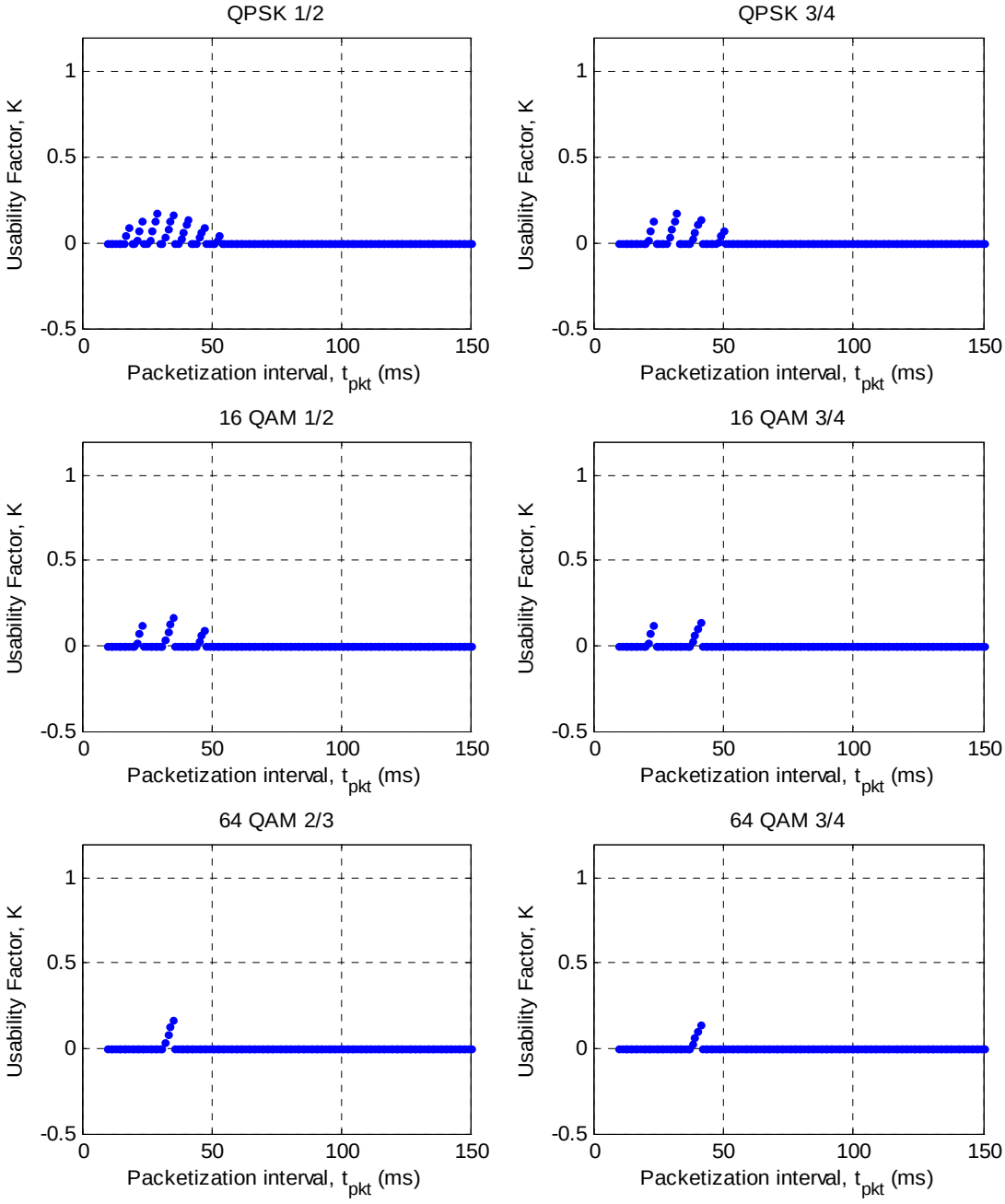

Figure 4-9 Usability Factor, K, for various packetization intervals for BER = $10^{-3.5}$. The constraints used are: the maximum latency of 100ms, the maximum bandwidth usage of 80kbps and the maximum packet loss rate of 1%.

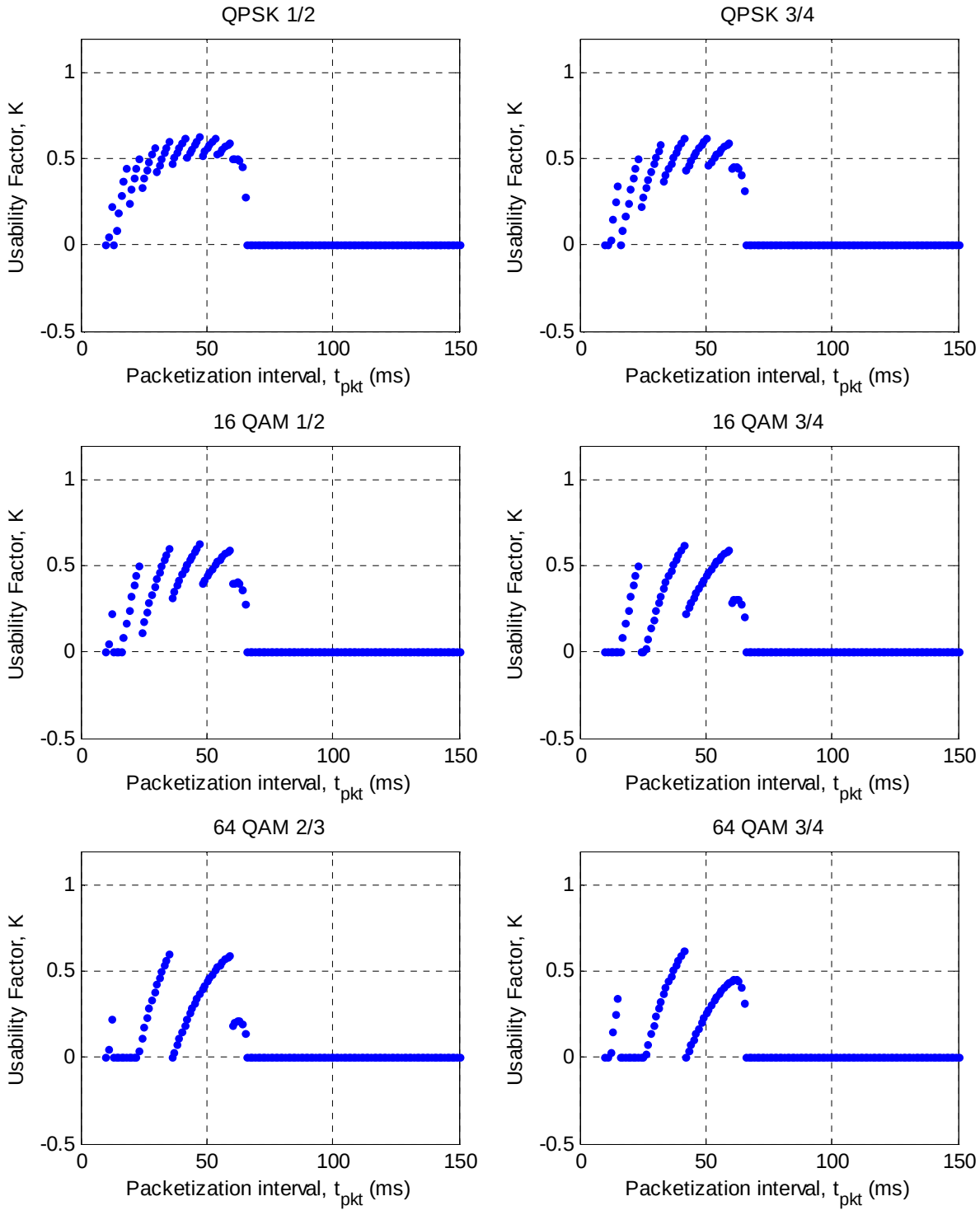

Figure 4-10 Usability Factor, K, for various packetization intervals for BER = 10^{-4}. The constraints used are: the maximum latency of 100ms, the maximum bandwidth usage of 80kbps and the maximum packet loss rate of 1%.

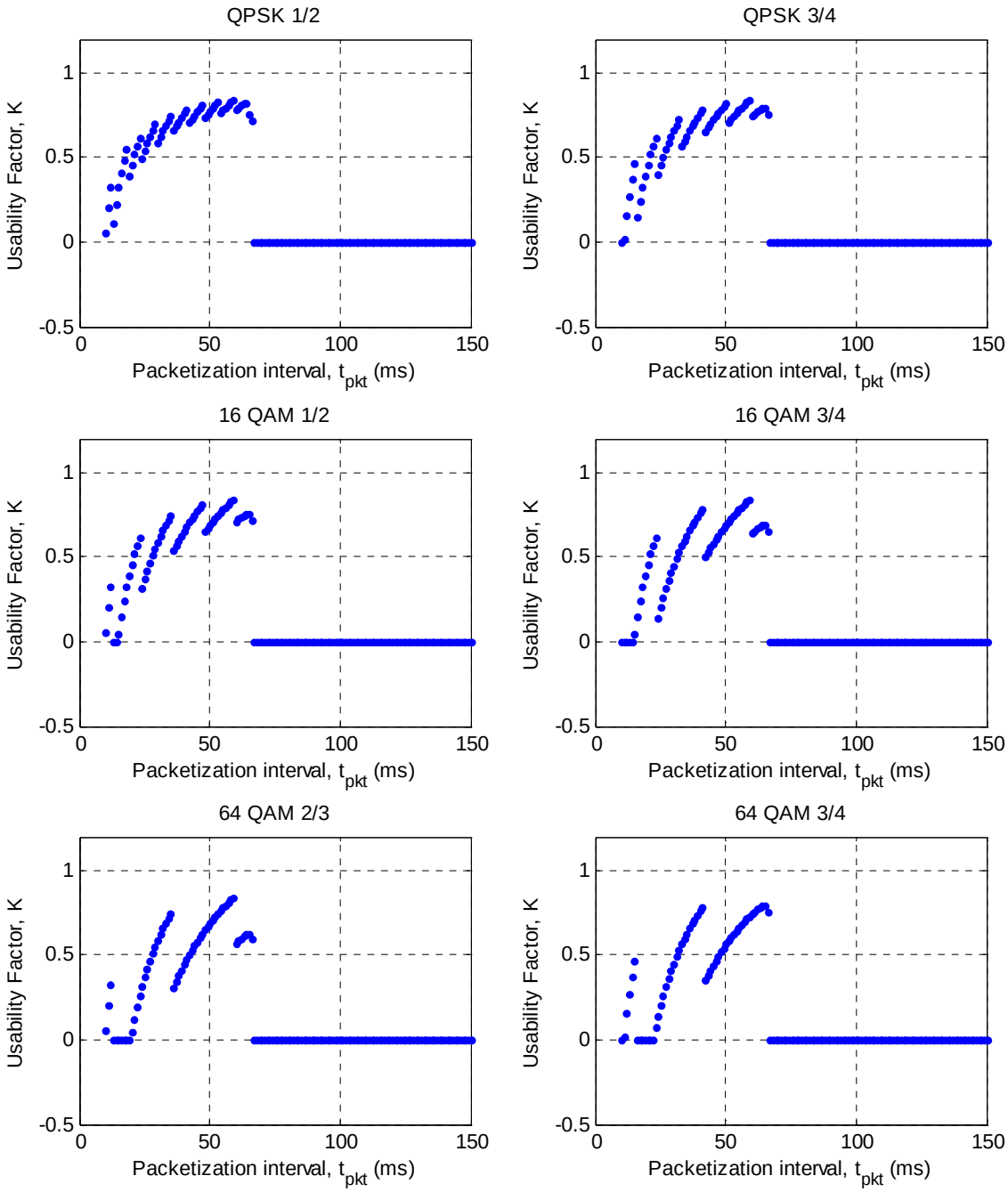

Figure 4-11 Usability Factor, K, for various packetization intervals for BER = 10^{-5}. The constraints used are: the maximum latency of 100ms, the maximum bandwidth usage of 80kbps and the maximum packet loss rate of 1%.

4.3.3 Derived Lookup Tables

On the basis of the results obtained in section 4.3.2, we may summarise these plots into lookup table by quantizing. In the scenario we use three levels of usability: High (H), Medium (M) and Low (L). L implies one or more of the constraints will not be satisfied and is a reject ranking. Figure 4-12 shows the quantized version of two plots derived from Figure 4-11. This makes the lookup table simple, and less memory is needed to store it if required to store in hardware.

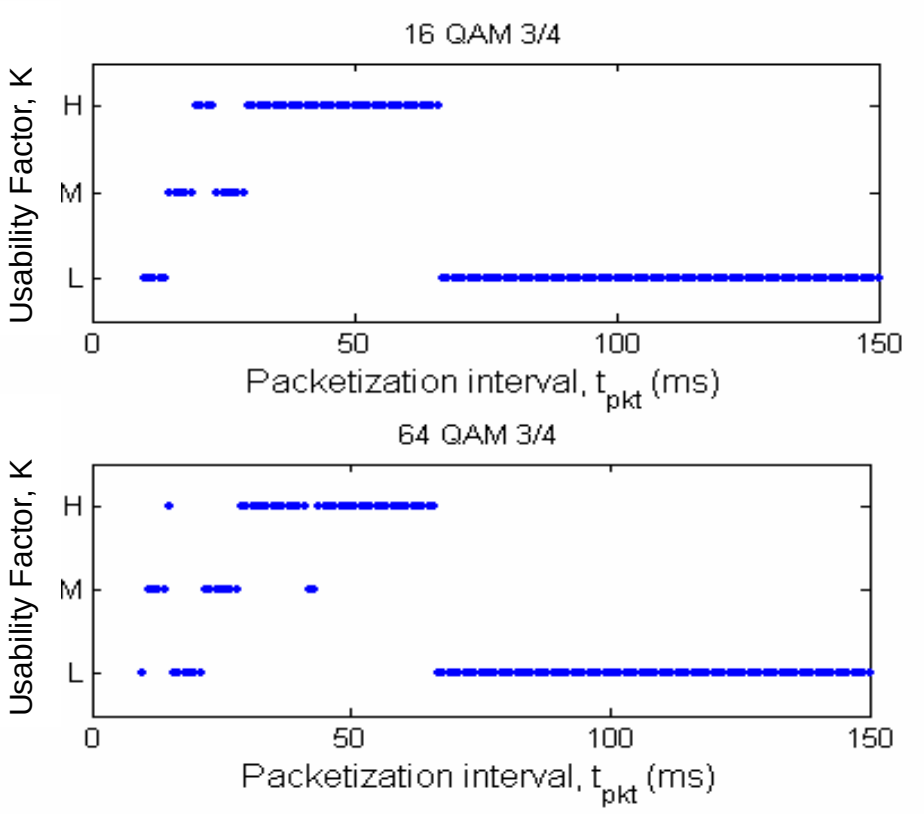

Figure 4-12 Quantized Usability Factor of various packetization intervals. H, M and L indicate High, Medium and Low usability respectively. BER = 10^{-5}

As an example, if a SS using 64 QAM 3/4 requests a 20 ms interval, the BS could respond with a counter request for an interval between 29 ms and 41 ms. The SS requested what is a very commonly used packetization interval for VoIP applications. The BS responds with

1) 41 ms – the more efficient option. However it has a higher latency and P_{loss},.

2) 29 ms – less efficient than option (1) but has lower latency and P_{loss}.

Based on channel conditions which the BS has knowledge of, the estimated delay to the destination, and any other QoS requirements of the flow, it can select the most appropriate value.

4.3.4 Increase in the Number of Users

The increase in the total number of supported simultaneous VoIP users (with reference to the t_{pkt} = 20 ms case) was given in (4.17). This ratio is converted to a percentage increase and plotted in Figure 4-13. The negative portion of the curves represents a reduction in the total users relative to the 20 ms case.

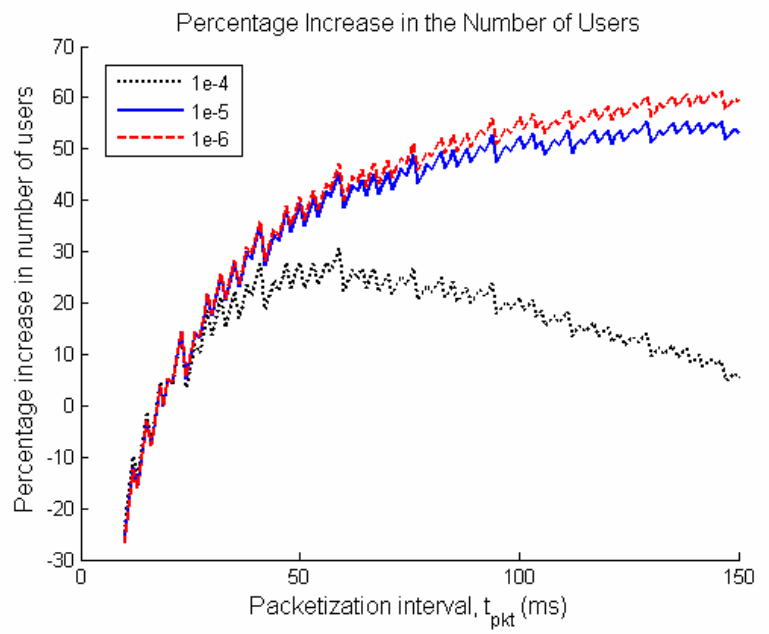

Figure 4-13 Percentage increase in the number of users for a fixed amount of UL resources, with all possible modulation schemes in use. Three values of BER are compared.

If we compare Figure 4-13 and Figure 4-14, it can be seen that the increase is more in this scenario where only the highest four modulation schemes are used. The curves are jagged. This is because at higher modulation schemes, OFDM symbols consist of a large number of bits. Hence most t_{pkt} , the packet can be represented by a small number of symbols. A small change in packet size can have a relatively large effect on the number of symbols required.

Poor choices of t_{pkt} can cause almost a whole symbol to be unused which can have a significant effect on the total cell throughput. This in turn can cause the number of sustainable users to vary significantly for a small change in t_{pkt}. It is also

important to note that in this scenario at BER = 10^{-5} only t_{pkt} values up to 67 ms are allowed, Figure 4-11. The *K* value is zero for all intervals above 67 ms. Based on (4.12), 67 ms would give a minimum latency of 100 ms which is our latency constraint. Similarly for the other BER values plotted the cutoff value should be read from the *K* value plot at the corresponding BER

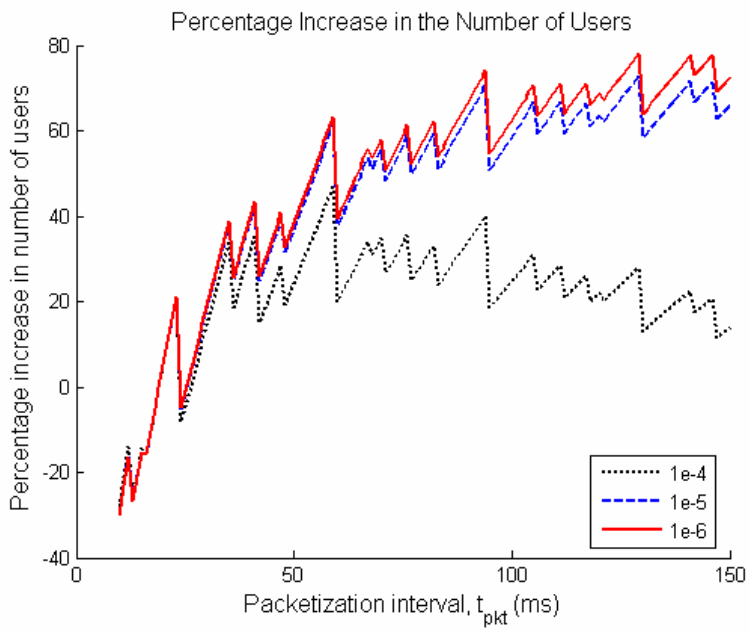

Figure 4-14 Percentage increase in the number of users for a fixed amount of UL resources, with only the highest four modulation schemes in use. Three values of BER are compared.

4.4 Simulation Study

A scenario is simulated using the simulation model described in detailed in Chapter 3. However some specific modifications to the standard such as, the UGS retransmission strategy and, dynamic service addition/change process which have not previously been discussed, are included in the preceding sections.

The goal of the simulation study is to demonstrate the practical validity of the proposed scheme, to use the best packetization interval for the prevailing conditions. In order to simulate variations in conditions we have introduced SS mobility, which impacts the receive/transmit channel conditions by dynamically changing the distance between the BS and the SS. The simulation is dependant on the function of the ranging process of 802.16, which we have shown to be accurately modelled in

Chapter 3. A discussion of details of the simulation scenario, assumptions made, limitations, input parameters, output parameters and simulation results follows.

4.4.1 Simulation Scenario

A single mobile SS is initialized very close to the BS location. It then moves at constant speed towards the cell boundary. A cell radius of 2 km is used and time to traverse this distance is 80 s. These are values chosen to easily observe the periodic ranging process and step down of modulation schemes within a short simulation time. Periodic ranging is carried out every 10 s which is adequate for this relatively high speed mobility scenario. In addition ranging is also done on demand, whenever the SS detects a drop or rise in SNR which crosses predefined thresholds. We stress that Fixed WiMAX does not support vehicular mobility. This simulation is only used as a means to demonstrate the process of actively managing the packetization interval.

A 32 kbps Constant Bit Rate (CBR) packet flow from the SS to the BS is used to simulate the UL stream of an uncompressed VoIP session. In the first case, the SS uses a 20 ms packetization interval throughout the simulation. The simulation is repeated with our modification to the MAC layer. The SS's application begins with a 20 ms packetization interval. The resulting MAC layer throughout is compared.

4.4.2 Simulator Modifications

The CBR application was modified in order to accommodate dynamic packet size and interval adjustments as and when indicated by the MAC layer. The application by default only allows these values to be set at the start of the flow. When the application begins it will communicate with the MAC layer using the defined interface. The SS will then initiate the service allocation process with the BS.

Each separate flow is allocated a unique CID, based on source and destination port numbers. These are known to the MAC layer by accessing the IP header of the packets. The SS then transmits a DSA-REQ to the BS, listing the allocated CID and requested flow parameters. The rest of the DSA handshake process follows until, a set of parameters are found which are agreeable to both parties.

When the SS MAC detects a drop in the SNR (or an increase if moving toward the BS), it initiates ranging. The BS on receiving the RNG-REQ checks if any VoIP

connections using UGS scheduling are active for that SS. If so, it initiates a DSC process for each connection as soon as ranging is completed. A certain amount of hysteresis allowance is included in the system to prevent flapping of modulation schemes. This could happen in the case of a SS being at the boundary between two annuluses, with varying reception conditions. The theory of this is out of the scope of this work. The specification of SNR thresholds is covered by the standard.

4.4.3 Assumptions

In the simulation the SS accepts the recommended packetization interval parameter without requiring further negotiations. The application produces a constant bitrate stream of packets, and not a bursty stream full of talk spurts.

In order for this scheme to be successful, we assume that the other end point of the VoIP session is using an application, with a codec able to change packetization intervals on demand.

We have approximated the 3-way hand shaking process during ranging, which comprises request-response-acknowledgement (RNG-REQ, RNG-RSP and RNG-ACK) to a 2-way process omitting the acknowledgement.

The BER of the link used is zero. This is assumption is based on the fact that we are only interested in demonstrating the process of dynamic packetization interval change, and the increase in efficiency gained by it, and not whether the BS can accurately assess the BER of the UL from the SS. Another important point is that, the ranging process makes sure BERs are within acceptable limits, so that a situation should not arise where the SS is on a modulation scheme which is too high for the conditions.

4.4.4 Simulation Results

Figure 4-15 gives the resulting bandwidth usage obtained from the simulation. The nominal curve (red) shows the actual bitrate of the application, which is fixed for the duration of the simulation run, at 32 kbps.

The MAC layer throughput of a SS using a t_{pkt} = 20 ms, and travelling from very near the centre of the cell as described in the simulation scenario is given by the blue curve. At the highest burst profile the flow is extremely inefficient, and uses close to three times the nominal bitrate. This improves as the SS steps down in burst

profiles, as it moves away from the BS. At the last profile tested at, it used approximately 57 kbps.

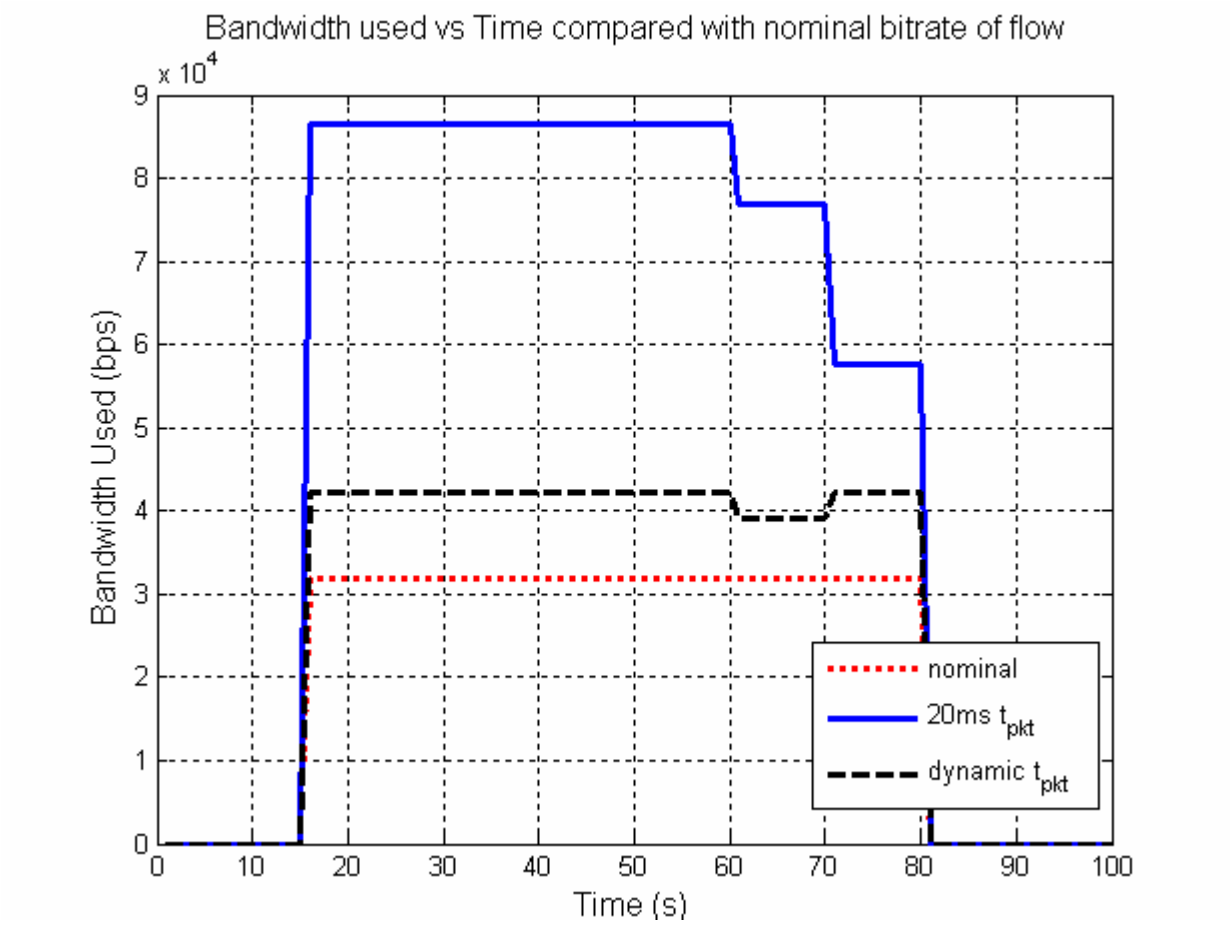

Figure 4-15 Comparison of bandwidth used for 20 ms t_{pkt} and dynamic t_{pkt}. The nominal bitrate of the flow (32 kbps) is also shown.

The black curve gives the MAC layer throughput of a SS configured to use the proposed scheme (also enabled at the BS). Initially the application starts with a t_{pkt} = 20 ms which we assume to be the default for most VoIP applications. This is increased to t_{pkt} = 41 ms during the DSA handshake by the request of the BS. It is assumed that the SS will accept the parameters proposed by the BS. The simulation time is denoted by t. As the SS moves away from the BS, while t < 60 s it is able to use 64QAM 3/4. At t = 60 s, after periodic ranging, it steps down to the next lower modulation scheme, 64QAM 2/3. The optimal t_{pkt} = 59 ms. Efficiency has increased some what at this particular interval. At t = 70 s, another step down occurs to 16QAM 3/4. Once again t_{pkt} = 41 ms is selected to be the optimal value. The t_{pkt} values used are taken from Figure 4-11, since K values for BER= 0, are very similar to those for

BER = 10^{-5}. Actually usable *K* values for a higher BER are a subset of those usable at a lower BER.

We reiterate that fixed WiMAX cannot support the mobility scenario given in the above simulation, but is used simply to observe the response of the modified MAC layer, to adaptive modulation and coding (AMC) in a short time span.

4.5 Conclusion

The efficiency of bandwidth usage is affected by the packet size in IEEE 802.16 systems using an OFDM physical layer. Since an integer number of symbols is used to transport a PDU, if the packet size is not close to an integer multiple of the symbol size, there will be wastage of bandwidth. This is more pronounced when the packet size is relatively small, such as, in VoIP applications. We have applied this phenomenon to VoIP, and analyzed the effect the packetization interval has on efficiency. The packet error rate and latency have also been factored into the analysis, as these are parameters which define the QoS of a VoIP flow. It was shown that by careful selection of the packetization intervals for VoIP, the amount of bandwidth wasted on overheads can be minimized, and as a result the number of supported users can be increased.

A selection criteria for the packetization interval based on three constraints, i.e., packet loss rate, bandwidth usage and latency, was introduced. A parameter which we call the Usability Factor, *K*, was defined. This is an index, which tells the BS how suited a given packetisation interval is, to the prevalent conditions, and the QoS requirements of the service flow.

A new retransmission strategy for UGS flows was introduced. In this scheme the BS provides implicit notification of failure, through a retransmit opportunity in the following frame. The BS continues to provide retransmit opportunities, until the packet is received free of errors, or, the deadline for that packet is reached, or, the SS flushes the packet cache and is no longer able to retransmit. This fast retransmit scheme is factored in to the analysis, as well the simulation model.

Modifications proposed for the MAC layer operation are shown to be able to change the packetization interval during call setup and also during periodic ranging or ranging on demand. This modification can be accommodated in the existing ranging

process, and Dynamic Service Addition/Change handshaking process, so no extra overhead is introduced. The Usability Factor can be quantized with a certain level of granularity, and stored as a lookup table. Using these tables at the BS, makes selecting an optimal interval simple and fast. A proof of concept was also given based on a simulation scenario, which shows positive results in terms of increased efficiency.

Chapter 5

ARQ for Real-Time Downlink Traffic

Wireless channels are prone to errors, and this holds true for the OFDM physical layer used in Fixed WiMAX. Much research has been done to improve the reliability of wireless links by upper layer techniques. Local retransmission is one of the most commonly used at the MAC layer in networks ranging from WLAN to 3G; WiMAX also has included an optional retransmission method based on Automatic Repeat Request (ARQ) and Hybrid ARQ (HARQ). HARQ mitigates the effect of impairments due to the channel and external interference, by effectively employing time diversity along with incremental transmission of parity codes (subpackets in this case). In the receiver, previously erroneously decoded subpackets and retransmitted subpackets are combined, to correctly decode the message. HARQ is only available with the OFDMA physical layer so is not within the scope of this work. The transmitter decides whether to send additional subpackets, based on feedback messages received from the receiver (Yaghoobi 2004).

In (Zhi and Jong-Moon 2004), an analysis has been done on the effects of ARQ on real-time traffic using the concepts of ARQ capacity, and effective capacity. It considers an error free feedback channel with timely feedback for all transmissions. It also only takes into account the DL bandwidth used for ARQ and not the UL bandwidth used for the ARQ feedback messages.

The strategy of (Uhlemann, Aulin et al. 2002) is to combine different coding and decoding methods with ARQ techniques, in order to fulfil the application requirements within a wireless real-time communication system. These requirements are formulated as two QoS parameters: (1) delivery deadline and (2) probability of correct delivery before the deadline, leading to a probabilistic view of real-time communication. An application can negotiate these QoS parameters, thus creating a flexible and fault-tolerant scheme.

A multiple retransmission based error recovery scheme was proposed (Hu, Zhu et al. 2001). In this work the receiver actively advertises a list of negative acknowledgements periodically. The scheme allows for multiple retransmissions based on end-to-end latency. There is however a substantial delay in recovering lost packets.

Video streaming applications have different requirements than voice applications due to varying compression methods. Compression is achieved by sending the most important frame (base line information) followed by less important frames (incremental information), which combine to show the progression of the video. Much work has been done in developing robust ARQ schemes in this context (Po-Chin, Zhi-Li et al. 2000; Uhlemann, Aulin et al. 2002; Chia-Hui, Ray et al. 2003; Min and Gang 2003; Seferoglu, Altunbasak et al. 2005).

Other schemes deal with providing QoS for real-time services using ATM by channel dependant adaptive coding, and multi-copy retransmission. Real-time ATM is similar in nature to a voice service because ATM cells are small in size (Chang Wook, Chung Gu et al. 1999).

The schemes described above as well as many of the others assume that ARQ feedback is always sent by the receiver, and received by the transmitter in a timely fashion, without any impact on available bandwidth. In this work we analyze the standard ARQ method taking into account both downlink, and uplink traffic. Then we propose a novel ARQ scheme. We also provide an analytical model for this scheme, and investigate its operation through simulation.

5.1 Operation of IEEE 802.16 ARQ

The ARQ mechanism is a part of the MAC layer itself, which is optional for implementation. When implemented, ARQ may be enabled on a per-connection basis. The per-connection ARQ will be specified and negotiated during the beginning of the flow. A connection cannot have a mixture of ARQ and non-ARQ traffic. Similar to other properties of the MAC protocol, the scope of a specific instance of ARQ is limited to one unidirectional connection. In other words a specific ARQ instance is attached to a specific CID. The ARQ scheme used in Fixed WiMAX will be referred to as the standard scheme from this point onward.

5.1.1 ARQ Block Usage

A MAC SDU is logically partitioned into blocks whose length is specified by the connection TLV parameter, ARQ_BLOCK_SIZE. When the length of the SDU is not an integer multiple of the connection's block size, the final block of the SDU is formed using the SDU bytes remaining after the final full block has been determined. Once an SDU is partitioned into a set of blocks, that partitioning remains in effect until all blocks of the SDU are successfully delivered to the receiver, or the SDU is discarded by the transmitter state machine.

Sets of blocks selected for transmission or retransmission are encapsulated into a PDU. A PDU may contain blocks that are transmitted for the first time as well as those being retransmitted. Fragmentation shall occur only on ARQ block boundaries. If a PDU is not packed, all the blocks in that PDU must have contiguous block numbers.

If ARQ is enabled at the connection, Fragmentation and Packing subheaders contain a Block Sequence Number (BSN), which is the sequence number of the first ARQ block in the sequence of blocks following the subheader. It is a matter of transmitter policy whether or not a set of blocks once transmitted as a single PDU should be retransmitted also as a single PDU. Figure 4-1 illustrates the use of blocks for ARQ transmissions and retransmissions. Two options for retransmission are presented, (1) with rearrangements of blocks, and (2) without rearrangements of blocks.

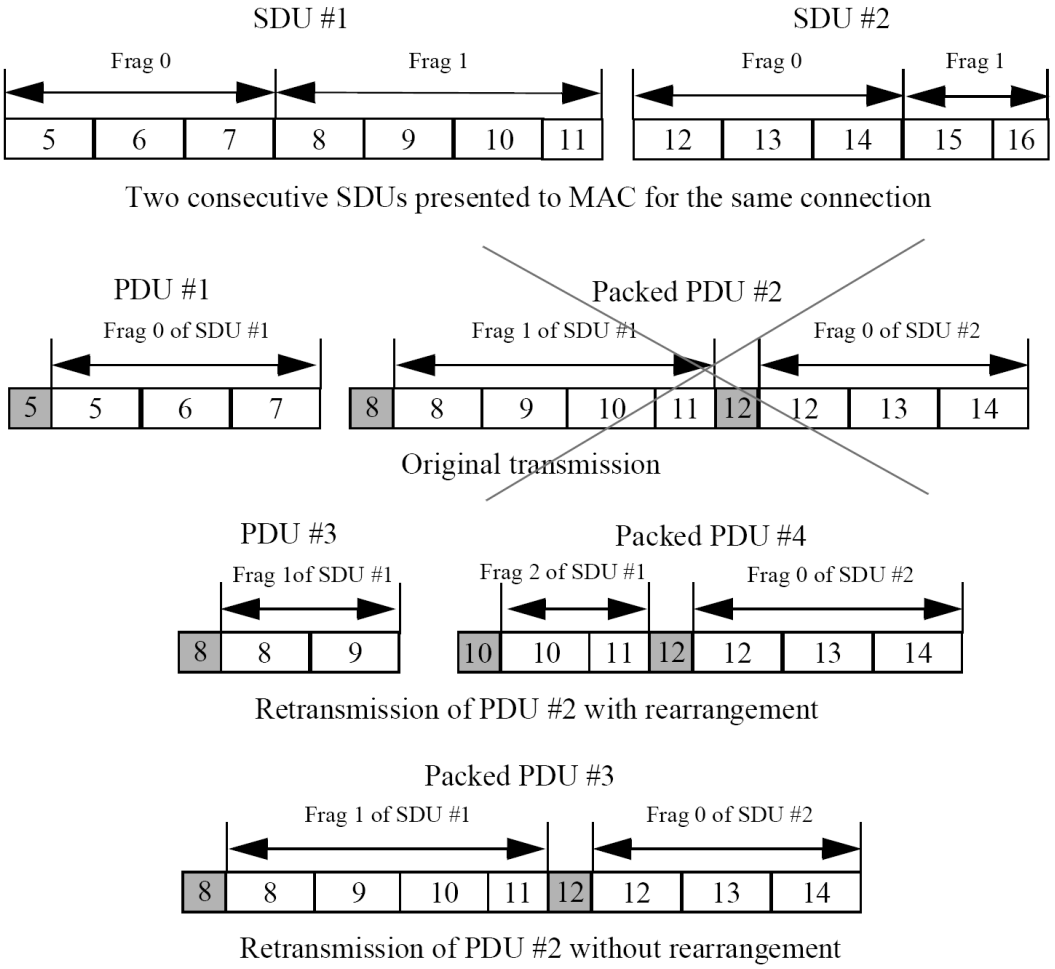

Figure 5-1 A broadband wireless alternative to PSTN based on WiMAX.

5.1.2 ARQ Acknowledgement Types

Acknowledgements are in the form of bit maps. The value of each bit signifies correct or incorrect receipt of an ARQ block. Two types of acknowledgement maps are defined in the standard.

1) Selective ACK Map

Each bit set to one indicates the corresponding ARQ block has been received without errors. The bit corresponding to the BSN value in the Information Element (IE), is the most significant bit of the first map entry. The bits for succeeding block numbers are assigned left-to-right (MSB to LSB) within the map entry.

2) Sequence ACK Map

Each bit set to one, indicates the corresponding block sequence has been received without error. The MSB of the field corresponds to the first sequence length field in the descriptor. The bits for succeeding length fields are assigned left-to-right within the map entry.

5.1.3 ARQ-enabled connection setup and negotiation

Connections are set up and defined dynamically through the Dynamic Service Addition/Change (DSA/DSC) class of messages. CRC-32 shall be used for error detection of PDUs for all ARQ-enabled connections.

5.1.4 Sequence number comparison

Transmitter and receiver state machine operations include comparing BSNs and taking actions based on which is larger or smaller. In this context, it is not possible to compare the numeric sequence number values directly to make this determination. Instead, the comparison shall be made by normalizing the values relative to the appropriate state machine base value and the maximum value of sequence numbers, *ARQ_BSN_MODULUS*, and then comparing the normalized values.

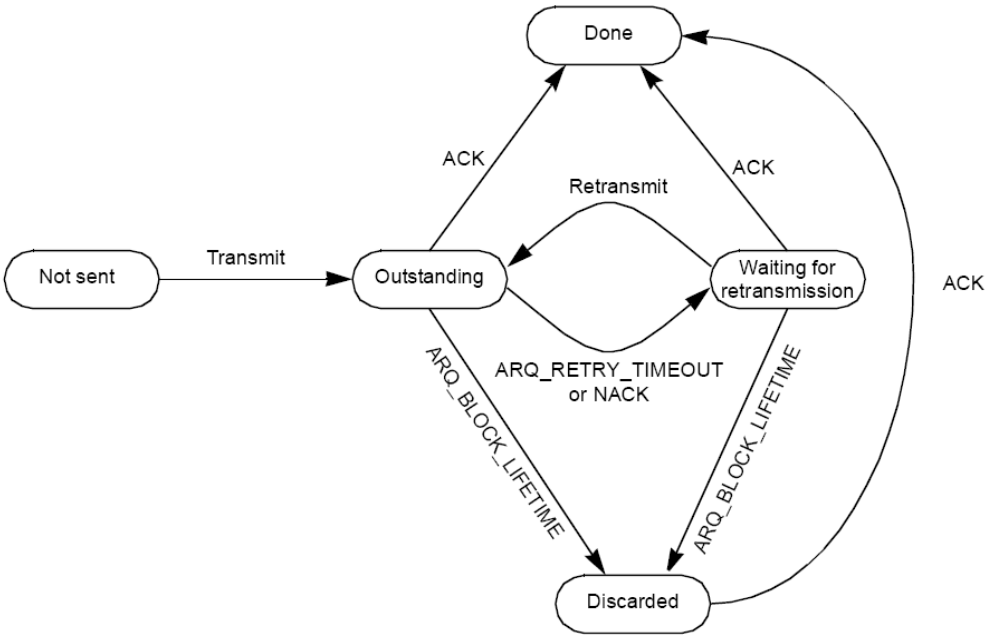

Figure 5-2 State machine of the transmitter.

5.1.5 Transmitter state machine

An ARQ block may be in one of the following four states — not-sent, outstanding, discarded, and waiting for retransmission. Any ARQ block begins as not-sent. After it is sent it becomes outstanding for a period of time termed *ACK_RETRY_TIMEOUT*. While a block is in outstanding state, it is either acknowledged and discarded, or transitions to waiting-for-retransmission after *ACK_RETRY_TIMEOUT* or NACK. An ARQ block can become waiting-for-retransmission before the *ACK_RETRY_TIMEOUT* period expires, if it is negatively acknowledged. An ARQ block may also change from waiting-for-retransmission to discarded, when an ACK message for it is received or after a timeout *ARQ_BLOCK_LIFETIME*, as shown in Figure 5-2.

For a given connection the transmitter shall first handle (transmit or discard) blocks in “waiting-for-retransmission” state and only then blocks in “non-sent” state. When blocks are retransmitted, the block with the lowest BSN shall be retransmitted first.

5.1.6 Receiver state machine

The state machine process of the receiver is given in Figure 5-3. When a PDU is received, its integrity is determined based on the CRC-32 checksum. If a PDU passes the checksum, it is unpacked and de-fragmented, if necessary. The receiver maintains a sliding-window defined by ARQ_RX_WINDOW_START state variable, and the ARQ_WINDOW_SIZE parameter. When an ARQ block with a number that falls in the range defined by the sliding window is received, the receiver shall accept it. ARQ block numbers outside the sliding window shall be rejected as out of order. The receiver should discard duplicate ARQ blocks (i.e., ARQ blocks that where already received correctly) within the window.

The sliding window is maintained such that the *ARQ_RX_WINDOW_START* variable always points to the lowest numbered ARQ block that has not been received or has been received with errors. When an ARQ block with a number corresponding to the *ARQ_RX_WINDOW_START* is received, the window is advanced. The timer associated with *ARQ_SYNC_LOSS_TIMEOUT* shall be reset. As each block is

received, a timer is started for that block. Timers for delivered blocks remain active and are monitored for timeout until the BSN values are outside the receive window.

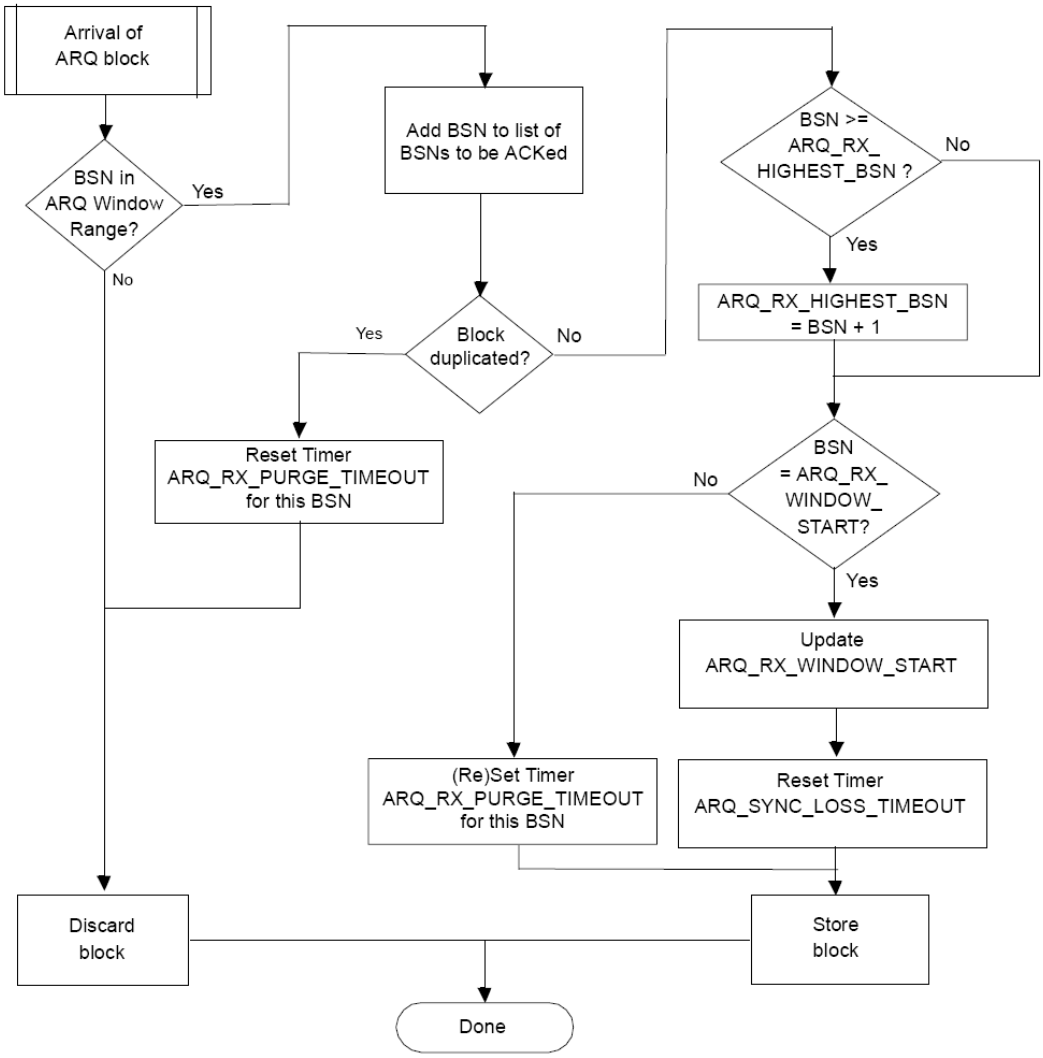

Figure 5-3 State machine of the receiver.

When *ARQ_RX_WINDOW_START* is advanced, any BSN values corresponding to blocks that have not yet been received residing in the interval between the previous and current *ARQ_RX_WINDOW_START* value shall be marked as received and the receiver shall send an ARQ Feedback IE to the transmitter with the updated information. Any blocks belonging to complete SDUs shall be delivered. Blocks from partial SDUs shall be discarded. When a discard message is received from the transmitter, the receiver shall discard the specified blocks, advance *ARQ_RX_WINDOW_START* to the BSN of the first block not yet received after the

BSN provided in the Discard message, and mark all not received blocks in the interval from the previous to new *ARQ_RX_WINDOW_START* values as received for ARQ feedback IE reporting. For each ARQ block received, an acknowledgment shall be sent to the transmitter. Acknowledgment for blocks outside the sliding window shall be cumulative. Acknowledgments for blocks within the sliding window may be either for specific ARQ blocks (i.e., contain information on the acknowledged ARQ block numbers), or cumulative (i.e., contain the highest ARQ block number below which all ARQ blocks have been received correctly) or a combination of both (i.e., cumulative with selective). Acknowledgments shall be sent in the order of the ARQ block numbers they acknowledge. The frequency of acknowledgment generation is not specified here and is implementation dependent. A MAC SDU is ready to be handed to the upper layers when all of the ARQ blocks of the MAC SDU have been correctly received within the time-out values defined.

When *ARQ_DELIVER_IN_ORDER* is enabled, a MAC SDU is handed to the upper layers as soon as all the ARQ blocks of the MAC SDU have been correctly received within the defined time-out values and all blocks with sequence numbers smaller than those of the completed message have either been discarded due to time-out violation or delivered to the upper layers.

When *ARQ_DELIVER_IN_ORDER* is not enabled, MAC SDUs are handed to the upper layers as soon as all blocks of the MAC SDU have been successfully received within the defined time-out values.

When an acknowledgment is received, the transmitter shall check the validity of the BSN. If BSN is not valid, the transmitter shall ignore the acknowledgment. When a cumulative acknowledgment with a valid BSN is received, the transmitter shall consider all blocks in the interval as acknowledged and update its sequence number counters. When a selective acknowledgment is received, the transmitter shall consider as acknowledged all blocks so indicated by the entries in the bitmap for valid BSN values. The bitmap entries are processed in increasing BSN order. A bitmap entry not indicating acknowledgement shall be considered a NACK for the corresponding blocks. When a cumulative with selective acknowledgment and a valid BSN is received, the transmitter performs the actions described above for cumulative

acknowledgment, followed by those for a selective acknowledgment. All timers associated with acknowledged blocks shall be cancelled. (IEEE 802.16 WG 2004).

5.2 Analytical Modelling of ARQ

We consider a system model where the forward as well as reverse wireless links have a probability of bit errors. The model specified in (Zhi and Jong-Moon 2004) is modified as shown in Figure 5-4. Here $A(t)$, $R(t)$ and C, define the packet arrival process, feedback arrival process and service capacity, respectively. Since the number of OFDM symbols per frame is fixed, C is not a time varying quantity. This is in contrast to throughput, which is dependant on the modulation scheme and impacts the number of bits per symbol.

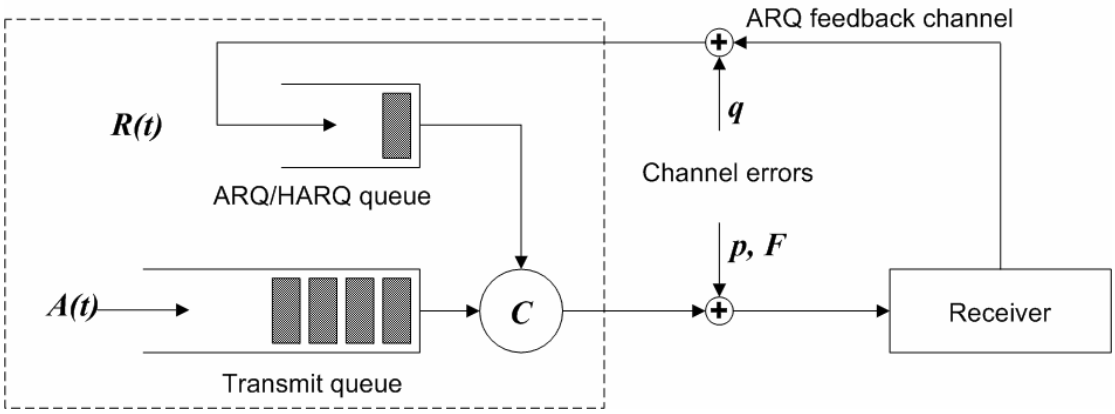

Figure 5-4 Block diagram of system model. Forward channel errors are given by p or F and reverse channel errors are given by q.

5.2.1 ARQ Model

5.2.1.1 Transmitter Model

We consider a selective repeat ARQ model with the following features. These steps are from the BS's perspective, see Figure 5-5.

1) Once a packet is transmitted by the BS on the DL an opportunity will be given in the UL of the following frame to the SS to provide feedback.
2) If no feedback is received or if the feedback is in error itself another slot will be provided in the following frame. This process will be repeated until feedback is received or a timeout happens and the BS will move on to the next packet.

3) If a NACK is received from the SS, the packet/ARQ block will be retransmitted in the following frame and a feedback opportunity given.
4) If an ACK is received the ARQ blocks will be removed from the cache and the packet will be deemed successfully delivered.
5) When the packet deadline is reached the ARQ blocks will be removed from the cache.

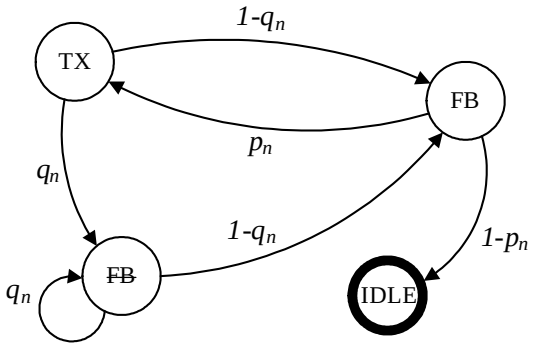

Figure 5-5 State diagram of the transmitter (BS). TX is transmitting state. States FB and ~~FB~~ represent receiving and not receiving feedback, respectively. All transitions occur at the frame boundary.

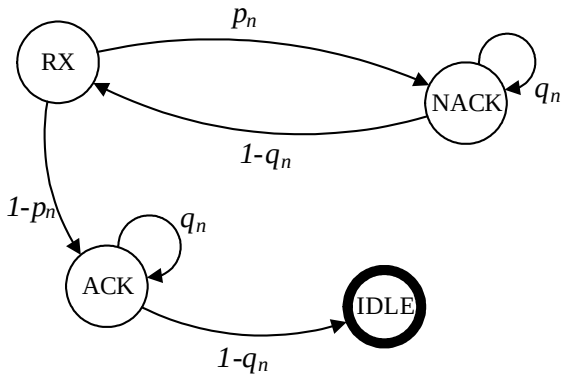

Figure 5-6 State diagram of the receiver (SS). RX is receiving state. States All transitions occur at the frame boundary.

5.2.1.2 Receiver Model

From a SS's point of view the ARQ process can be detailed as follows, see Figure 5-6.

1) A SS expects a DL packet from the BS based on the DL-MAP received at the beginning of the frame. We will assume the SS receives the DL-MAP without errors.

2) If the packet is received without any detectable errors a positive ARQ feedback (ACK) is transmitted to the BS in the provided slot in the UL channel of the following frame.
3) If the packet was received in error (or not received) a negative acknowledgement will be transmitted.
4) The SS waits in idle mode for retransmission of the previous erred packet or for the next packet of the sequence.

The SS plays a passive role in this form of ARQ mechanism, as it merely reacts to the instructions of the BS. While requiring less intelligence from the SS, it also means wasted resources because the BS does not have any information on the packet reception status of the SS. This requires the BS to give feedback opportunities to all SSs regardless of the reception status.

The state transition diagrams for a SS and BS using selective repeat ARQ without packet combining, are given in Figure 5-5 and Figure 5-6. The subscript n signifies the number of frames elapsed since the packet reached the head of line (HOL) of the transmit queue. When n reaches a threshold which marks the deadline for the packet, both state diagrams will transition to the IDLE state, irrespective of their current states. p_n is the packet error rate on the forward link (DL) during the n^{th} frame and q_n is the packet error rate on the reverse link (UL) during the n^{th} frame. We model the channel at the packet level as either a good packet or a bad packet.. As reported by (Hong Shen and Moayeri 1995; Zorzi and Rao 1997), the two-state Markovian model suffices for modelling these packet states.

$$\underbrace{\begin{bmatrix} R & A & N & I \end{bmatrix}}_{S_n} \underbrace{\begin{bmatrix} 0 & \bar{p}_n & p_n & 0 \\ 0 & q_n & 0 & \bar{q}_n \\ \bar{q}_n & 0 & q_n & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{T} = \underbrace{\begin{bmatrix} \bar{q}_n N \\ \bar{p}_n R + q_n A \\ p_n R + q_n N \\ I + \bar{q}_n A \end{bmatrix}^T}_{S_{n+1}} \tag{5.1}$$

.

$$S_n T_n = S_{n+1}$$

$$S_1 \prod_{i=1}^{n} T^i = S_{n+1}$$

The state diagram in Figure 5-5 can also be represented in matrix form using the recurrence relation given in (5.1). R, N, A and I, represent the probabilities of being in RX, NACK, ACK and IDLE states respectively. S_n is the state probability

vector in the n^{th} frame and T is the state transition matrix. $T(4,4) = 1$ is used so that $S_{n+1}(4)$ always has the cumulative probability of completed transmissions.

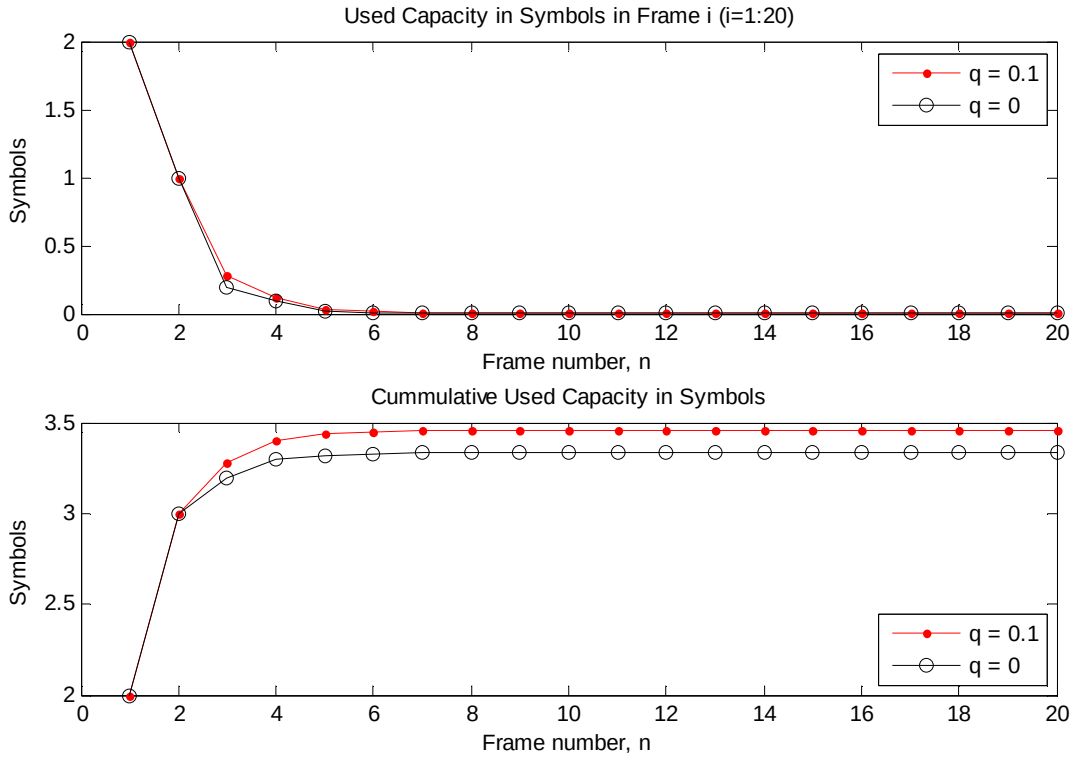

Figure 5-7 Comparison of resource usage both when UL errors are present, and when UL is considered error free. DL packet error rate, p=0.2. Assume that 2 symbols are needed for DL and 1 symbol for feedback message on UL.

5.2.2 Constant Channel Conditions with an Error Free UL

Most of the previous work in this area assumes an error free ARQ feedback channel. This can be justified by the results in Figure 5-7. We have used the recurrence relation in (5.1) to calculate the mean resource (symbol) usage and the cumulative resource (symbol) usage. We use an exaggerated forward channel error rate of 0.2. There is only a small increase in resource usage in the non-error free case. This assumption makes the state diagram in Figure 5-5 redundant. The receiver and transmitter state machines will now be synchronized as any feedback sent by the SS will be received error free on the UL.

5.2.3 Analysis of Resource Usage

The channel capacity used up by the NACKs and ACKs has, not been considered in previous work. In an OFDM system, we will assume that each feedback message being a MAC header plus the feedback MAC management message, will use up one OFDM symbol on the UL. This message will be less than 10 Bytes in size. The derivations to follow from here forward will use the approximation given in section 5.2.2.

If D is the delay experienced by the packet as it reaches the HOL position and D_{max} is the deadline for reception at the SS, we can define the maximum number of frames available for transmission, n_d as,

$$n_d = \frac{\lfloor D_{\max} - D \rfloor}{T_f} \quad . \tag{5.2}$$

Here T_f is the frame duration. The total number of OFDM symbols required for UL feedback ($C_{UL,FB}$), and DL ARQ ($C_{DL,A}$) are given in (5.3) and (5.4) below. For a single SS, $S_{DL,E}^{(i)}$, $S_{DL,A}^{(i)}$ and $S_{UL,FB}^{(i)}$ are the number of symbols needed for initial transmission, ARQ retransmissions, and feedback respectively, using the i^{th} burst profile.

$$\begin{aligned} C_{UL,FB} &= S_{UL,FB}^{(i)} \sum_{i=0}^{n_{fb}} p^i \quad , n_{fb} = \left\lfloor \frac{n_d}{2} \right\rfloor - 1 \\ &= S_{UL,FB}^{(i)} \frac{(1-p^{n_{fb}})}{1-p} \end{aligned} \quad . \tag{5.3}$$

$$\begin{aligned} C_{DL,A} &= S_{DL,A}^{(i)} \sum_{i=0}^{n_a} p^i \quad , n_a = \left\lfloor \frac{n_d - 1}{2} \right\rfloor \\ &= S_{DL,A}^{(i)} \frac{p(1-p^{n_a})}{1-p} \end{aligned} \quad . \tag{5.4}$$

The total capacity used for this transmission, C is the sum of the effective capacity (C_E), $C_{UL,FB}$ and $C_{DL,A}$. If the ratio $C_{UL,FB}$:C is evaluated after an even number frames we have,

$$\frac{C_{UL,FB}}{C} = \frac{1}{1 + S_{DL,E}^{(i)} / S_{UL,FB}^{(i)}} , \tag{5.5}$$

which is a constant value for any set of symbol sizes. This is also its asymptotic value. For example, consider that most of a typical cell can be covered by the two highest

modulation schemes (2/3 QAM64 and 3/4 QAM64). At the highest modulation scheme, a VoIP packet can be transported using a single OFDM symbol (using 256 subcarriers). In this case, it can be shown that feedback bandwidth, $C_{UL,FB}$ can be between 33% and 50% of the used capacity regardless of packet error rate.

5.3 Proposed ARQ Scheme

In order to increase the efficiency of the ARQ mechanism for relatively small real-time packet flows, we propose Contention based Negative Acknowledgement ARQ which we will refer to as C-ARQ, (Perera and Sirisena 2006).

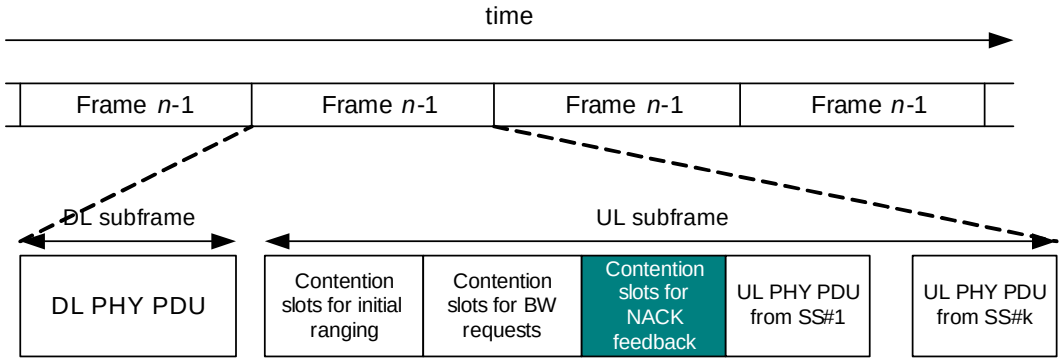

Figure 5-8 Modification of frame structure with additional contention channel for negative feedback of C-ARQ

5.3.1 Operation of C-ARQ

The operation of C-ARQ is described as follows.

1) The BS transmits packets on the DL channel to *n* SSs in a given frame.
2) Most subscribers will receive the packets without errors and will transition to idle mode. No explicit feedback is required for positive acknowledgements. The BS assumes correct reception until explicitly notified otherwise.
3) Any SSs which received the packets with errors (the probability of a packet error is given by *p*) will transmit a NACK over the NACK transmit channel, Figure 5-8, in the UL section of following frame. They will pick a slot randomly for this. The BS will allocate $\lambda.pn$ slots for this purpose. We will discuss the selection of λ in detail in following sections.
4) When a NACK is received by the BS the packet will be retransmitted in the next frame.

5) Those SSs who sent a NACK but did not receive a retransmission (due to collision of the NACKs) will resend a NACK in the next frame using the same random slot selection procedure. This process can be repeated until the packet deadline has been exceeded, or buffer constraints cause the SS to forward what is already received to the upper layers.

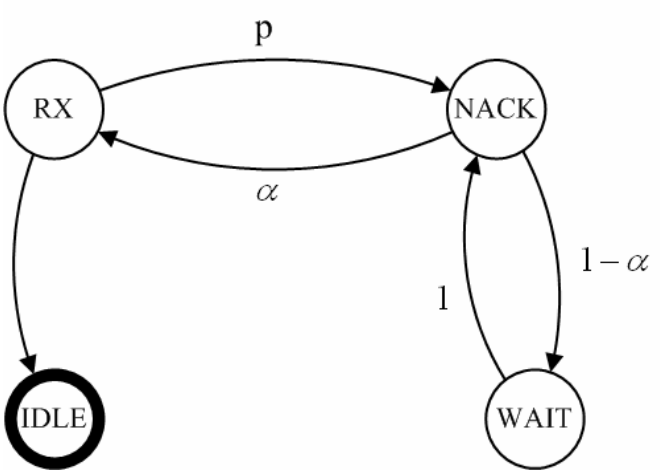

Figure 5-9 State diagram of C-ARQ scheme at the receiver (SS) end. α is the probability of successful feedback transmission using contention. All transitions occur at the frame boundary.

The state diagram for the above operation is given in Figure 5-9. We do not require analysis for the transmitter (BS) state machine because in this scheme the BS plays a reactive role responding to NACKs received from the SSs. λ is the number of contention slots allocated per SS (proportion of contention slots to SSs). $\lambda.pn$ is the total number of required slots for n number of users with p packet error rate. Here α is the probability that a transmitted NACK will be successfully received by the BS without a collision.

5.3.2 Analysis of Resource Usage by C-ARQ

The state diagram given in Figure 5-9 is represented in matrix form in (5.6). As said previously the states *R* and *N* refer to the Receive state (RX) and NACK transmit state respectively. The state *W* is when a SS has transmitted a NACK, did not receive a retransmission in the subsequent frame and is now waiting for the next contention based feedback opportunity to retransmit the NACK. As before, S_n represents the current state vector, *T* represents the state transition matrix and S_{n+1} represents the next state vector.

$$\underbrace{\begin{bmatrix} R & N & W & I \end{bmatrix}}_{S_n} \underbrace{\begin{bmatrix} 0 & p & 0 & \bar{p} \\ \alpha & 0 & \bar{\alpha} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{T} = \underbrace{\begin{bmatrix} \alpha N \\ pR + W \\ \bar{\alpha} N \\ I + \bar{p} R \end{bmatrix}^T}_{S_{n+1}} \tag{5.6}$$

$$S_n T = S_{n+1}$$

$$S_1 T^n = S_{n+1}$$

.

Equations (5.7) and (5.8) give the probability of being in NACK or RX state respectively, in the n^{th} frame. α is the proportion of SSs who successfully transmitted a NACK to the BS.

$$N^{(n)} = \begin{cases} p(\alpha p + \bar{\alpha})^i \ \ where\ i = \left\lfloor \dfrac{n-2}{2} \right\rfloor & ,n\ even \\ 0 & ,n\ odd \end{cases} \tag{5.7}$$

.

$$R^{(n)} = \begin{cases} 0 & ,n\ even \\ \alpha p(\alpha p + \bar{\alpha})^i \ \ where\ i = \left\lfloor \dfrac{n-2}{2} \right\rfloor & ,n\ odd \end{cases} \tag{5.8}$$

.

Now we quantify the total resource usage. Let us define the three components C_E, $C_{DL,A}$ and $C_{UL,FB}$:

C_E – resources used for the initial transmission on the DL. We specify the number of resource units for a single transmission by $S_{DL,E}$.

$C_{DL,A}$ – resources used for retransmission on the DL. We specify the number of resource units for a single transmission by $S_{DL,A}$.

$C_{UL,FB}$ – resources used for feedback on the UL. We specify the number of resource units for a single transmission by $S_{UL,FB}$.

The total resources used, C is the sum of the initial transmission, retransmissions and all bandwidth allocated for the feedback channel. This is given below.

$$\begin{aligned} C &= C_E + C_{DL,A} + C_{UL,FB} \\ &= S_{DL,E}^{(i)} + S_{DL,A}^{(i)} \sum_{j=3}^{n} R^{(j)} + \lambda S_{UL,FB}^{(i)} \sum_{j=1}^{n} N^{(j)} \end{aligned} \tag{5.9}$$

The resources used for feedback as a proportion of the total resources used, up to the n^{th} frame is,

$$\frac{C_{UL,FB}}{C} = \frac{\lambda S_{UL,FB}^{(i)} \sum_{j=1}^{n} N^{(j)}}{S_{DL,E}^{(i)} + S_{DL,A}^{(i)} \sum_{j=3}^{n} R^{(j)} + \lambda S_{UL,FB}^{(i)} \sum_{j=1}^{n} N^{(j)}} \quad . \tag{5.10}$$

We find a lower bound for this ratio as n tends to infinity. The validity of this approximation will be investigated in the following section with a numerical example.

$$\frac{C_{UL,FB}}{\lim_{n\to\infty} C} = \frac{p\lambda S_{UL,FB}^{(i)}}{\alpha\bar{p} S_{DL,E}^{(i)} + p\lambda S_{UL,FB}^{(i)} + \alpha p S_{DL,A}^{(i)}} \tag{5.11}$$

It is fair to assume $S_{DL,E}$ and $S_{DL,A}$ are an equal number of symbols since both values are related to DL transmissions by the BS. We do not allow modulation schemes to be changed during the process of retransmitting a packet. Based on this we simplify (5.11) further.

$$\frac{C_{UL,FB}}{\lim_{n\to\infty} C} = \frac{1}{\left(\frac{\alpha}{p\lambda}\right)\left(S_{DL,E}^{(i)} / S_{UL,FB}^{(i)}\right)+1} \quad . \tag{5.12}$$

5.3.3 Optimal Selection of α and λ

These two parameters are closely linked and one has to be fixed to calculate the other. More precisely the number of contention slots assigned by the BS will determine the value of α. If λ is increased arbitrarily by the BS, we may achieve a very small collision rate in contention at the expense of the bandwidth used on the contention slots. It is obvious from (5.12) that the proportion of feedback bandwidth is proportional to λ, and inversely proportional to $\alpha k + 1$, where k is a constant. Our requirements here, are to obtain a satisfactory rate of packet completion while reducing overheads for feedback.

We assume the DL packet error rate p can be accurately estimated by the BS so that it will be known what percentage of packets will be erroneously received. If n packets were transmitted on the DL and n_r of those were received with errors by the SSs we can assume that n_r SSs will contend for feedback slots in the following frame. If there are M feedback slots provided by the BS, using the theory of occupancy (Johnson and Kotz 1977) we can find the expected number of slots with r SSs transmitting, (5.13).

$$. \quad a_r = M\binom{n_r}{r}\left(\frac{1}{M}\right)^r\left(1-\frac{1}{M}\right)^{n_r-r} \tag{5.13}$$

As per the C-ARQ scheme all SS with receive errors will transmit NACKs using contention. A value of r more than 1, implies two or more stations transmitting feedback in the same slot. This is a collision scenario. Then the expected number of collisions during the feedback contention period will be,

$$\begin{aligned} E[n_c] &= \sum_{r=2}^{n_r} r a_r \\ &= n_r - n_r\left(1-\frac{1}{M}\right)^{n_r-1} \end{aligned} \quad . \tag{5.14}$$

Hence the expected proportion of successful feedback transmissions α, which is referred to as *Throughput* in (Sung-Min and Jae-Hyun 2005) is,

$$\begin{aligned} \alpha &= \frac{n_r - E[n_c]}{n_r} \\ &= \left(1-\frac{1}{M}\right)^{n_r-1} \end{aligned} \quad . \tag{5.15}$$

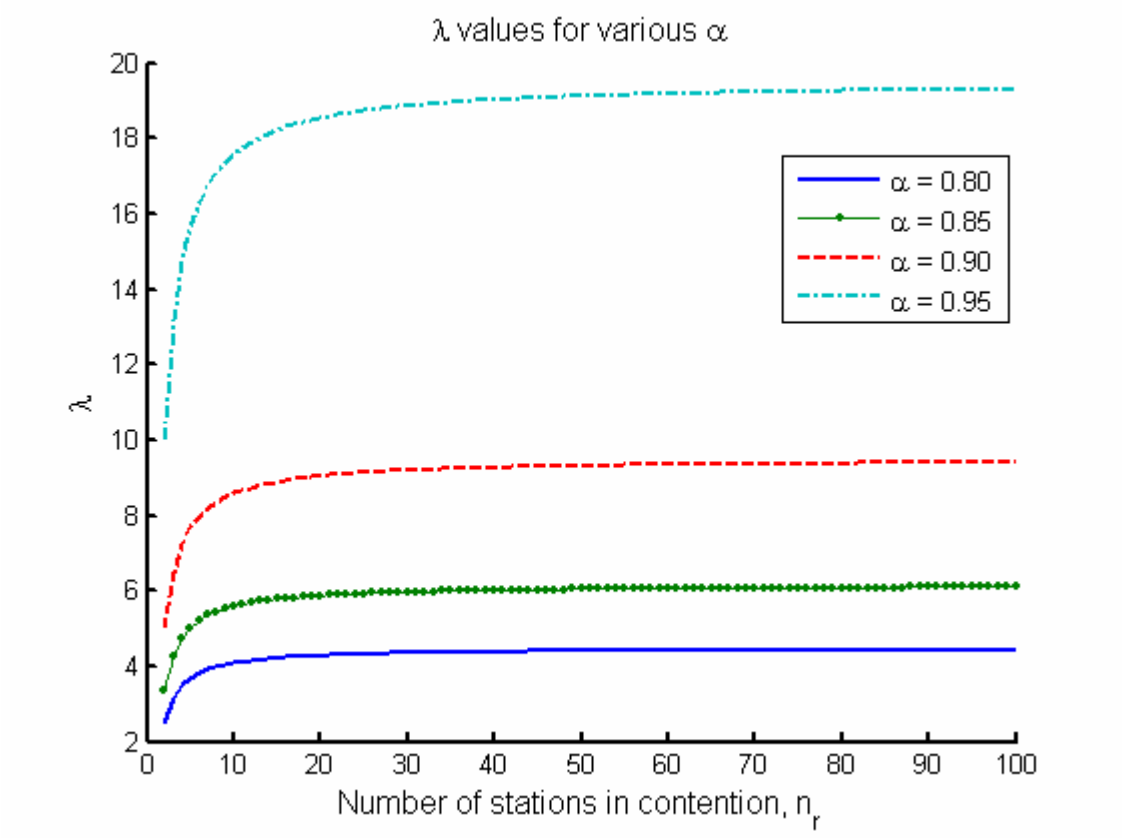

Figure 5-10 λ_{min} for a range of n_r values and 4 different α values.

For a given (or required) α, the minimum λ (λ_{min}) value can be found as the ratio of contention slots per contender. The expression for M is obtained from (5.15).

$$\lambda_{\min} = M \Big/ n_r = \frac{1}{n_r[1-\alpha^{1/(n_r-1)}]} . \tag{5.16}$$

The λ of (5.16) is plotted in Figure 5-10. Our interest lies in the initial section of the curves which represent a few tens of erroneous receptions.

We can also find an approximate upper bound $\lambda_{min,up}$ (such that $\lambda_{min,up} > \lambda_{min}$ for all n_r) by finding the limit of λ_{min}.as n_r tends to infinity.

$$\lim_{n_r \to \infty} \lambda_{\min,up} \approx \frac{1}{1-\alpha} . \tag{5.17}$$

Now we compare the overhead in the standard feedback scheme, and our proposed scheme. Consider (5.5) and (5.12). Lower feedback overhead is achieved in the proposed scheme when the condition in (5.18) is satisfied.

$$. \ \lambda < \alpha \Big/ p \tag{5.18}$$

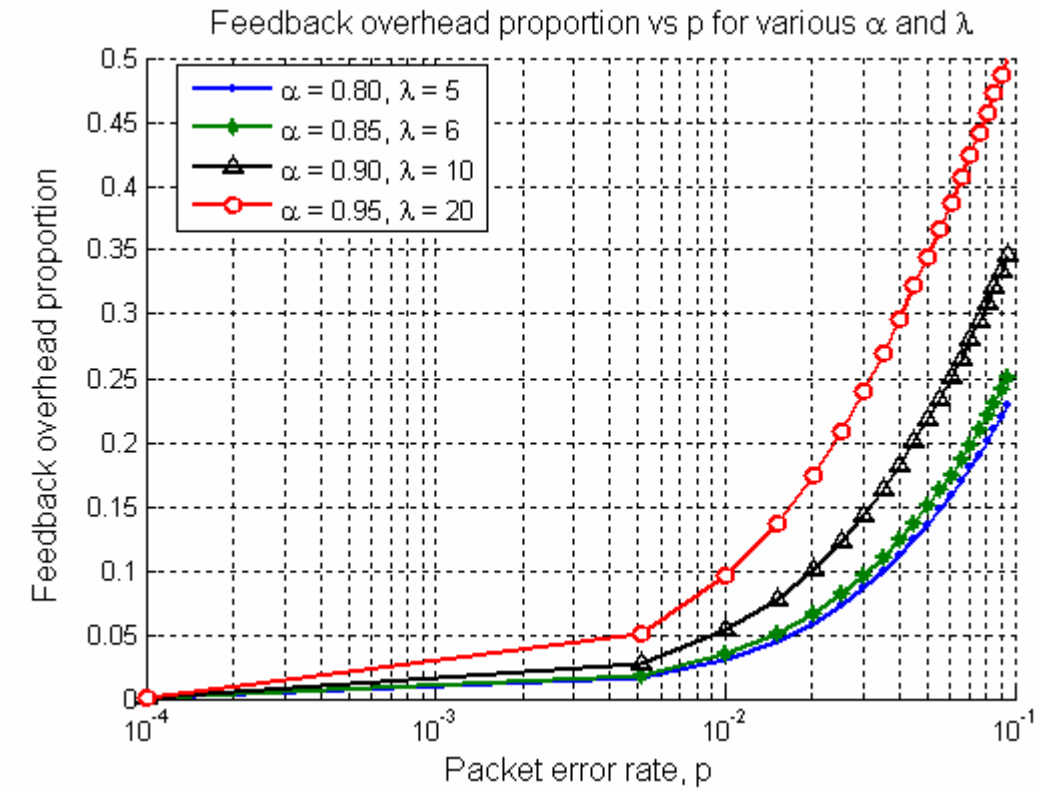

Figure 5-11 Feedback overhead proportion for a range of *p* values and α values.

5.3.4 Analytical Comparison of Performance

We select α from 0.95 to 0.80 (95% to 80% throughput on the contention feedback) to evaluate the feedback overhead for this scheme. The optimal value of λ should be calculated dynamically but here for simplicity of the analysis we use these fixed values. As can be seen in Figure 5-11 the overhead used increases with the packet loss rate. This can be expected due to the fact that our ARQ feedback scheme only sends NACKs which use a contention channel. Higher the packet loss rate, the more NACKs need to be sent and more contention opportunities given. However the overhead percentage remains at a lower value than that of the standard schemes. This is also using approximations for λ which give values that are larger than actually required. When p is less than 10^{-2}, the overheads used is approximately 10% which is very much lower than the standard. At *p=0.1* we can still keep the overhead percentage below the threshold of the standard scheme, by using a lower λ value. This however increases the time taken to reach the required level of completion determined by the QoS parameters of the flow.

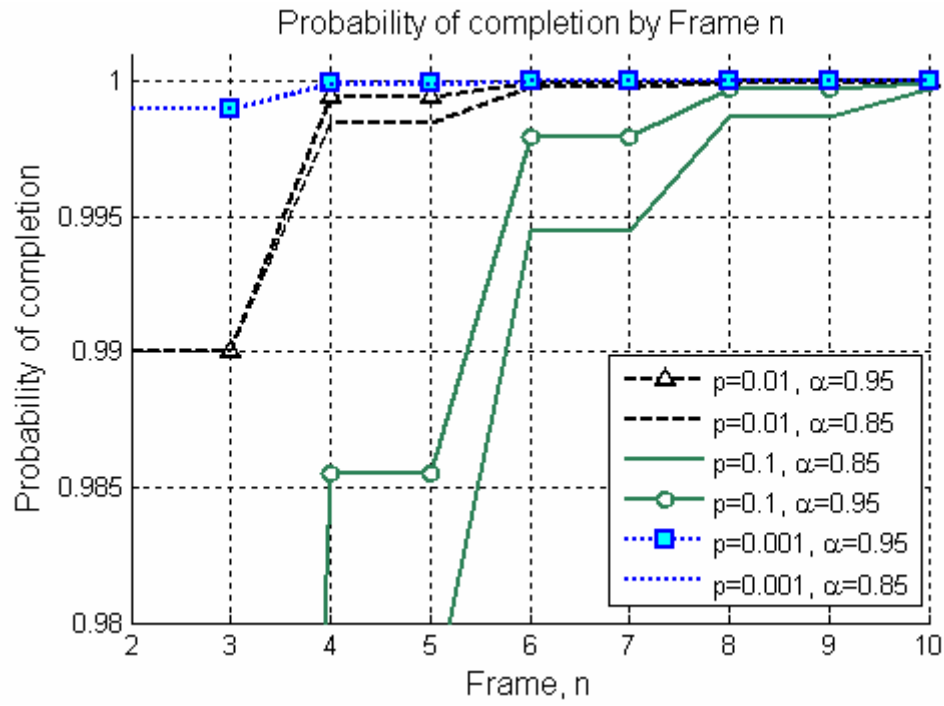

Figure 5-12 Probability of completion of C-ARQ scheme against the elapsed frame number. 3 different *p* values are shown.

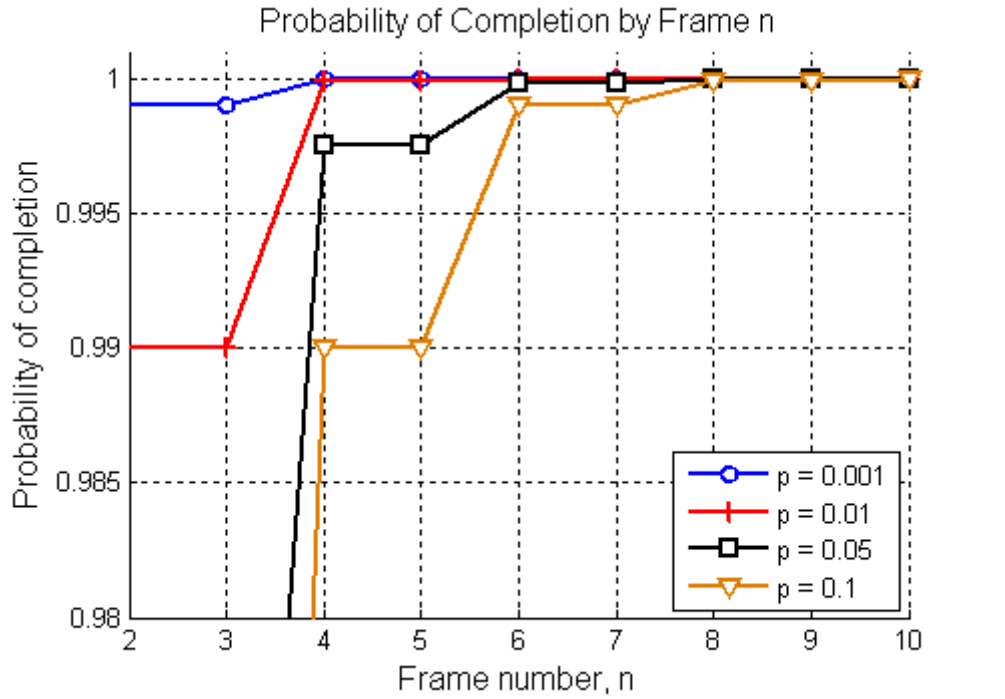

Figure 5-13 Probability of completion of standard ARQ scheme against the elapsed frame number.

We compare the time required in frames to reach a certain level of successful delivery. In real-time flows the application is able to cope with a relatively high packet loss rate as long those delivered are done so with minimal latency.

In Figure 5-12, if we assume a packet delivery guarantee of 99% (Hattingh and Szigeti 2004), 6 frames are required at $\lambda=0.85$ and $p=0.1$ at an overhead of 24%. The standard scheme requires 4 frames in the same situation with an overhead of 33%, Figure 5-13.

The results summarized in Table 5-1 show that our scheme is more efficient for most of the reasonable packet error rates. The shaded rows represent p-α combinations where the efficiency is less than the current method. Although a 10% packet loss rate is considered here it is only for completeness sake. A step down in burst profiles will be done (initiated by the BS or the SS) according to the IEEE 802.16d so that data transfer occurs in a more efficient manner.

Table 5-1 The overhead percentage and frame delay is summarized for a few p values and 3 selected α values.

p	α,λ	Overhead %	99% delivery delay(frames)
0.001	0.85, 6	0.3	1
	0.90, 10	0.5	1
	0.95,20	1	1
0.01	0.85, 6	3	3
	0.90, 10	5	3
	0.95,20	9	3
0.1	0.85, 6	26	8
	0.90, 10	35	7
	0.95,20	51	6

5.4 Simulation Study

A simulation study has been done to verify the effectiveness of the proposed ARQ scheme. We build upon the functionality of the basic simulation model, described in Chapter 3, to accommodate the modifications of C-ARQ

5.4.1 Simulation Scenario

The scenario is defined as follows.

- 100 users spread out within the cell with uniform aerial density.
- Traffic flows are from sources external to the cell area. There is a constant bit rate (CBR) flow to each of the 100 users. The packet size used is consistent with the interpacket interval and header sizes of a VoIP flow.
- Interpacket durations of 20 ms, 30 ms and 40 ms are simulated. There are two implications due to the interpacket duration.

We use a data stream of 8kbps. As the duration increases so does the packet size. Certain users, depending on the burst profile will need more OFDM symbols to receive the packet. The lower the modulation scheme the higher the number of symbols and vice versa. In our simulated cell users belong to one of the highest 4 burst profiles. At 20 ms and 30 ms SS using the highest 3 burst profiles can receive the packets using one OFDM symbol. (C_E = 1, $C_{DL,A}$ = 1, $C_{UL,FB}$ = 1). This represents

80% of the cell area. The SS using the lowest burst profile will require two symbols to receive a packet of the same size (C_E = 2, $C_{DL,A}$ = 2, $C_{UL,FB}$ = 1). This represents 20% of the cell area. When the interpacket interval of 40 ms is used these cell area percentages are changed to 21% and 79% respectively with a larger area of the cell requiring two symbols per packet.

The other effect is that the number of available frames in which to recover a lost packet changes. At 20 ms there are 4 frames for recovery while at 40 ms there are 9 frames. At the same time, longer recovery periods cause more jitter. However the larger the packet the more important each one is.

- Overheads per frame are continuously recorded and the mean value is calculated for a 20 minute period excluding the transient period at the beginning of the simulation.
- Packet drop rates of 0.001, 0.005, 0.01, 0.05 and 0.1 are simulated.
- α values of 0.8, 0.85, 0.9 and 0.95 are simulated. Lambda is calculated based on these.
- The system uses a TDD frame structure with a frame duration of 4 ms.

5.4.2 Assumptions and limitations

- A packet will be retransmitted by the BS if a request is received from the SS (NACK) as many times as possible until the next packet of the flow becomes available, ready for DL. The maximum number of retransmissions depends on the interpacket duration.
- Packet drops at the SS are artificially introduced at the MAC layer by a random packet error generator just as they are received from the PHY. We apply the errors only to data packet of the flow under test. All management type packets such as ranging, dynamic service addition/change/deletion and map messages will not be subject to packet drops. This makes it easier to isolate the performance of C-ARQ.
- Ranging is carried out at the time a SS powers on and then every 10 seconds.
- The aggregate packet loss rate of all the SS is estimated by the BS and a moving total of possible erred packets with a window size equal to the interpacket gap is used to calculate the number of contention slots for NACK feedback.
- The UL is assumed to be error free.

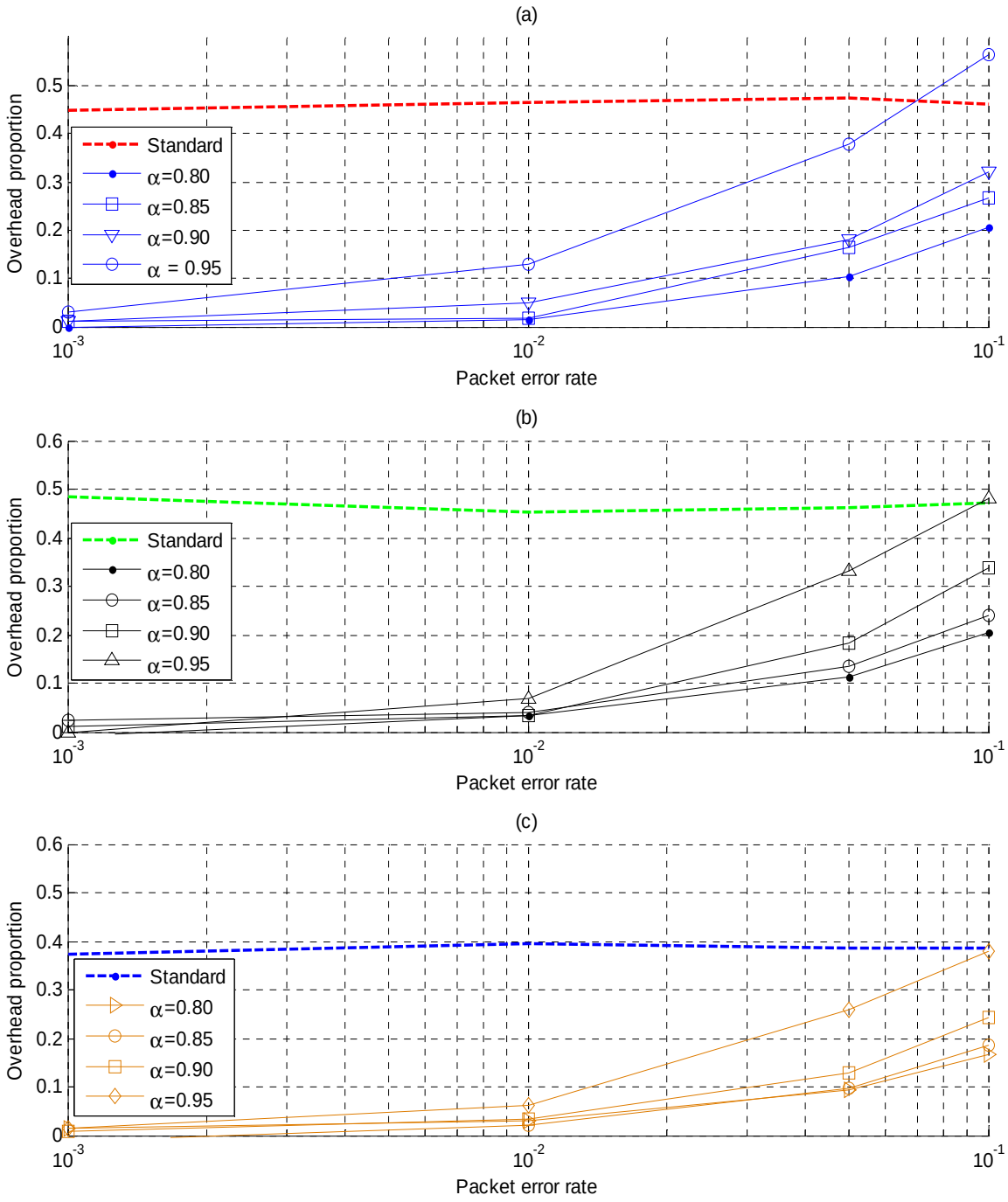

Figure 5-14 Feedback overhead proportion of standard scheme compared with the C-ARQ scheme. (a) 20 ms interpacket duration (b) 30 ms interpacket duration (c) 40 ms interpacket duration.

5.4.3 Simulations Results

The simulation results show the feedback overhead percentage for the combinations of the test matrix detailed in the Simulation Scenario section.

The subplots (a) and (b) show an equal overhead percentage for the standard scheme. This is the expected behaviour as per (5.5). The results for all three test cases follow the form of Figure 5-11 as given in the analysis section.

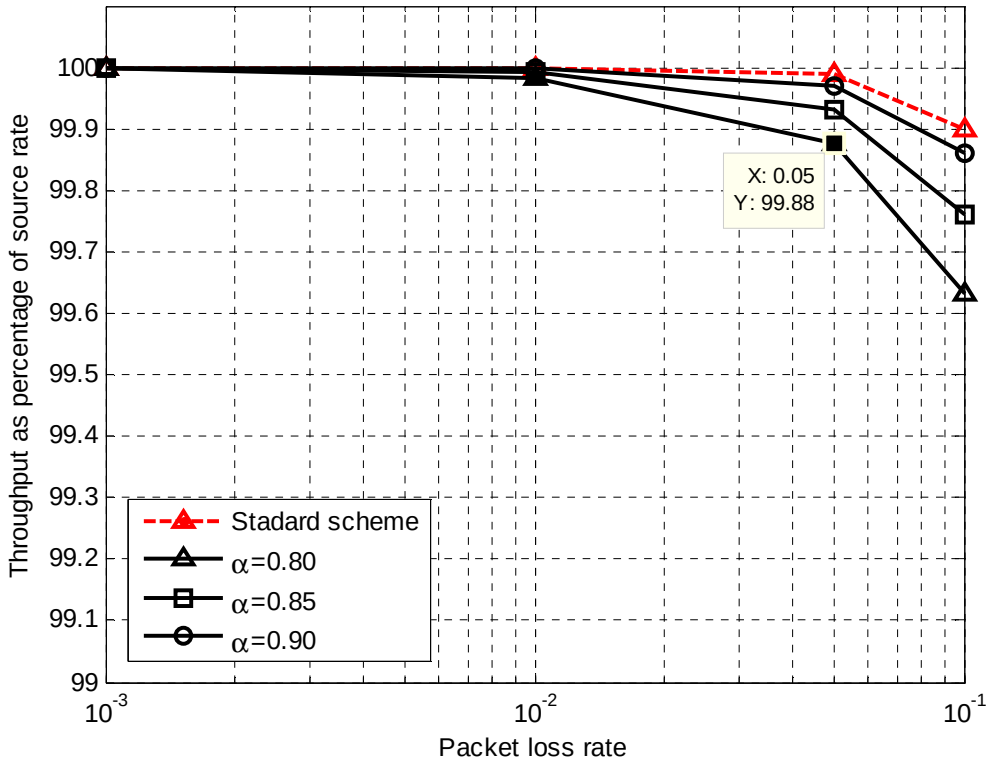

Figure 5-15 SS Throughput as a percentage of source traffic rate for standard scheme compared with the C-ARQ scheme.

The throughput as a percentage of source traffic can be interpreted as the completion rate of delivery. C-ARQ shows almost 99.9% rate of completion for packet error rates up to 5%, Figure 5-15. It should be noted that the standard scheme performs better than C-ARQ in terms of completion rate. This is however at the expense of a comparatively much higher overhead, as described in previous sections.

5.5 Conclusion

In this thesis we have analyzed the ARQ scheme used in WiMAX in the context of downlink real-time flows of small packets such as in VoIP. Irrespective of the packet error rate of the link, a substantial proportion of the bandwidth is used for the feedback messages. On the uplink it is not possible to concatenate packets of different subscriber stations. Transmitting many small packets makes the feedback process very bandwidth hungry and inefficient. However without ARQ, we cannot guarantee any level of QoS. To improve on this without sacrificing performance we have proposed a contention based negative acknowledgement ARQ (C-ARQ) scheme. The defining difference between this scheme and many others as well as the standard scheme implemented in WiMAX, is the feedback mechanism. Subscribers who received erroneous packets contend to send feedback to the base station. We have analytically proven the viability of this scheme in terms of overhead bandwidth usage and rate of completion. The analysis has been validated with simulations which show a very good performance improvement. It has been shown that for packet error rates lower than 10%, our scheme is more efficient at the expense of a small increase in delay.

Chapter 6

Contention Based Access for Best Effort Traffic

According to the standard (IEEE 802.16 WG 2004), four classes of scheduling have been defined on the UL. Of these, UGS and rtPS can be used to schedule uplink real-time traffic such as VoIP and streaming variable rate video. Lower priority traffic such as file transfer and web browsing (e.g. FTP and HTTP), would be serviced as BE traffic. On the DL, the BS has the responsibility of scheduling the packets so that they reach the SS before the expiration deadline. This of course depends on the type of service flow the packets belong to, and whether the BS can differentiate these flows based on fields in the network layer or transport layer headers, or pattern matching of the payload itself. BE traffic being of the lowest priority means that only the remaining bandwidth (BW) after servicing all higher priority flows will be allocated to it. WiMAX employs a contention based method with exponential backoff to facilitate multiple access for BE traffic. BW requests need to be sent to the BS using contention. This adds overhead in the form of contention slots and latency due to collision and backoff. To guarantee a certain level of success in contention, the BS needs to allocate a suitable number of contention slots (Sung-Min and Jae-Hyun 2005).

We concentrate on analysing the service of BE flows over Fixed WiMAX, while reducing contention overheads and BW wastage due to underutilization. The rest of the chapter is organized as follows. An overview of the operation of the BE

service class is given followed by a Markov Chain model. Next, we provide an approximate throughput analysis of TCP bulk transfer flows using the BE class. Simulation results and validations of the model are also provided, as well as results showing the application of this analysis to increase efficiency for TCP based traffic.

6.1 Previous Work on Contention Based Access Techniques

Here we list some of the most widely used MAC protocols which have comparable contention resolution techniques, namely IEEE 802.11 (Wi-Fi) and Data-Over-Cable Service Interface Specification (DOCSIS).

6.1.1 IEEE 802.11 Contention Resolution

In the 802.11 protocol, the fundamental mechanism to access the medium is the distributed coordination function (DCF), which is a random access scheme, based on the carrier sense multiple access with collision avoidance (CSMA/CA) protocol. Retransmission of collided packets is managed according to binary exponential backoff rules. The standard also defines an optional point coordination function (PCF), which is a centralized MAC protocol able to support collision free and time bounded services.

DCF describes two techniques to employ for packet transmission. The default scheme is a two-way handshaking technique called basic access mechanism. This mechanism is characterized by the immediate transmission of a positive acknowledgement (ACK) by the destination station, upon successful reception of a packet transmitted by the sender station. Explicit transmission of an ACK is required since, in the wireless medium, a transmitter cannot determine if a packet is successfully received by listening to its own transmission. In addition to the basic access, an optional four way handshaking technique, known as request-to-send/clear-to-send (RTS/CTS) mechanism has been standardized. Before transmitting a packet, a station operating in RTS/CTS mode “reserves” the channel by sending a special RTS short frame. The destination station acknowledges the receipt of an RTS frame by sending back a CTS frame, after which normal packet transmission and ACK response occurs. Since collision may occur only on the RTS frame, and it is detected by the lack of CTS response, the RTS/CTS mechanism allows increasing the system

performance by reducing the duration of a collision when long messages are transmitted. As an important side effect, the RTS/CTS scheme designed in the 802.11 protocol is suited to combat the so-called problem of Hidden Terminals (Bianchi, Fratta et al. 1996), which occurs when pairs of mobile stations are unable to hear each other. This problem has been specifically considered in (Kai-Chuang and Kwang-Cheng 1995; H. S. Chhaya 1997), which, in addition, studies the phenomenon of packet capture.

In the literature, performance evaluation of 802.11 has been carried out either by means of simulation (Kanjanavapastit and Landfeldt 2003; Yang 2003; Bruno, Conti et al. 2005; Khanna, Gupta et al. 2005) or by means of analytical models with simplified backoff rule assumptions. In particular, constant or geometrically distributed backoff window has been used in (Wanang and Yin 2005; Amjad and Shami 2006) considered an exponential backoff limited to two stages (maximum window size equal to twice the minimum size) by employing a two dimensional Markov chain analysis. In (Bianchi 2000), an extremely simple model that accounts for all the exponential backoff protocol details, and allows to compute the saturation (asymptotic) throughput performance of DCF for both standardized access mechanisms (and also for any combination of the two methods) has been devised. The core contribution of (Bianchi 2000) is the analytical evaluation of the saturation throughput, in the assumption of ideal channel conditions. The analysis, assumes a fixed number of stations, each always having a packet available for transmission, i.e., *saturation* conditions where the transmission queue of each station is always nonempty.

6.1.2 DOCSIS Contention Resolution

The Data-Over-Cable Service Interface Specification (Cable Television Laboratories Inc 2007), is a communication protocol designed for high-speed packet-based communications over cable television communication networks, but which can also form the basis of wireless communication systems. At the user end of the communication system a headend is linked to several Customer Premises Equipment units (CPEs) through several communication links. The headend acts as a distribution hub into a wide area network for the CPEs to which it is linked. Each CPE may

manage more than one service flow. In a cable television communication network the headend, communication link, and CPE are a Cable Modem Termination System (CMTS), a coaxial cable, and a Cable Modem (CM) respectively. In a wireless communication network the headend, communication link, and CPE are a BS, a radio channel, and a SS respectively.

Under DOCSIS, the headend transmits Bandwidth Allocation Map (MAP) messages on a regular basis to all CPEs to which it is linked. A MAP message is a management message that the headend uses to announce and allocate transmission opportunities to the CPEs. This is similar to the MAP messages used in WiMAX. A section of the frame is dedicated for broadcast type bandwidth requests, in which case any CPE may transmit a request, for any service flow. The CPE calculates a random offset based on parameters communicated to it by the headend. After expiry of the random offset the CPE transmits a request for bandwidth for a data-transmission for a service flow. The request indicates the number of bytes remaining in the transmission queue of the service flow. The random offset is necessary to minimize request collisions between different CPEs transmitting requests in response to the same broadcast-type Request IE. When the headend receives the request, the headend attempts to schedule the requested data transmission within the interval between transmission of the next two MAP messages (the next MAP period). If the requested, data transmission can be accommodated within the next MAP period, the headend alerts the CPE that the request was granted using a Data Grant IE in the next MAP message. If the request for transmission cannot be accommodated within the next MAP period, the headend alerts the CPE that the request is pending using a zero length Data Grant IE (referred to as a Data Grant Pending IE) in the next MAP message. When the headend receives a request from a CPE, the head unit must send either a Data Grant IE or a Data Grant Pending IE to the CPE in the subsequent MAP message. This procedure of explicit success notification is in contrast to the method used in WiMAX.

If requests from more than one CPE arrive at the headend simultaneously, the requests will collide. All of the requests will be lost, unless the headend is implementing a capture effect in which case all but one request will be lost. The headend does not detect the lost requests, and is therefore unable to send a Data Grant

IE or Data Grant Pending IE to a CPE whose request was lost. When a CPE does not receive a Data Grant IE or a Data Grant Pending IE in the MAP message subsequent to the transmission of the request, the CPE realizes that a contention arose and that its request was lost. According to the truncated binary exponential back-off algorithm, which is used by DOCSIS, the CPE must double its backoff window after every contention failure. The CPE continues this contention resolution method until it receives a Data Grant IE or Data Grant Pending IE, a maximum number of retries have been attempted (at which point the packet of data is discarded), or the CPE receives a unicast-type Request IE in which to transmit the payload. Work has been done to analyse through simulation the contention resolution scheme in terms of access delay and attainable throughput, (Heyaime-Duverge and Prabhu 2002; Jianxin and Speidel 2003; Wei-Tsong, Kun-Chen et al. 2006). Further analysis has been carried out in similar fashion using Markov chains, (Seung-Eun, Oh-Hyeong et al. 2006; Kai-Chien and Wanjiun 2007).

6.2 Analysis of Best Effort Service Class in WiMAX

In this section we provide a detailed description of how the BE service class operates. A two dimensional Markov chain model is used to accurately reflect the contention mechanism described in the standard. Based on the Markov model we have produced a mathematical model to analytically express channel availability to SSs, efficiency of access, and latency.

6.2.1 Detailed Operation

The operation of the BE class can be broken down in to two area, (1) the process of transmitting a bandwidth request (BW-REQ), and (2) the process of contention resolution in the case of collision of the transmitted BW-REQ.

6.2.1.1 Requesting Bandwidth

As stated before, the BE class is based on a contention based request-grant mechanism. There are two contention methods described in the standard. The mandatory method is based on non-subchannelization capable equipment, while the optional method requires the BS as well as SSs to be subchannelization capable.

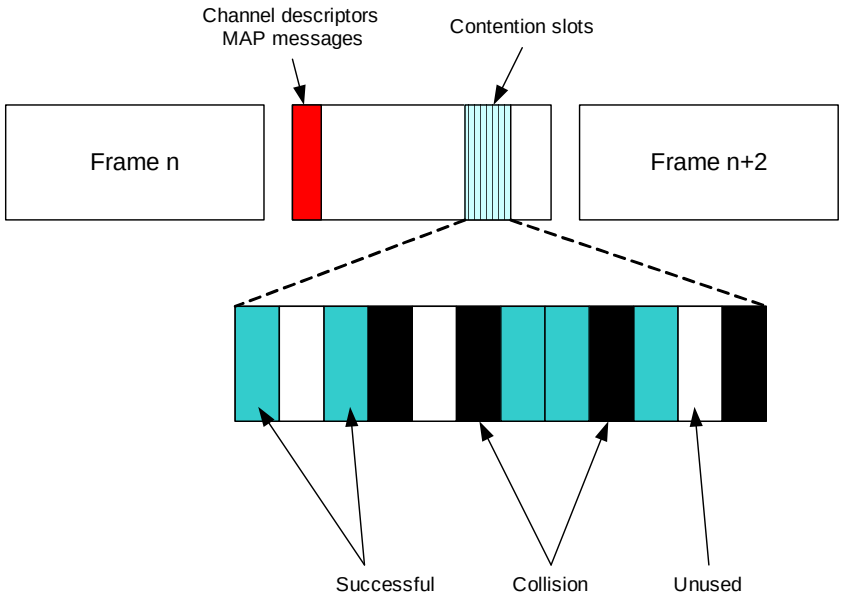

Figure 6-1 Contention slots in a frame. Collisions and successes can be randomly distributed in the contention region. Some slots remain unused.

When a packet becomes available to be transmitted at the SS MAC layer it will be queued in the appropriate queue depending on its IP precedence. All SSs will receive the broadcast part of the frame which contains among other frame information, the UL-MAP. The SS will scan through the UL-MAP for an element known as the 'REQ Region-Full' which is signified by the Uplink Interface Usage Code of 4 (UIUC = 4). This region consists of contention slots. The width of a slot (in time) is explicitly communicated to all SSs by the BS using the Uplink Channel Descriptor (UCD) message. The SS randomly picks a slot out of those given in the frame and transmits a BW-REQ PDU. A BW request PDU is constructed with a preamble and a 6 Byte bandwidth request MAC header, which includes the CID (which uniquely identifies the SS and the UL flow) and the aggregate BW required in Bytes. No payload is included in this message.

No sensing of the medium is done as in Wi-Fi. If contention was successful, it is implicitly made know to the SS by a BW grant, denoted by a 'Data Grant Burst Type IE' in the UL-MAP, which contains the Basic CID of the SS. This IE contains the start time and duration in OFDM symbols of the allocation. It is now up to the SS to transmit its PDU within the specified allocation. BW-REQs received through contention will be serviced as per delay requirements of the flow or agreed upon

values of maximum latency, although for BE traffic this ultimately depends on the available BW after higher priority traffic has been serviced.

6.2.1.2 Contention Resolution

The SS considers the contention transmission lost, if no data grant has been given within BW-REQ timeout period (denoted by T16). The SS shall now increase its backoff window by a factor of two, as long as it is less than the maximum backoff window. The SS shall randomly select a number within its new backoff window and repeat the deferring process described above.

This retry process continues until the maximum number (i.e., Request Retries for bandwidth requests and Contention Ranging Retries for initial ranging) of retries has been reached. At this time, for BW-REQs, the PDU shall be discarded. The maximum number of retries is independent of the initial and maximum backoff windows that are defined by the BS and is set by the BS.

For bandwidth requests, if the SS receives a unicast Request IE or Data Grant Burst Type IE at any time while deferring for this CID, it shall stop the contention resolution process and use the explicit transmission opportunity.

The BS has much flexibility in controlling the contention resolution. At one extreme, the BS may choose to set up the Request Backoff Start and Request Backoff End to emulate an Ethernet-style backoff with its associated simplicity and distributed nature, as well as its fairness and efficiency. This would be done by setting Request Backoff Start = 0 and Request Backoff End = 10 in the UCD message. At the other end, the BS may make the Request Backoff Start and Request Backoff End identical and frequently update these values in the UCD message so that all SS are using the same, and hopefully optimal, backoff window. The UCD message also contains the following parameters relevant to the contention process.

Request Backoff Start: $CW_{\min}$ Initial backoff window size for contention, expressed as a power of 2 in the range of 0–15.

Request Backoff End: $CW_{\max}$ Final backoff window size for contention, expressed as a power of 2 in the range 0–15.

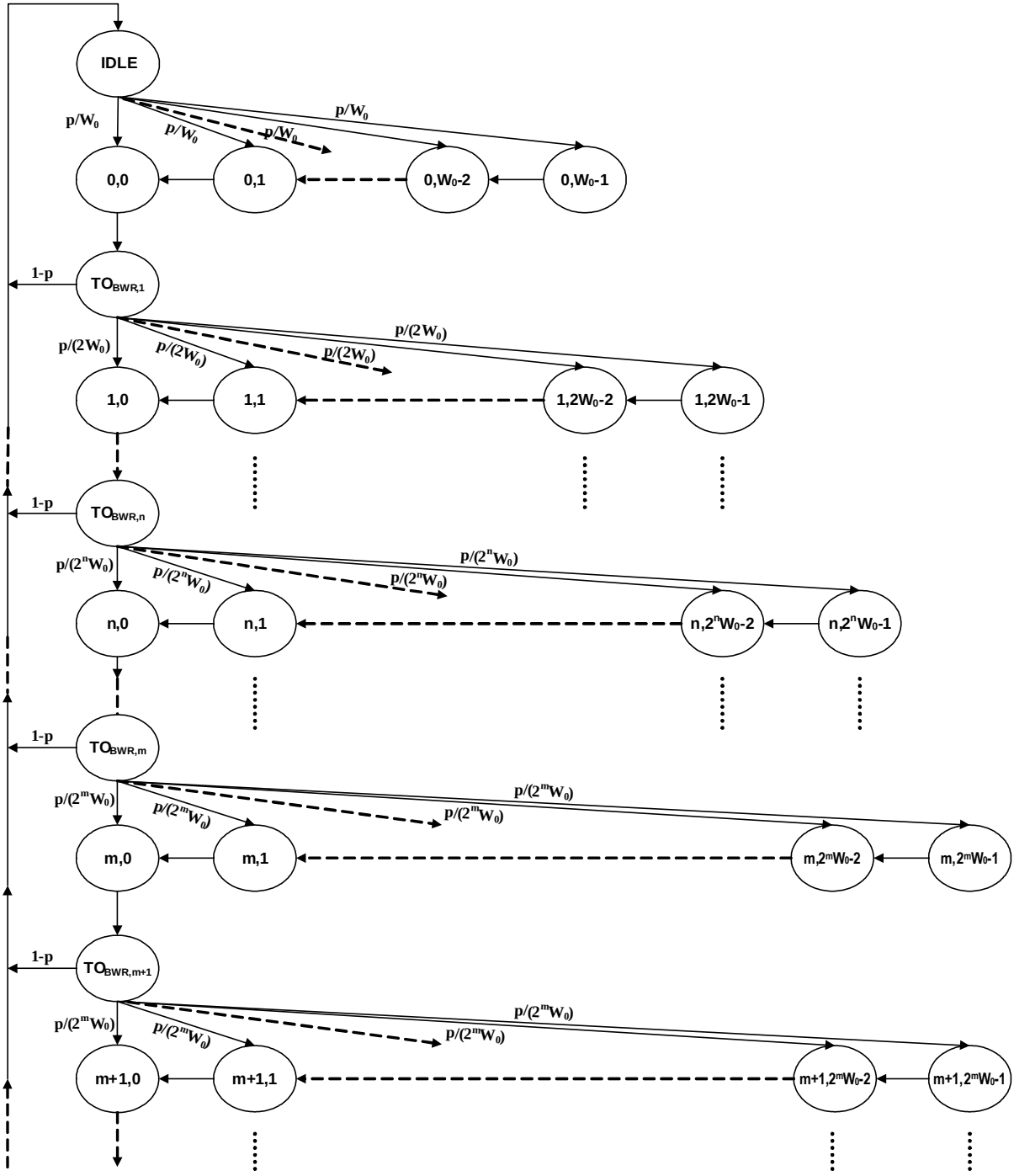

Figure 6-2 Two dimensional Markov chain for the contention and backoff algorithm used in REQ Region-Full type contention. State transitions occur with frame transitions, except from the TO states which will contain in integer number of idle/wait frames.

6.2.2 Markov Chain Model

We develop a two-dimensional discrete time Markov chain to accurately model the contention resolution process described above. This is given in Figure 6-2. The notations used are defined as follows.

p – probability of a collision during contention

W_0 – initial backoff window which is given by $2^{CW_{\min}}$

$TO_{,BWR,n}$ – BW-REQ time out state after nth contention attempt. Since state transitions take place on the frame transitions this state actually represents multiple states. We have given an expanded view of this in Figure 6-3.

The SS randomly chooses the number of frames to backoff from the range [0,W_0-1]. After each unsuccessful request, the contention window is doubled, up to a maximum value. W_i is the window for the ith retransmission of the BW-REQ. Once a BW allocation is received the window is reset to W_0.

6.2.3 Mathematical Representation of Markov Model

The state $\{s(t), b(t)\}$ in the model are define as follows. $s(t)$ is the backoff stage of a bandwidth request at time t. This is equal to the number of collisions suffered by the request so far. $b(t)$ represents the backoff counter at time t. When the backoff counter is decremented to zero the SS may send/resend the request. In Bianchi's model (Bianchi 2000) the behaviour of a single station is modelled using a Markov chain and the stationary probability τ that a station transmits in a generic time slot is obtained. The state transition probabilities are defined by $P\{i_1, k_1 \mid i_0, k_0\}$ where,

$$\begin{aligned} i_0 &= s(t) \\ k_0 &= b(t) \\ i_1 &= s(t+1) \\ k_1 &= b(t+1) \end{aligned} \tag{6.1}$$

We define the individual transition probability groups.

$$\left.\begin{aligned} &P\{0, k \mid IDLE\} = 1/W_0, && k \in [0, W_0 - 1] \\ &P\{i, k \mid TO_{BWR,k}\} = p/W_i, && i \in [1, m_{\max}], k \in [0, W_i - 1] \\ &P\{i, k \mid i, k+1\} = 1, && i \in [1, m_{\max}], k \in [0, W_i - 1] \\ &P\{IDLE \mid i, 0\} = 1 - p, && i \in [0, m_{\max} - 1] \\ &P\{IDLE \mid m_{\max}, 0\} = 1 \end{aligned}\right\} \tag{6.2}$$

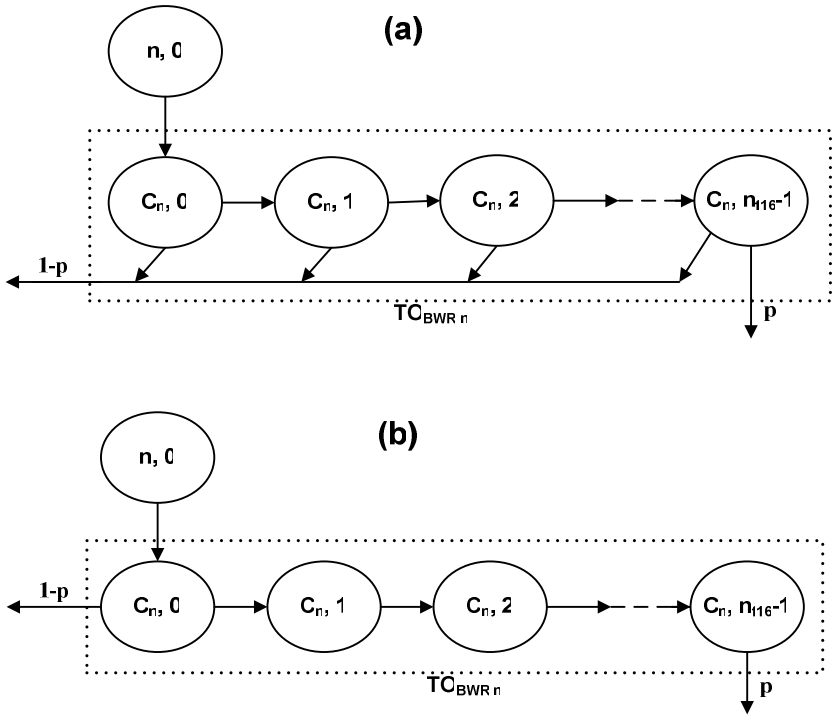

Figure 6-3 The states which make up $TO_{BWR,n}$ are shown here. The BS could service a received request in the immediately following frame or, any frame before the T16 timeout, depending on BW usage at the time. This is approximated by (b) in which all received BW requests are serviced in the next frame. n_{f16} is the T16 timeout period expressed in frames

As in (Bianchi 2000), let $b_{i,k}$, where $i \in [0, m)$ and $k \in [0, W_i)$, be the stationary distribution of the chain.

$$b_{i,k} = \lim_{t \to \infty} P\{ s(t) = i, b(t) = k \} \tag{6.3}$$

In addition, for every $b_{i,0}$ state there are n_f states $c_{i,j}$ where the SS waits to be serviced. Figure 6-3(a) gives the accurate state diagram and Figure 6-3(b) gives the approximation used for the analysis, which is that all requests received in a frame are serviced in the following frame. We substitute,

$$\begin{aligned} c_{i,0} &= b_{i,0} \\ c_{i,j} &= p.c_{i,0} \quad , \quad j \in [i, n_f - 1] \end{aligned} \tag{6.4}$$

where n_f is the number of frames a SS waits in idle mode for a T16 to expire, i.e., $n_f = ceil\left(T_{16}/T_f\right)$. In other words states $c_{i,j}$ are idle states. Imposing the normalizing condition on all the states and expressing $b_{i,k}$ in terms of $b_{i,0}$, we get

$$
\begin{aligned}
1 &= \sum_{i=0}^{m}\left(\sum_{j=0}^{n_f-1} c_{i,j} + \sum_{k=0}^{W_i-1} b_{i,k}\right) \\
&= \sum_{i=0}^{m}\left((1-p+p.n_f).b_{i,0} + \sum_{k=0}^{W_i-1} b_{i,k}\right) \\
&= \sum_{i=0}^{m}\left((1-p+p.n_f).p^i b_{0,0} + p^i b_{0,0}.\frac{W_i+1}{2}\right)
\end{aligned}
\tag{6.5}
$$

This gives,

$$
b_{0,0} = \frac{2(1-p)(1-2p)}{(1-2p)+pW_0[1-(2p)^m]+2(1-p+p.n_f)(1-2p)(1-p^{m+1})} \tag{6.6}
$$

We can now express the probability that a SS transmits a BW-REQ in a frame, τ. As any transmission occurs when the backoff counter is equal to zero, regardless of the backoff stage, τ is given as,

$$
\begin{aligned}
\tau &= \sum_{i=0}^{m} b_{i,0} = b_{0,0}\left(\frac{1-p^{m+1}}{1-p}\right) \\
&= \frac{2(1-2p)(1-p^{m+1})}{(1-2p)(W_0+1)+pW_0[1-(2p)^m]+2(1-p+p.n_f)(1-2p)(1-p^{m+1})}
\end{aligned}
\tag{6.7}
$$

where m is the retransmission limit. If we assume infinite retries, the above equation can be simplified to,

$$
\tau = \frac{2}{(W_0+1)+pW_0\sum_{i=1}^{\infty}(2p)^i + 2(1-p+p.n_f)} \tag{6.8}
$$

The assumption that the maximum retry limit is not a finite value can be justified because in the case where a SS has been unsuccessful in the past n times it would not be productive to simply drop the head-of-line packet and move onto the next one. The probability of success will not improve by attempting to obtain BW for the next packet instead of the current one. The summation in the denominator cannot be simplified further because $p = 0.5$ does not give us a finite solution.

If n_s, n_c and F represent the number of successful contenders, number of total contenders and number of BW-REQ slots, the number of successful contenders is given as (6.9), see (Sung-Min and Jae-Hyun 2005) and (Johnson and Kotz 1977).

$$
n_s = n_c\left(1-\frac{1}{F}\right)^{n_c-1} \tag{6.9}
$$

We can find the value of n_c such that n_s is maximized for a given value of F.

$$n_c = -1 \Big/ \ln\left(1 - \frac{1}{F}\right) \tag{6.10}$$

Numerically this is approximately equal to F. So when the number of competing stations is equal to the number of contention slots, the number of successful contenders is at a maximum.

$$n_c \approx F \tag{6.11}$$

This is the optimal value of n_c in terms of resource usage or contention successes per frame. The maximum number of successful stations can be approximated as shown below based on the linear curve fitting. C_f is the constant of proportionality.

$$\begin{aligned} n_{s,\max} &\approx F\left(1 - \frac{1}{F}\right)^{F-1} \\ n_{s,\max} &\approx F.C_f \\ C_f &\approx 0.37 \end{aligned} \tag{6.12}$$

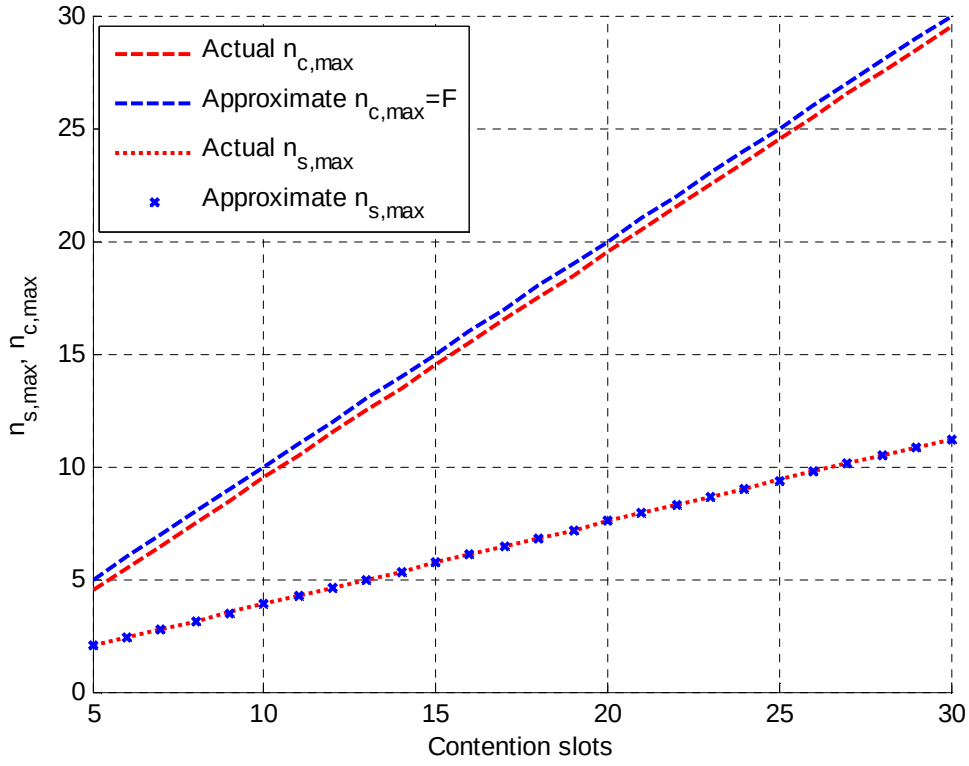

Figure 6-4 Analytical values for the number of maximum contenders and successes compared with the approximation $n_{c,max} = F$.

The approximation given in (6.7) can be justified by the plots given in Figure 6-4. While the number of optimum contenders shows a constant offset of ½ the number of maximum successes is equal for intents to the analytical value.

As stated before, if τ is the probability of a station transmitting a BW-REQ in any given frame, the number of contenders in any given frame is τ times the number of active stations.

$$n_c = n_a \tau \tag{6.13}$$

Here n_a is the number of active stations.

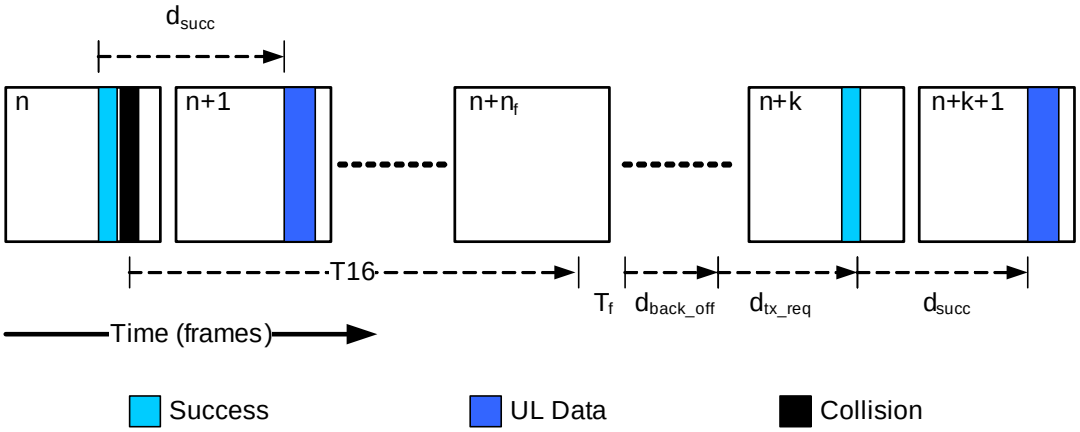

Figure 6-5 .Components of total delay in the contention delay.

6.2.4 Delay Analysis

The contention delay experienced by a SS consists of four components. These are

1) The d_{ackoff} period, d_{back_off}
2) The idle/wait period, T16, d_{idle}
3) The time between the T16 timer expiration and beginning of backoff period, d_w
4) The time between transmitting a successful BW-REQ and being granted BW, d_{succ}

The mean contention delay d_c in seconds is given as:

$$d_c = E\left[d_{back\_off} + d_{idle} + d_w + d_{succ}\right]$$

$$d_c = E\left[d_{back\_off}\right] + E\left[d_{idle}\right] + E\left[d_w\right] + E\left[d_{succ}\right] \quad . \tag{6.14}$$

$$d_c = \frac{\bar{p}W_0T_f}{2} + \sum_{n=2}^{\infty} \bar{p}p^{n-1}\left\{\frac{T_f}{2}\left[2W_0\left(2^{\min(n-1,\,n_{\max})} - 1\right) + n - 1\right] + (n-1)T_{16}\right\}$$

Assuming an infinite number of retries and $p \neq 0.5$ we can obtain a bounded finite value for the average contention delay as

$$d_c \xrightarrow{n\to\infty} \frac{\bar{p}W_0}{2}T_f + \frac{\bar{p}T_fW_0}{1-2p} + \frac{T_f + 2T_{16}}{2(1-p)} - \left(\frac{T_f}{2} + T_fW_0 + T_{16}\right) \tag{6.15}$$

where, $\bar{p} = 1 - p$ and T_f. is the frame duration. If $n_{f,16}$ is T16 expressed in frames, the delay in number of frames is,

$$d_{c,f} \xrightarrow{n\to\infty} \frac{\bar{p}W_0}{2} + \frac{\bar{p}W_0}{1-2p} + \frac{1 + 2n_{f,16}}{2(1-p)} - \left(\frac{1}{2} + W_0 + n_{f,16}\right) \quad . \tag{6.16}$$

The minimum number of frames between requests n'_f is,

$$n'_f = d_{c,f} + k_s \quad . \tag{6.17}$$

Here k_s is the fixed delay in frames required to process and service the request. This gives the maximum number of active stations that can be supported as,

$$n'_{SS} = n_{s,\max} \times (d_{c,f} + k_s) \quad . \tag{6.18}$$

The average packet rate per station (packets per second) on the UL is,

$$R = \frac{1}{T_f\,(d_{c,f} + k_s)} \quad . \tag{6.19}$$

6.2.5 Validation of Markov Chain Model

The Markov model described in the previous section is validated by means of the WiMAX simulator for QualNet 3.9.5. Unless otherwise stated the parameters used are given in the following table.

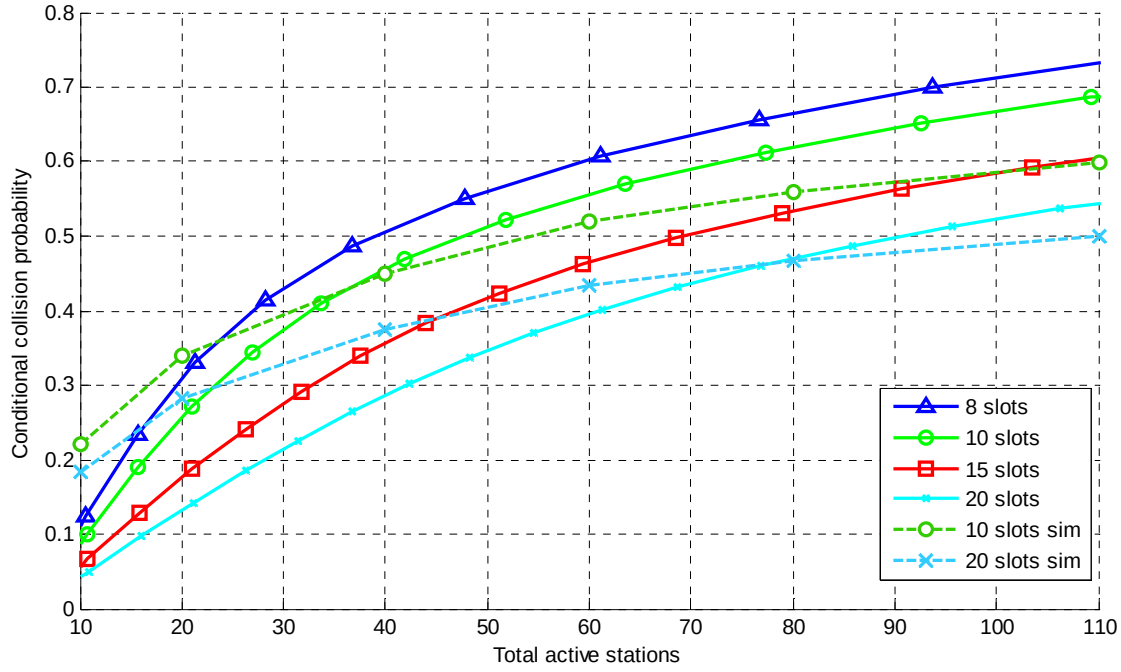

Figure 6-6 Conditional collision probability (p) of a transmitted BW-REQ is show for different numbers of contention slots in a frame.

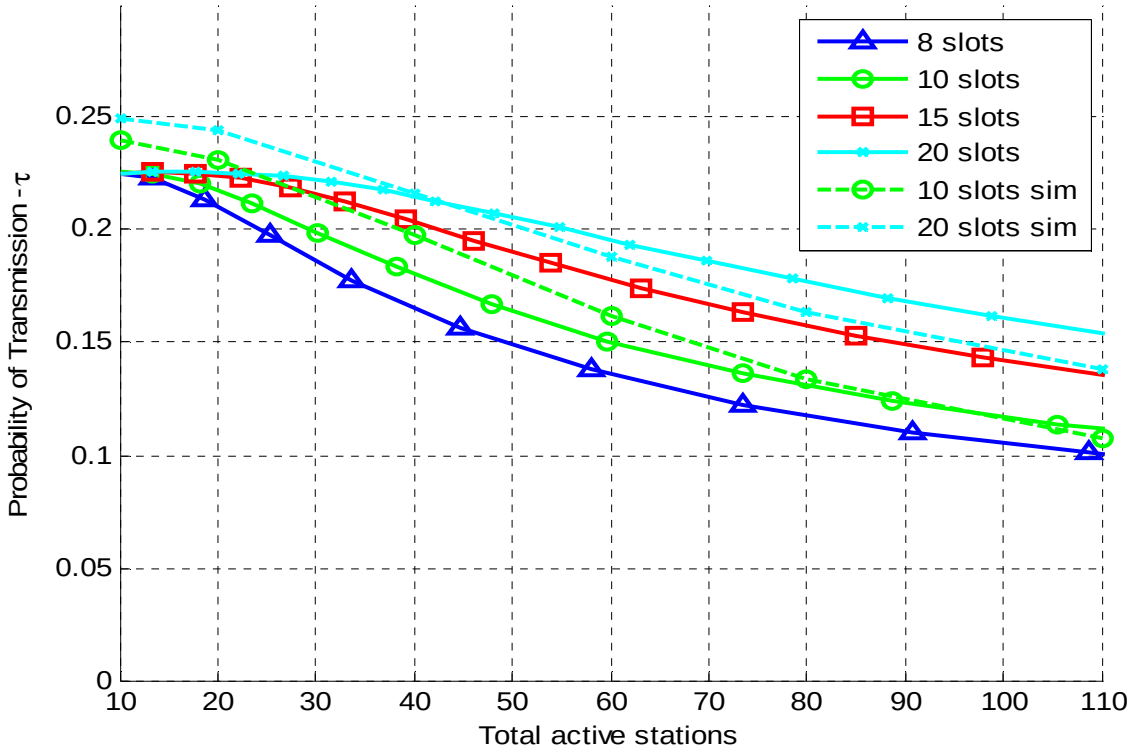

Figure 6-7 Probability of transmitting a BW-REQ in a frame is shown for different numbers of contention slots in a frame.

Table 6-1 Parameters used in the simulation and the analysis.

W_0	2
T_f	4 ms
T16	4*T_f = 16 ms
m	50

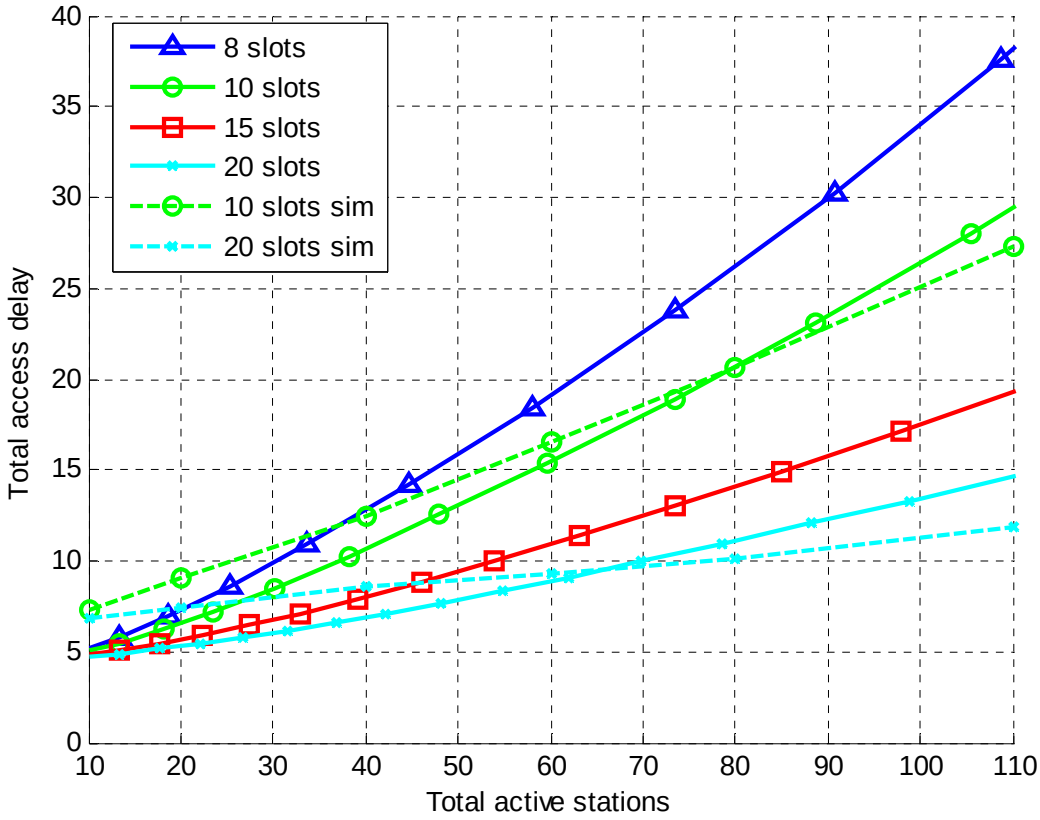

Figure 6-8 Total access delay in frames for different numbers of contention slots in a frame.

We attempt to verify that the model is able to correctly calculate the conditional probability of collision, *p*, which is a fundamental parameter of the system. Figure 6-6 gives analytical results for four different values of contention slots per frame, and the simulated results for two values (10 and 20 slots). No restrictions have been placed on UL or DL bandwidth available. The network loads range from approximately 15% to 85% utilisation. The SS are assumed to have a non-empty transmit buffer at all times. This is achieved by configuring a CBR traffic source at

each SS with a very high packet rate. In the simulation we use an interpacket interval of 5 ms. The DL is not congested and the only DL load is due to routing, ranging and dynamic service MAC messages. The probability that a SS will transmit a BW-REQ in a given frame, τ, is obtained from the same simulation and plotted in Figure 6-7. This plot also shows a close likeness between the simulation and the analytical results.

The access delay obtained from the analytical model (6.14), is compared with the access delay from the simulation model. The parameters are unchanged from the previous case. The results match up extremely well. It is also clear that a near linear relationship exist between the access delay and the number of active stations for a fixed set of parameters and contention slots.

6.3 Throughput of TCP Based Flows

In the preceding sections we have modelled the contention resolution process mathematically. Here we use the model to predict the DL throughput of a TCP based traffic flow. TCP being a reliable transport protocol it is clear that the UL access rate and delay will have a direct effect on its DL throughput.

6.3.1 Approximate Steady State Throughput

In steady state we may assume that the data packet rate on the DL is approximately $(2 \cdot R)$, twice the packet rate on the UL (R), which gives a maximum MAC layer throughput per station T_{max}, (6.20). Here PDU_{max} is the maximum PDU size that can be transported across the service. This assumption is based on TCP based traffic flows using a delayed acknowledgement scheme, which is the commonly used configuration (Stevens 1994). This scheme is recommended by RFC 1122 and RFC 2581 as the acknowledgement scheme which should be used.

$$T_{\max} = 2 \cdot R \cdot PDU_{\max} \tag{6.20}$$

The maximum values of throughput, number of frames between requests, successful stations per frame etc are mean values. We would expect the packet rate to fluctuate about the mean. So in order for this model to be valid the system would need to provide a dynamic number of slots for DL packets. Alternatively a fixed number of data slots on the DL with a buffer at the MAC layer will have the same effect. In

practice the amount of bandwidth allocated to BE flows is dependant on bandwidth used up by higher priority flows.

Consider x downlink traffic flows. k is the amount of resources (weighted number of OFDM symbols) for a data packet. $F_{x,p}$ is the number of contention opportunities required to give x successes with a failure rate of p. The total number of resources per frame used for x DL flows, is the sum of the DL, the UL (feedback) and the contention slots, (6.21).

$$S_{x,dl} = kx + \frac{x}{2} + F_{x/2,p} \tag{6.21}$$

Total used resources needs to be less than the maximum available, S_{max} which gives

$$S_{\max} \geq kx + \frac{x}{2} + F_{x/2,p} \tag{6.22}$$

From (6.9) we get

$$F_{x/2,p} = \left[1 - \exp\left(\frac{\ln(\bar{p})}{x/2\bar{p} - 1}\right)\right]^{-1}, \tag{6.23}$$

which can be closely approximated by a linear relationship for a given p.

Using the model defined by equations (6.8), (6.16), (6.18) and (6.23) we are able to predict the required number of contention slots for a given DL packet rate and estimate the maximum number of active SSs that can be supported. The number of active stations is

$$n_{a,\max} = \frac{x/2}{\bar{p}\tau} \tag{6.24}$$

Equations (6.23) and (6.24) are both functions of p, x and solving them numerically we find an approximate value for x which is the DL packet rate (and approximately twice the required UL success rate).

We can use the above analysis to dimension the system. Given a certain amount of resources we need to partition it for the DL/UL such that the DL packet rate (hence the throughput) is maximized. Conversely given a required DL packet rate, x, we need to allocate F such that the achieved value of x is as close as possible to the required value while keeping resource usage for contention overhead at a minimum.

6.3.2 Model Validation – Simulation Scenario

The IEEE 802.16d MAC layer has been coded for the simulation package QualNet 3.9.5 as described previously. We use a 256 subcarrier OFDM physical layer using a TDD frame structure. The frame duration is 4 ms and OFDM symbol duration is 12.5μs. The base station is connected via a wired link to an Ethernet subnet on which reside FTP servers which act as data sources. For simplicity we have placed all the SSs close to the BS so that they all use the same modulation scheme. If required to be more realistic the SSs could be placed spread out in the cell and the factor k in (6.21) and (6.22) can be a weighted average based on SS count using each modulation scheme.

The simulation is run for a duration of 10 minutes, with a varying number of active SSs from 20 to 60. The total available frame duration for BE traffic has been limited to 1 ms of the full frame duration of 4 ms in order to introduce throughput saturation at the BS. The network links which provide backhaul to/from the cell are not under congestion. This facilitates our assumption of the relationship between downstream and upstream packet rates.

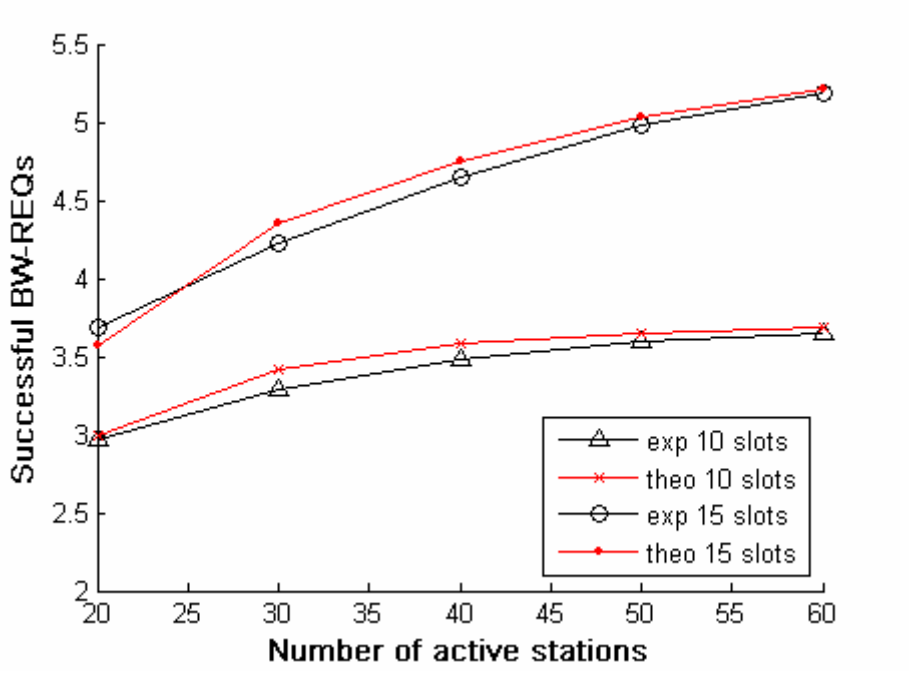

Figure 6-9 Average number of BW-REQ received by the BS with 10, 15 contention slots. (exp – experimental, theo – theoretical)

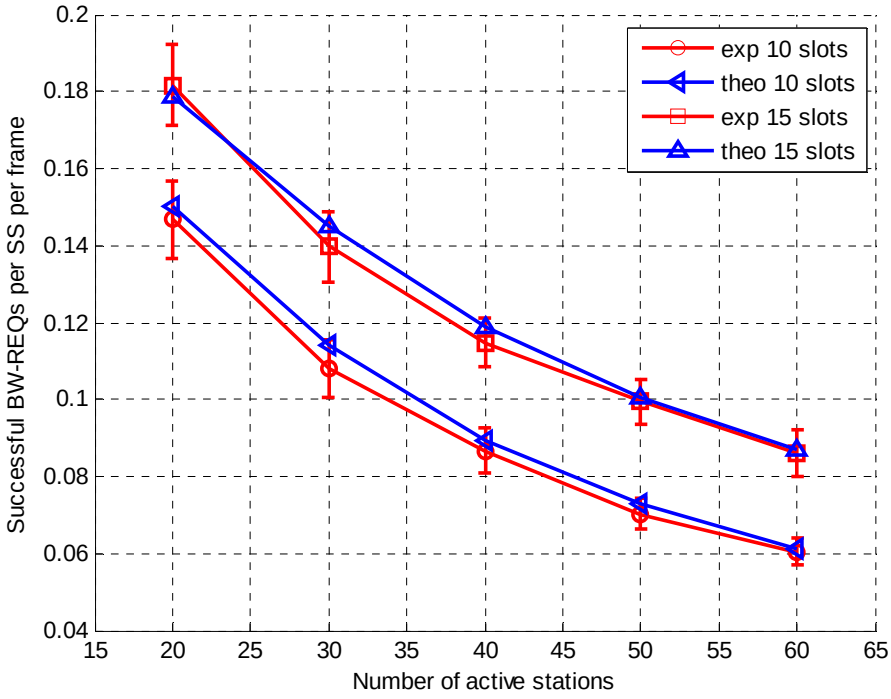

Figure 6-10 Average number of BW-REQ received by the BS per SS with 10, 15 contention slots. (exp – experimental, theo – theoretical). The error bars are for 1 standard error. The number of active stations is the number of samples.

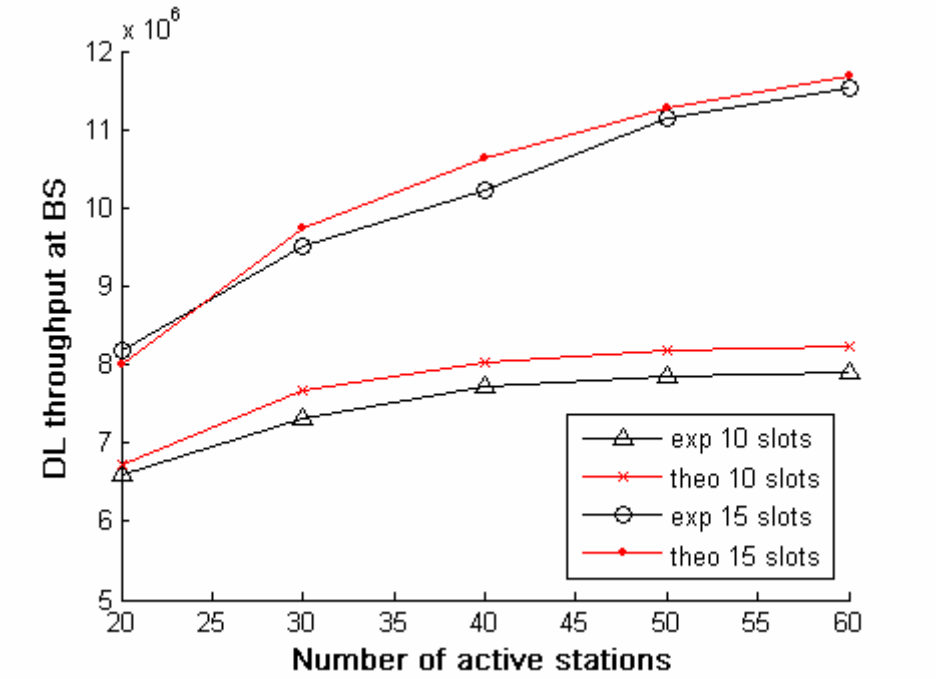

Figure 6-11 Average MAC layer throughput on the DL at the BS with 10, 15 contention slots. (exp – experimental, theo – theoretical)

Figure 6-9 shows the average number of BW-REQs received by the BS in a frame. The number of BW-REQs received from individual SSs is given in Figure 6-10. This is the mean value over the number of active stations. The theoretical results match very closely with the simulation results. Figure 6-11 shows the MAC layer DL throughput at the BS. The predicted values are always slightly more than the experimental values because the model does not take into account the time taken for routing related transmissions, and periodic ranging carried out by the SSs (Ranging is initially done when a SS joins the cell and then periodically to maintain link quality). Occasionally when packets are lost due to buffer overflows along the transport network, TCP would go into recovery phase and then slow start. During these times the delayed acknowledgment scheme is not used. The simple TCP approximation does not take this into account. This also contributes to the theoretical throughput being higher than the experimental throughput. Never the less, the results produced by the analytical model can be deemed to be accurate.

6.4 Throughput of UDP Based Flows

UDP is an unreliable transport protocol which does not use any feedback from the receiver. The UL and DL flows are independent of one another unlike in TCP based flows. This makes the analysis less complex. The UL packet rate will be equal to the contention success rate. UDP is mainly used for the transport of real-time traffic. We have in previous chapters, discussed in depth about using WiMAX for real-time services.

6.5 Adaptive allocation of Contention Bandwidth

An attempt has been made to regulate TCP based FTP traffic by dynamically adjusting the number of contention opportunities in a frame. The BS implements a controller which can feedback either the throughput, number of successes or number of DL packets per frame or any such parameter into the process. Our simple moving average filter on the number of received BW-REQs shows it is able to maintain a high utilization of downlink resources while using the minimum number of contention slots. The goal of this exercise is to adjust the number of BW-REQ slots in order to

control DL throughput and to use the minimum number of BW-REQ slots for a required throughput, i.e., not waste BW through allocating too many BW-REQ slots.

6.5.1 Simulation Scenario

Consider the case where BW for BE traffic on the DL is limited. If the BS has prior knowledge of the number of competing stations in every frame, the number of contention slots can be selected in order to maintain an optimal number of successes, which will in turn dictate the number of UL packets. As stated above we assume TCP based FTP file transfers. The number of successes, in turn limits the number of acknowledgements sent back to the server nodes residing on the external network, Figure 6-12.

It is advantageous to control the traffic flow to the BS in this way. Else, packets will be dropped due to buffer overflow or proactive queue management techniques such as random early detection (RED). The overhead increases when a packet loss is detected at the SS due to DUP-ACKs being sent for every subsequently received packet, and not every other packet as usual when using delayed ACKs.

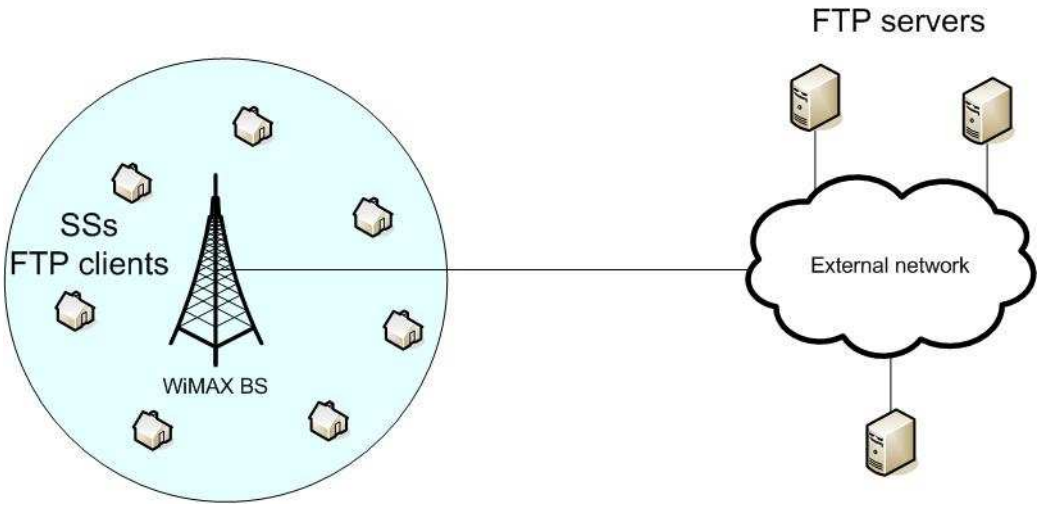

Figure 6-12 Simulation setup showing FTP servers and clients.

When F = 15 the average BW-REQs per frame received at the BS is more than 4, for more than 20 active SSs. i.e., the rate of ACKs is more than half the rate of the data packets (maximum data packets is 8 packets per frame in this scenario). This causes the servers to send data too fast and queues to build up and packets to be

dropped. We use a simple moving average filter on the number of BW-REQs received in previous frames and attempt to maintain the UL packet rate as close as possible to half the DL packet rate at the BS. Other methods such as dynamically modifying the T16 value, *Request Backoff Start* and *Request Backoff End* parameters depending on cell load can be considered viable. These values will be notified to the SSs through the UCD broadcast message. Enough time for approximately 8 full sized packets is given to the DL in each frame. This assumption is based on 25% of frame time being allocated to BE traffic.

At 50 s into the simulation 20 SSs begin FTP sessions with their external servers. At 100 s a further 10 SSs start FTP sessions. At 150 s the last 10 SSs join in. We allow a maximum of 15 contention slots. Based on the *F* values obtained from the model, the DL packet rate is maintained as close as possible to 8, while using a minimum number of contention slots.

6.5.2 Simulation Results

In our simulator we have implemented a 100 point (400 ms window) moving average filter at the BS for received BW-REQs, i.e., for the last 100 frames. A simple controller is used to regulate the number of contention slots in a frame in order to maintain an UL packet rate of approximately 4 packets per frame. This would imply 4 TCP acknowledgements per frame. If we assume all SSs will reach steady state, then the average DL packet rate will be approximately 8 packets per frame.

From equations (6.20) to (6.24) it can be calculated that for an average of 4 uplink packets with n_a = 20, 30, 40 (number of active stations), we need F = 22.9, 12.8, 11.6 (contention slots).The simulator maintains the number of contention slots based on the moving average of the number of BW-REQs received. The mean value of the number of contention slots in Figure 6-13, agrees well with the calculated values.

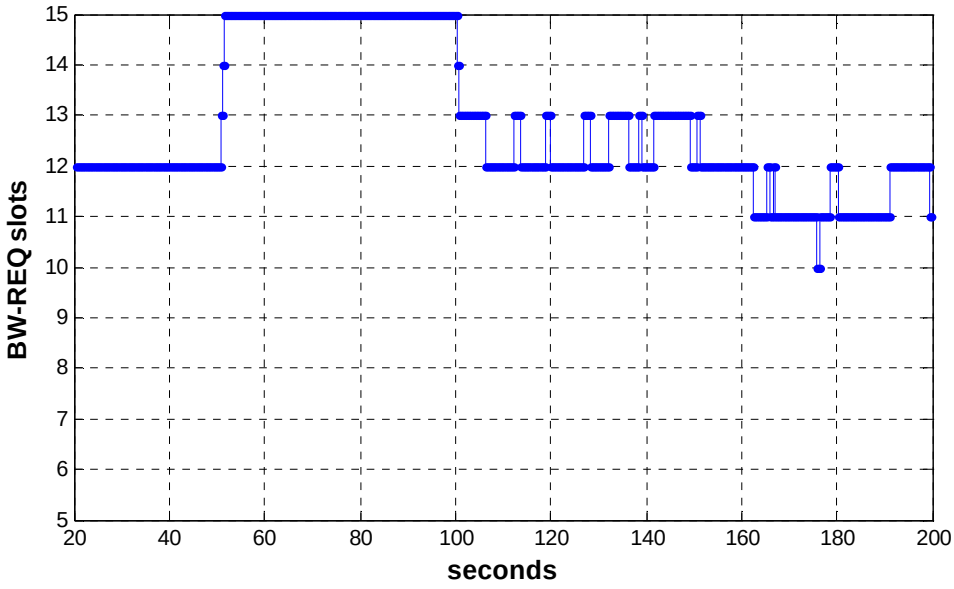

Figure 6-13 Shows the number of contention slots given by the BS. At 50 s the first of 20 FTP servers begin downloading data. At 100 s the next 10 begin, and at 150 s the last 10 begin transferring.

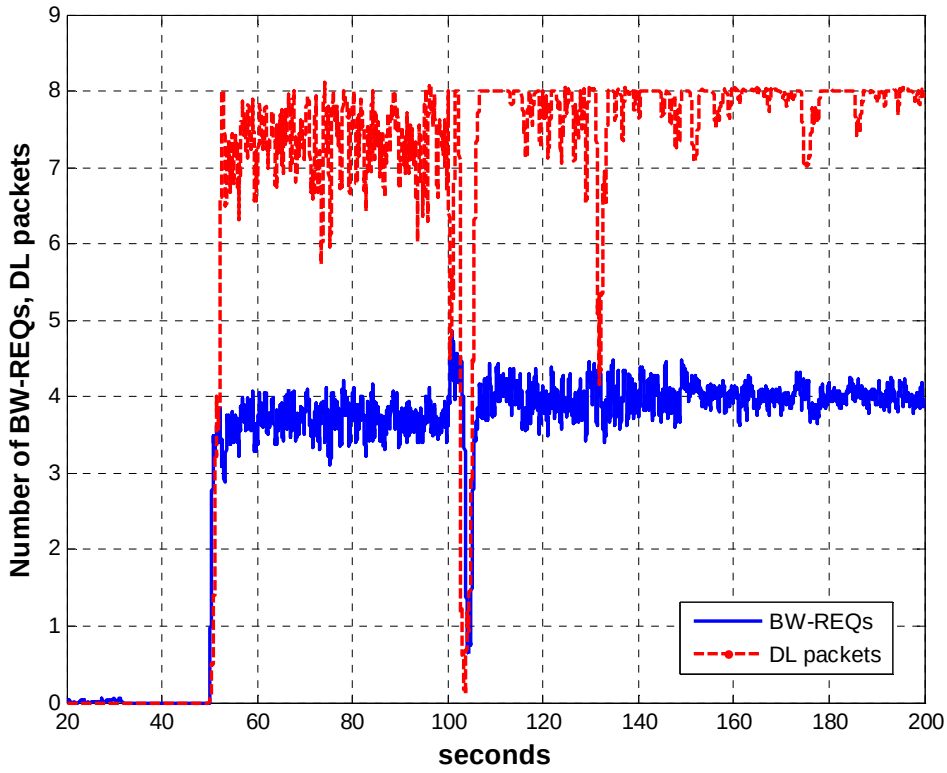

Figure 6-14 Shows the number of DL packets and the moving average of received BW-REQ at the BS.

The deep drops in the DL packet rate are due to periodic ranging and routing overhead, Figure 6-14. Other than those drops, packet throughput has been maintained at very close to maximum of 8 packets per second. The packet rate has a direct correlation to throughput. A utilisation in the range 0%-100% is directly comparable to a packet rate of 0 – 8 packets per second. 8 packets per second implies a maximum utilisation, i.e. maximum throughput.

6.6 Conclusion

Contention based access is a very important part of many MAC protocols to service low priority traffic with a dynamic load. This provides a means of sharing limited bandwidth among a large number of stations in a stochastically fair method. We have shown that some similarities are present in the IEEE 802.11 distributed coordination function and the DOCSIS standard random access method with the contention based access in IEEE 802.16d. In this chapter we have presented an analytical model for the mandatory contention mechanism of IEEE 802.16d. An analytical model has been developed using discrete time Markov analysis. The two-dimensional Markov chain includes all states a subscriber station goes through in its request/backoff procedure. Any state in the chain represents a stage in the backoff process. We have made allowance for ‘idle’ states when a subscriber waits for the base station to grant it bandwidth. This analysis differs from previous work in that there is no explicit acknowledgement to the subscriber of failure. This of course is a characteristic of the standard.

From the Markov chain model we have derived expressions for access delay and access rate. From theory of occupancy we have derived analytical expression for the number contenders, the number of successes for a given number of contention slots and active stations. Optimal values of these parameters have been derived based on increasing contention success for a fixed number of contention slots. Our simulation scenario has shown that the analysis is accurate in terms of collision rates and access delay. The analytical expression need to be numerically solved to produce values which are then compared with corresponding output from the QualNet3.9.5 WiMAX simulator.

In addition to a general scenario we have investigated the ability to predict throughput of TCP based downlink traffic flows using the model. Our TCP model is only accurate enough to model steady state TCP flow using the delayed ACK scheme. The comparison of analytical and simulation results shows much promise. An attempt has been made to regulate TCP based FTP traffic by dynamically adjusting the number of contention opportunities in a frame. The BS implements a controller which can feedback the throughput, number of successes or number of DL packets per frame or any such parameter to into the process. Our moving average filter on the number of received BW-REQs shows it is able to maintain a high utilization of downlink resources while using the minimum number of contention slots. We suggest as future work a controller that takes into account the round trip time (RTT) of the TCP flows in order to better regulate the number of slots.

Chapter 7

Polling Based Access for Best Effort Traffic

In telecommunication systems, traffic such as file transfer, e-mail, peer-to-peer and web browsing (most traffic transported using TCP) would be serviced as low priority best effort traffic. On the downlink the base station has the responsibility of scheduling the packets so that they reach the subscriber stations before the expiration deadline. This of course depends on the type of service flow the packets belong to and whether the base station can differentiate these flows based on information available in the packet headers or pattern matching of the payload itself. Best effort traffic being of the lowest priority means that only the remaining bandwidth after servicing all higher priority flows will be allocated to it. On the uplink bandwidth requests need to be sent to the base station using contention. This adds overhead in the form of contention slots and variable latency due to collision and backoff. We have proposed using an enhanced version of the nrtPS service class in Fixed WiMAX, as an alternative to contention based access for low priority traffic. The wireless resource utilization is improved and overheads are greatly reduced in bulk data transfer as well as bursty traffic scenarios.

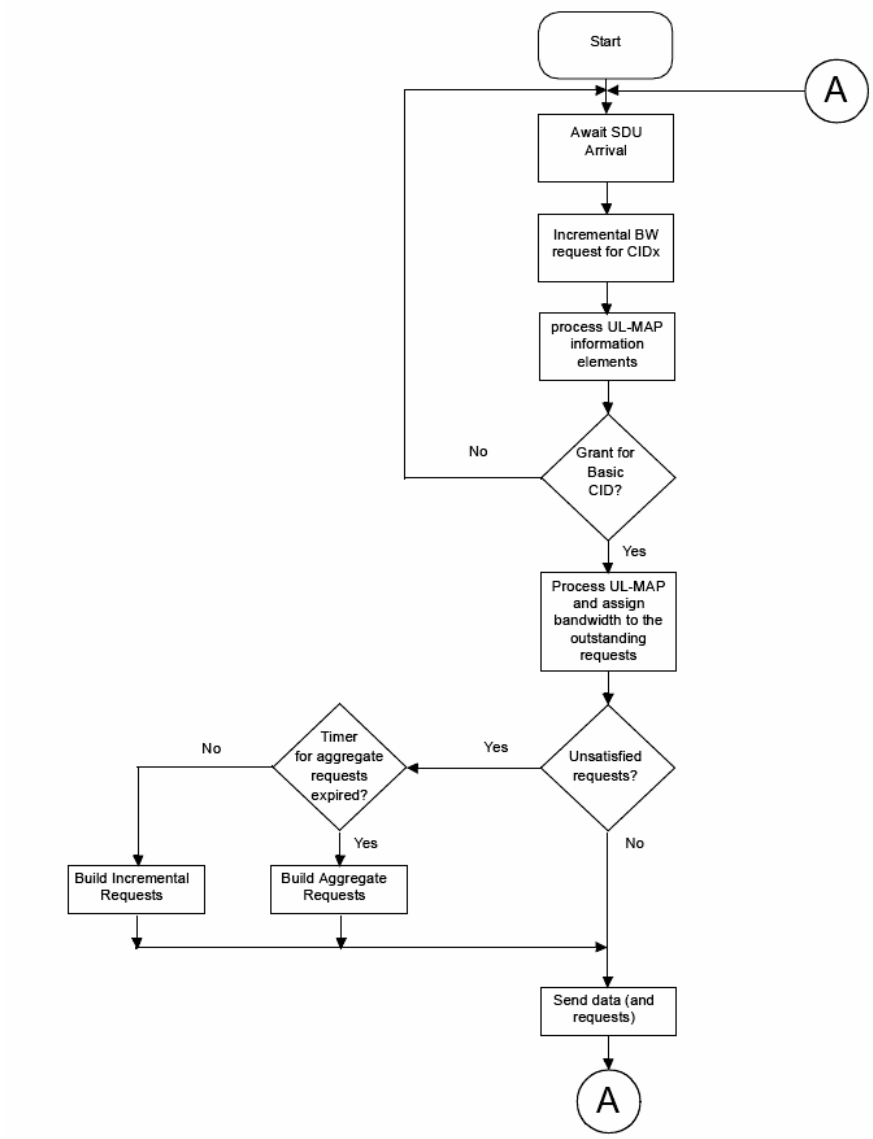

Figure 7-1 The process of allocating granted BW, to needy connections by the SS.

7.1 Overview of Polling Mechanisms

Polling is the process by which the BS allocates bandwidth to the SSs, specifically for the purpose of making bandwidth requests. These allocations may be to individual SSs or to groups of SSs. Allocations to groups of connections and/or SSs actually define bandwidth request contention IEs. The allocations are not in the form of an explicit message, but are contained as a series of IEs within the UL-MAP. Polling is done by the BS on a SS basis. Bandwidth is always requested by the SS on a CID basis, and bandwidth is allocated on a SS basis. The SS then apportions the received BW among the connections, Figure 7-1.

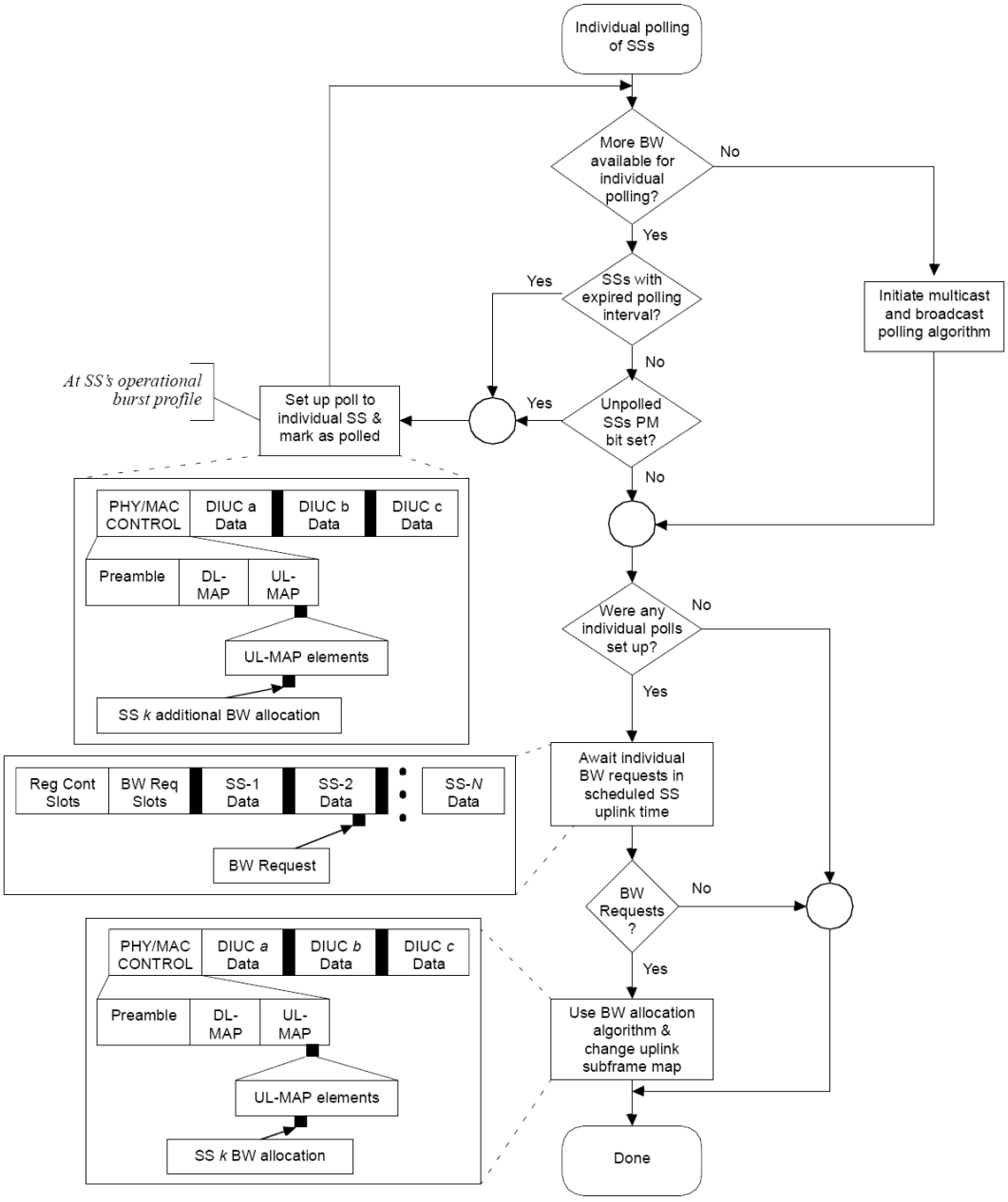

Figure 7-2 The process of unicast polling of a SS and the information exchange between the BS and the SS.

7.1.1 Unicast Polling

When a SS is polled individually, Figure 7-2, no explicit message is transmitted to poll the SS. Instead, the SS is allocated, in the UL-MAP, bandwidth sufficient to respond with a BE-REQ. If the SS does not need bandwidth, the allocation is padded by transmitting a known null sequence of bits or by transmitting a PDU with a MAC

header and a payload of zeros. SSs that have an active UGS connection of sufficient bandwidth, shall not be polled individually unless they set the ‘Poll Me’ bit (7.1.3) in the header of a packet on the UGS connection. This saves bandwidth over polling all SSs individually. Unicast polling is normally done on a per-SS basis by allocating a Data Grant IE directed at its Basic CID.

7.1.2 Multicast and broadcast

If insufficient bandwidth is available to individually poll many inactive SSs, some SSs may be polled in multicast groups or a broadcast poll may be issued. Certain CIDs are reserved for multicast groups and for broadcast messages. As with individual polling, the poll is not an explicit message, but bandwidth allocated in the UL-MAP with the appropriate UIUC and CID. The difference with unicast polling is that, rather than associating allocated bandwidth with an SS’s CID, the allocation is to a multicast or broadcast CID.

The information exchange sequence for multicast and broadcast polling is shown in Figure 7-3. When the poll is directed at a multicast or broadcast CID, an SS belonging to the polled group may request bandwidth during any request interval allocated to that CID in the UL-MAP by a Request IE. In order to reduce the likelihood of collision with multicast and broadcast polling, only SSs needing bandwidth reply. They shall apply the contention resolution algorithm as described in previous sections to select the slot in which to transmit the initial bandwidth request. Zero-length BW-REQs shall not be used in multicast or broadcast Request Intervals.

The SS shall assume that the transmission has been unsuccessful if no grant has been received in the number of subsequent UL-MAP messages specified by the parameter Contention-based Reservation Timeout. With a frame-based PHY with UL-MAPs occurring at predetermined instants, erroneous UL-MAPs may be counted towards this number. If the rerequest is made in a multicast or broadcast opportunity, the SS continues to run the contention resolution algorithm. The SS is however, not restricted to issuing the rerequest in a multicast or broadcast Request Interval. The effects and different applications of these two methods have been discussed in (Lidong, Weijia et al. 2007)

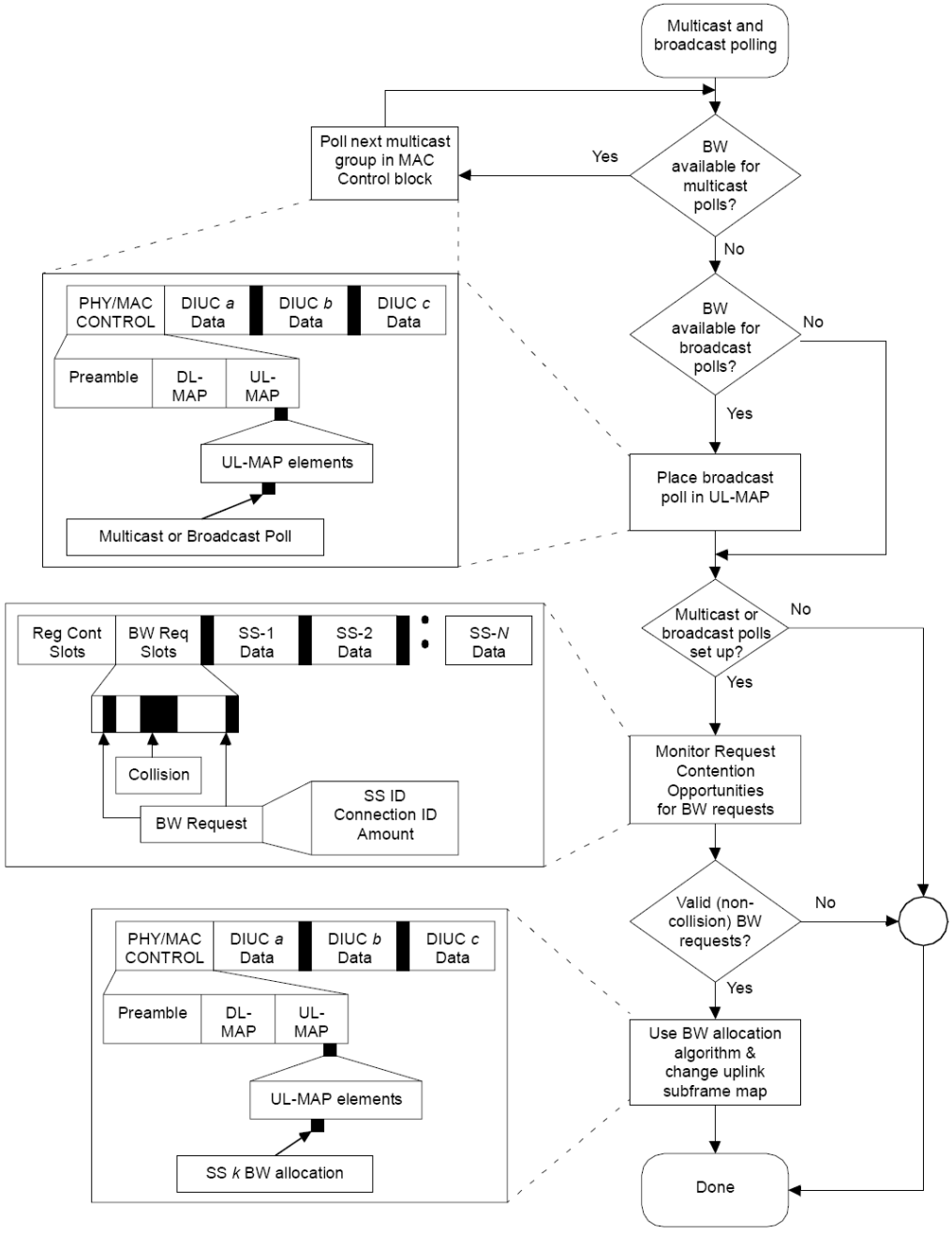

Figure 7-3 The process of multicast/broadcast polling a group of CIDs or SSs.

7.1.3 PM bit Usage

SSs with currently active UGS connections may set the PM bit (This it one of the bits in the Grant Management subheader and will be described in detail in a following section) in a MAC packet of the UGS connection to indicate to the BS that they need to be polled to request bandwidth for non-UGS connections. To reduce the bandwidth

requirements of individual polling, SSs with active UGS connections need be individually polled only if the PM bit is set (or if the interval of the UGS is too long to satisfy the QoS of the SS's other connections). Once the BS detects this request for polling, the process for individual polling is used to satisfy the request. The procedure by which an SS stimulates the BS to poll it is shown in Figure 7-4. To minimize the risk of the BS missing the PM bit, the SS may set the bit in all UGS MAC Grant Management subheaders in the uplink scheduling interval.

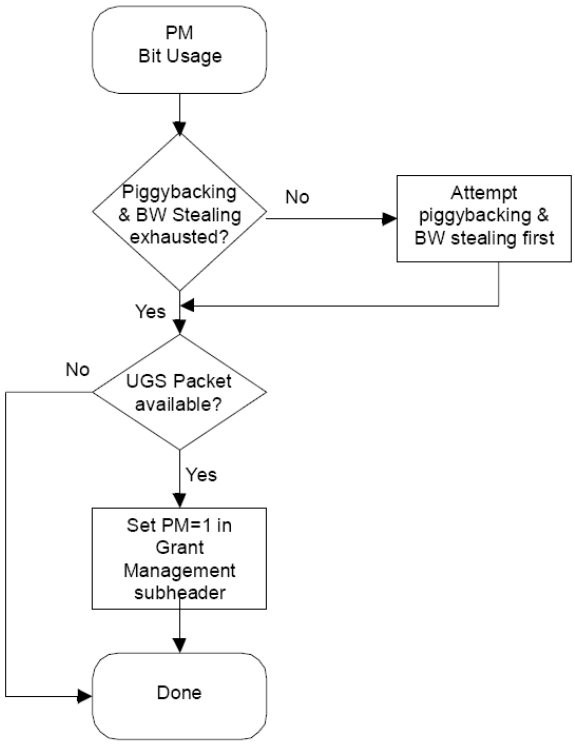

Figure 7-4 Poll Me bit usage by SS to inform the BS of polling requirements.

7.2 Operation of the Best Effort traffic class

As stated previously, the BE class is based on a contention based request-grant mechanism. When a packet is queued to be sent to the BS, the SS randomly picks a slot out of those given in the frame and transmits a BW request. No sensing of the medium is done like in Wi-Fi. If contention was successful, BW is granted by an entry in the UL-MAP (uplink map message, transmitted by the BS once every frame) which contains a connection identifier (CID) unique to that SS. If a BW grant is not received within a BW request timeout period (TOBWR) the SS defers contention as

per a truncated binary exponential back-off with a minimum contention window CW_{min}. The back-off window value gives the number of frames to back-off, regardless of how many request opportunities are available in each frame. Once the BW request is received by the BS it will be serviced as per delay requirements of the flow or agreed upon values of maximum latency, although for BE traffic this ultimately depends on the available BW.

7.2.1 Contention without Piggyback Requests

In order to analyse the performance of the BE BW-REQ service queue we need to define the arrival and departure processes of requests. The arrival process is due to a SS successfully transmitting a BW-REQ to the BS, either through contention, polling or piggybacking. The departure process is how the BS services the BW-REQs in the queue. Similar analysis using Markov Modulated Poisson Processes (MMPP) have been carried out in (Ganesh Babu, Le-Ngoc et al. 2001; Muscariello, Meillia et al. 2004). However this analysis does not involve modelling of the individual sources but an aggregate receive process at the BS.

Using F contention slots, n_c (out of a total of n_a) active stations contend to transmit BW-REQ to the BS. For a given n_c, the number of successes is $n_s \in [\,0, n_{s,\max}\,]$, where

$$n_{s,\max} = \begin{cases} n_c, & n_c \le F \\ F-1, & n_c > F \end{cases} \tag{7.1}$$

If p_s is the probability of a contention slot having only one BW-REQ transmit during it (probability of contention success)

$$p_s = \binom{F}{1}\left(\frac{1}{F}\right)\left(1-\frac{1}{F}\right)^{n_c-1} \tag{7.2}$$

The probability of having none or more than one BW-REQ in a contention slot (unused or collision) is $1-p_s$. Using the above we can say that,

$$PDF(n_s) = \binom{n_{s,\max}}{n_s} p_s^{n_s}\left(1-p_s\right)^{n_{s,\max}-n_s} \tag{7.3}$$

This is also the distribution of the arrival process of BW-REQs at the BS request queue. The mean of this process is given by $\lambda_a = n_{s,\max} \times p_s$. Let the arrival process $A(n) = a_n$ be the number of requests received by the BS at frame n.

$a \in [a_{\min}, a_{\max}]$, where a_{min} and a_{max} are the lowest and highest possible number of requests received. This distribution is given by (7.3).

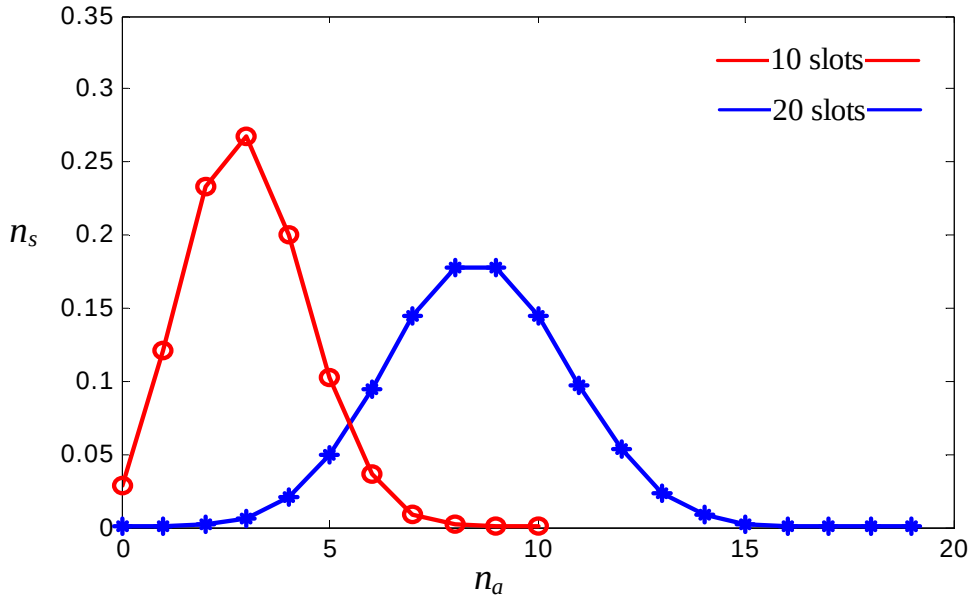

Figure 7-5 Probability distribution of n_s. Shows the two possible scenarios for $n_{s,max}$ given in (7.1), (1) when the maximum contenders is 10, and (2) when the maximum contenders is more than 20.

The departure process depends on the rate at which the BS can allocate BW (or frame time, we have used the term BW in this sense throughout this thesis) for the BW-REQs. The transmission time depends on two factors. These are (1) the packet size and (2) the UL burst profile of the user. If the departure process is given by $B(n) = b_n$ and b is the number of requests serviced by the BS in one frame, $b \in [b_{\min}, b_{\max}]$, where b_{min} and b_{max} are the lowest and highest possible number of requests served. The lowest comes about when all the requests are for packets of the maximum size at the lowest burst profile. Similarly the highest comes about when all the requests are for the smallest packet size at the highest burst profile. Burst profiles (of which the modulation scheme is the most important parameter) are related to signal quality of the link between the BS and a given SS. We can assume that SSs are uniformly distributed within the cell, and that the proportion of SSs using a given burst profile is equal to the proportion of the total cell area in which the given burst profile is used, a_i. The distribution of packet sizes, *ps*, can be obtained by observing

the traffic flow for a sufficient period of time which includes short term fluctuations. Let us assume that the service distribution, PDF(*b*), is approximately known and that its mean is given by μ_b. The expected packet size, *E*[*ps*] is found from the distribution of the packet sizes, where ps_i denotes the i^{th} packet size. s_{pp} denotes symbols per packet and a_i gives the proportion of the cell covered by the i^{th} burst profile. This leads to the UL and DL service rates, (7.4).

$$\begin{aligned} E[ps] &= \sum_i (ps_i \cdot p(ps_i)) \\ E[s_{pp}] &= \sum_i \left\lfloor \frac{E[ps] \cdot a_i}{n_s} \right\rfloor \\ E[N_{UL}] &= \frac{BW_{UL}}{E[s_{pp}]} = \mu_{b,UL} \\ E[N_{DL}] &= \frac{BW_{DL}}{E[s_{pp}]} = \mu_{b,DL} \end{aligned} \tag{7.4}$$

We define the queue by process X, where $X(t) = x_t$ gives the length of the queue at time *t*. With the frame structure of WiMAX it is more appropriate to consider discrete time process $X(n) = x_n$ where n is the nth frame. From the batch arrival and departure processes described above we can see that the BW-REQ queue at the BS can be modelled by a discrete time Markov chain. The maximum length of the queue is *Q*. Then,

$$X(n) = \min[X(n-1) + A(n), Q] - B(n) \ . \tag{7.5}$$

We need to use the *min* function because contention based BW-REQs are received at the beginning of the UL subframe, before any BW-REQs are serviced. Any overflow of the queue (dropping of requests) occurs at this time. The number of dropped requests, $D(n) = d_n$, is given by,

$$D(n) = \max[X(n-1) + A(n) - Q, 0] \ . \tag{7.6}$$

At the beginning of the frame the queue length is always less than or equal to $Q - b_{\min}$. However it could reach the maximum length during the contention part of the frame.

Matrix $\boldsymbol{P}$ is the QxQ state transition probability matrix where each element $p_{i,j}$ is the probability that the queue length changes from i to j by the end of the frame. $p_{i,j}$ can be defined in terms of probabilities from the PDFs of a_n and b_n.

$$p_{i,j} = \begin{cases} \Pr(b-a \le i-j) & , \quad j=0 \\ \Pr(b-a=i-j) & , \quad (j\neq 0)\ and\ (i \le Q-a) \\ \Pr(b-a=i-j) - \\ \sum_{k=Q-i+1}^{a_{\max}} \Pr(a>Q-i)\times\Pr(b=i-j+k) + \\ \Pr(a>Q-i)\times\Pr(b=Q-j) & , \quad (j\neq 0)\ and\ (i > Q-a) \end{cases} \tag{7.7}$$

The final probability terms such as $\Pr(b-a=i-j)$ can be calculated from the PDF of b-a. When j=0 (7.7) gives us the empty queue state probabilities while when ($j\neq0$) and ($i>Q$-a) it gives the probabilities of states where one or more BW-REQs have been dropped due to the queue being full.

The steady state probability vector of the queue, $\boldsymbol{\pi}$, can be obtained (Stewart 1994) by solving,

$$\pi = \pi \mathrm{P} \ \text{ and } \ \sum \pi_i = 1 . \tag{7.8}$$

Then the probability of dropping a BW-REQ is,

$$p_d = \sum_{i=0}^{Q} \left[\pi_i \times \Pr(a > Q-i)\right] , \tag{7.9}$$

with the expected number of drops per frame given by,

$$\mathrm{E}[d] = \sum_{i=0}^{Q} \left[\pi_i \times \Pr(a>Q-i)\times \left|a-(Q-i)\right|\right] . \tag{7.10}$$

The expected queue length $\mathrm{E}[X] = \sum i\pi_i$, and mean waiting time $\mathrm{E}[T_w] = \mathrm{E}[X]/\mu_a$ are easily obtained. The total access delay is the sum of (1) the time spent in backoff due to collisions or drops of BW-REQs, and (2) the time spent in the BW-REQ queue. Since there is a possibility of many consecutive collisions and/or drops, it is likely that a small proportion of SSs would experience long access delays. The access delay due to contention and the backoff procedure was detailed in Chapter 6, section 6.2. If a single SS is considered, during the access delay (time between two

consecutive BW grants) all other SSs should be serviced once on average. So the number of active SSs should be equal to,

$$\mathrm{E}[n_a] = \mathrm{E}[n_s] \times \mathrm{E}[T_{ad,f}] \,, \tag{7.11}$$

where $T_{ad,f}$ denotes the total service delay in frames. Since the left hand side of (7.11) is a function of n_s, when n_a is known we can numerically solve for n_s (as done in Chapter 6).

The UL packet rate per SS is $1/\mathrm{E}[T_{ad,f}]$, which can be used to estimate the throughput, using the expected value of packet sizes on the UL. In the case of a DL TCP based flow, the DL packet rate can be assumed to be roughly twice the UL rate. Using the expected value of DL packet size, we can then estimate the DL throughput.

7.2.2 Contention with Piggyback Requests

We consider the situation where the SS is allowed to piggyback bandwidth requests along with the transmitted payload on the UL. This is accomplished via the Grant Management Subheader which is inserted in to the PDU between the MAC header and the payload. The operation of this is described below.

Table 7-1 Grant management subheader format

Syntax	Size	Notes
Grant Management Subheader() {		
if (scheduling service type == UGS) {		
SI	1 bit	
PM	1 bit	
reserved	14 bits	Shall be set to zero
}		
else {		
PiggyBack Request	16 bits	
}		
}		

7.2.2.1 Grant Management Subheader

The Grant Management (GM) subheader is two bytes in length and is used by the SS to convey bandwidth management needs to the BS. This subheader is encoded differently, based upon the type of uplink scheduling service for the connection (as

given by the CID). The use of this subheader is defined in (IEEE 802.16 WG 2004). The Grant Management subheader format is shown in Table 7-1. The capability of Grant Management subheader at both BS and SS is optional. Here SI is the Slip Indicator bit which alerts the BS of the growth of the SDU queue at the SS. PM is the Poll Me bit which advices the BS to poll the respective SS. The combination of these two bits is used in the UGS service class. In this case, we use the GM subheader for nrtPS flows which utilize all 16 bits to convey the amount of BW required to the BS.

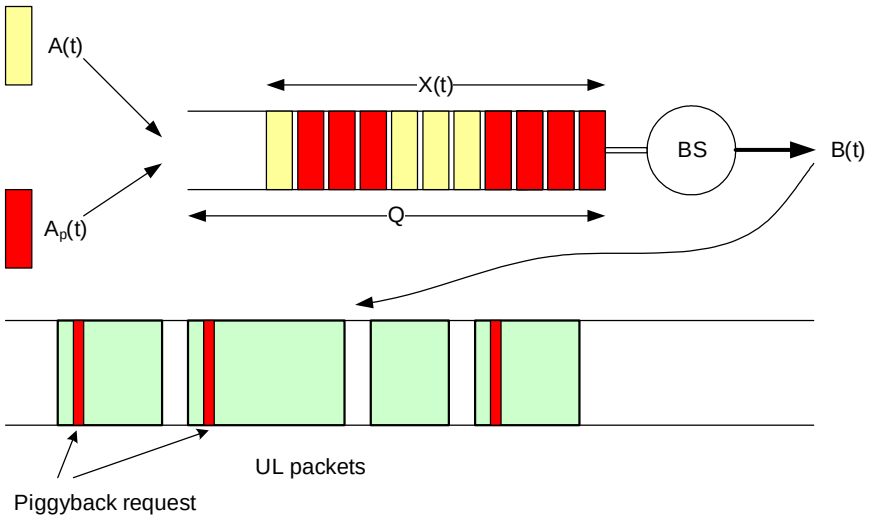

Figure 7-6 The BW-REQ queue process showing the arrival process of BW-REQs through contention and piggybacking.

7.2.2.2 Analysis of Queuing Model

Consider the case where BE can use piggyback requests. The queue process which is depicted in Figure 7-6, can now be given by,

$$X(n) = \min[X(n-1) + A(n), Q] - B(n) + A_p(n) \tag{7.12}$$

$A_p(n)$ is the additional component which represents the number of piggyback BW-REQs received in the current frame. The rest of the terms in the expression remain unchanged from the previous sections. The number of piggy back requests received in a frame can be at most equal to the number of UL packets received in that frame, $0 \le A_p(n) \le B(n)$. Under saturated conditions, it can be assumed that every SS will have a non-empty transmit queue. This implies that $A_p(n) = B(n)$ and that the queue growth depends on $A(n)$. None of the piggybacked requests will be dropped as

they will take up queue space created by servicing the head of line requests. When the queue overflows only the contention based REQs will be dropped. This introduces unfairness by allowing a SS with a BW-REQ already in the queue a better chance of getting another REQ into the queue. Since there is a possibility of many consecutive drops (and collisions) it is likely that a small proportion of SSs would experience long access delays or data rates which drastically fluctuate over time. We would however expect the aggregate cell throughput to be close to the peak value and have an approximately equal per SS throughput in the *long term*.

Under saturated conditions, the elements $p_{i,j}$ of the state transition probability matrix P are now given by,

$$p_{i,j} = \begin{cases} \Pr(a = j - i), & j < Q \\ \Pr(a \geq j - i), & j = Q \end{cases} \tag{7.13}$$

In a saturated system there will be no shrinking of the queue which means all elements $p_{i,j}$ where $j<i$ are zeros. As in the case without piggybacking we can solve (7.8) to find the steady state probability vector. It can be seen that for any non-zero values of a, the queue will keep growing until it is full. So the steady state Expected length of the queue is Q with a waiting time of $\mathrm{E}[T_w] = Q/\mu_b$ for successful entrants. If the maximum queue length, Q, is less than the number of active stations, those SS already having a BW-REQ in the queue have an advantage as they will be able to continuously transmit data through piggybacking. The unlucky SS to not have a BW-REQ in the queue when it reaches the full state will experience longer access delays which will be dependant on the flow of traffic to other SSs and the random backoff based contention resolution process.

7.3 The Non-Real-Time Polling Service (nrtPS)

It can be seen that most of the widely used access technologies for wired as well as wireless networks have more delay/throughput aware scheduling classes for higher priority flows. In terms of service class definition and basic operation of these classes, WiMAX has many similarities with Data-Over-Cable Service Interface Specifications (DOCSIS), including similar polling mechanisms. This is due to the WiMAX Mac layer being based on DOCSIS. PCF of 802.11 has been mentioned in previous

sections as well. In this section we provide an analysis of the operation of the nrtPS service class and investigate the use of nrtPS to service low priority BE traffic.

7.3.1 Operation of nrtPS

The nrtPS class is designed to support delay-tolerant data streams consisting of variable sized data packets for which a minimum data rate is required, such as FTP. The nrtPS offers unicast polls on a regular basis, which assures that the uplink service flow receives request opportunities even during network congestion. The BS typically polls nrtPS connections on an interval on the order of one second or a few hundred milliseconds. The mandatory QoS service flow parameters for this scheduling service are:

- Time Base
- Minimum Reserved Traffic Rate
- Maximum Sustained Traffic Rate
- Traffic Priority
- Request/Transmission Policy

The Request/Transmission policy specifies what kind of dynamic access is allowed when the poll based bandwidth is insufficient or the polling frequency is too low.

Using the same notation used for the BE queue we model the nrtPS request queue as below.

$$X(n) = X(n-1) + \lfloor \lambda_p,\ 0 \mid X(n-1) < n_a \rfloor - B(n) + A_p(n) \tag{7.14}$$

Where λ_p is the expected number of polls per frame and A_p is the piggyback receive process such that, $0 \le A_p(n) \le B(n)$. $[x, y|z]$ means that if z is true, the value of the expression is x, else the value is y.

The number of polls per second for all active stations is n_a/T_p, and $\lambda_p = n_a T_f / T_p$. Here T_p is the poll period in seconds and T_f is the frame duration. To sustain all stations at the minimum packet rate, which is at least equal to the rate of polling,

$$\begin{aligned} &\mu_b > n_a/T_p \text{ and} \\ &Q > n_a\ , \end{aligned} \tag{7.15}$$

must be satisfied.

The state transition probability matrix ***P*** is *Q*x*Q* as before, and its elements are defined as follows:

$$p_{i,j} = \begin{cases} \Pr[B(n) - A_p(n) > i + \lambda_p] , & j = 0, i \le n_a \\ \Pr[A_p(n) - B(n) = j - i - \lambda_p] , & j > 0, i < n_a \\ \Pr[A_p(n) - B(n) = j - i] , & j > 0, i = n_a \end{cases} . \quad (7.16)$$

From above when $i = n_a$

$$p_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} . \quad (7.17)$$

This implies that once the queue length reaches n_a it will remain at that value as long as the system is saturated.

The packet rate per second per SS is,

$$R = \frac{\mu_b}{n_a \cdot T_f} , \quad (7.18)$$

with an expected waiting time in the queue of $\mathrm{E}[T_w] = 1/R$, which is also equal to the mean access delay E[T_{ad}].

The amount of overheads used in nrtPS is much less than contention based access. The only overhead is for the Dynamic Service Addition (DSA) message sent at the start of the flow and the BW-REQ opportunities given to SSs which do not have data to send. Unlike in contention based access this class can provide a more consistent data rate with fewer fluctuations.

7.3.2 Comparison of Real-Time and Non-Real-Time Polling Service

The rtPS class is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as moving pictures experts group (MPEG) video, (Ganesh Babu, Le-Ngoc et al. 2001). The service offers real-time, periodic, unicast request opportunities, which meet the flow's real-time needs and allow the SS to specify the size of the desired grant. This service requires more request overhead than UGS, but supports variable grant sizes for optimum data transport efficiency.

In order for this service to work correctly, the Request/Transmission Policy setting shall be such that the SS is prohibited from using any contention request opportunities for that connection. The BS may issue unicast request opportunities as

prescribed by this service, even if prior requests are currently unfulfilled. This results in the SS using only unicast request opportunities in order to obtain uplink transmission opportunities (the SS could still use unsolicited Data Grant Burst Types for uplink transmission as well). All other bits of the Request/Transmission Policy are irrelevant to the fundamental operation of this scheduling service and should be set according to network policy. The key service IEs are the Maximum Sustained Traffic Rate, the Minimum Reserve Traffic Rate, the Maximum Latency and the Request/Transmission Policy.

Optimisation of the rtPS service class has been the topic of discussion in (Ben-Jye and Chien-Ming 2006; Zhang, Li et al. 2006) which aim to minimize access delay to provide a UGS-like service with variable size grant. A work on queue aware BW scheduling for a polling based broadband access system was done in (Niyato and Hossain 2005; Niyato and Hossain 2006). The main emphasis in this work was the modeling of traffic sources at the SS, and providing of timely feedback to the BS of queue status which would in turn help the BS to schedule more intelligently. The applicability of schemes, with such high OH, would in practice only be viable for high priority flows such as real-time video. Low latency reverse channel error feedback was assumed in the above work as well as by the authors (Hyogon, Sangki et al. 2005; Ben-Jye, Yan-Ling et al. 2007). The rtPS service class in contrast to nrtPS, is a higher value service with stricter QoS guarantees. It is clear that UGS and rtPS target the interactive voice and video space of the service spectrum.

7.3.3 Disadvantages of nrtPS

1) Once an nrtPS allocation is granted by the BS there is no way to end it when an idle period is detected. No procedure has been defined for notification of the end of a flow or end of an SS’s life time.
2) Some small packets can actually be transported in a single OFDM symbol. Since the allocation for the unicast poll is at minimum one symbol wide, it is more effective to use it for an SDU with a piggyback request tagged on. nrtPS is not flexible enough to do this.
3) In addition to the polls, bandwidth is required for contention because the request/transmission policy of nrtPS allows the SSs to contend for BW. How

long a SS waits for a unicast poll (if it waits at all) before it decides to use contention is also not specified in the standard. In this case how the polling scheme adapts to a received contention based BW-REQ is open to interpretation.

4) During periods of inactivity the polling service will still continue to provide polls at the preset frequency. This is a waste of BW and adds to the OH of the scheme.

7.4 Enhancements to nrtPS

In order to utilize nrtPS as a substitute for contention based access for dynamic BW requirements, the disadvantages given previously need to be addressed. We have enhanced the way the request/grant mechanism functions. The resulting polling scheme is referred to as 'Enhanced nrtPS' or e-nrtPS from this point forward, not to be confused with the service class ertPS defined in the 2005 standard, (IEEE 2005). The enhancements are detailed in the following sections.

7.4.1 Adaptive Poll Period

In order to minimize BW wastage during idle periods, the polling period is adaptively increased based on a truncated exponential algorithm. Let the default polling period be given by $T_{p,d}$. After a certain number of idle polls the period is increased by multiplying it by α ($\alpha > 1$). Once the period reaches a maximum value the increase is truncated (7.19).

$$T_p = \min(\alpha^n T_{p,d},\ T_{p,\max}) \tag{7.19}$$

If the SS has no data to send on the uplink (UL) it will send a dummy packet consisting of a generic MAC header or a BW-REQ MAC header with a zero bytes request. These messages keep the connection alive during relatively short idle periods, Figure 7-7. Consider the process of web browsing where information is transferred to and from the SS in bursts which are followed by periods of inactivity where the user reads downloaded information. The speed by which the poll period is retarded depends on the value of α.

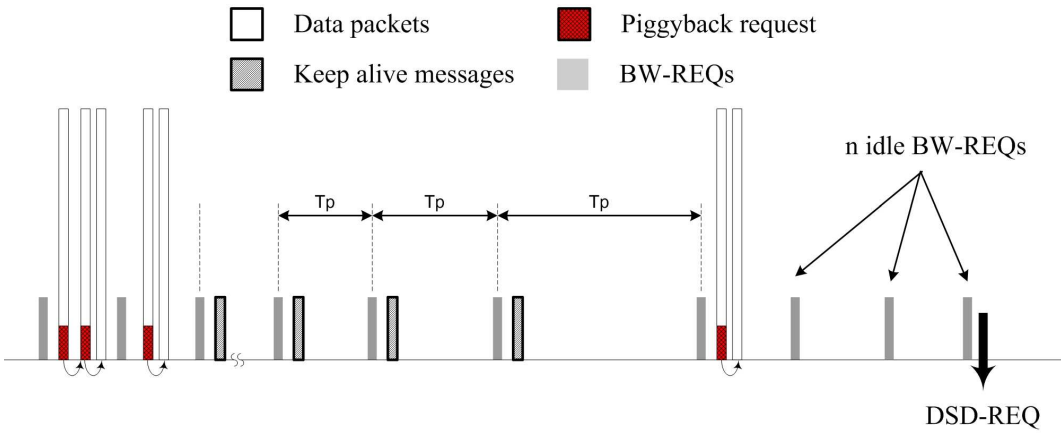

Figure 7-7 The proposed modifications to nrtPS to save BW during idle periods and to discontinue the connection during extended idle periods when the SS is offline. The grey bars denote BW-REQ opportunities, the hatched bars denote keep alive messages and the taller bars denote data packets.

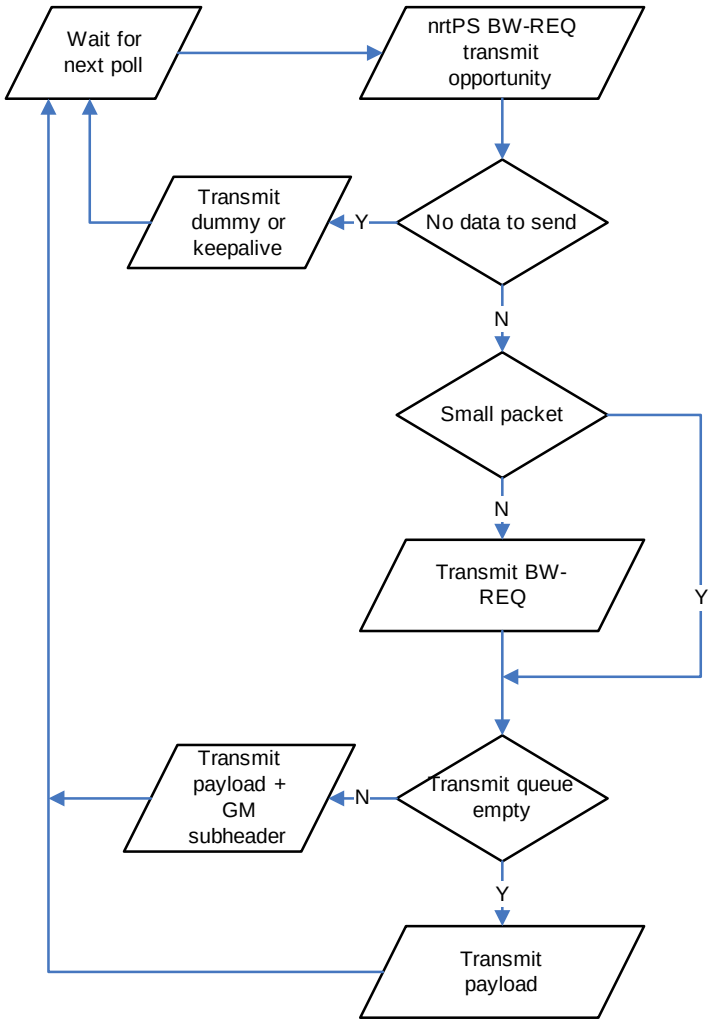

Figure 7-8 The enhanced nrtPS poll procedure from the SS's perspective.

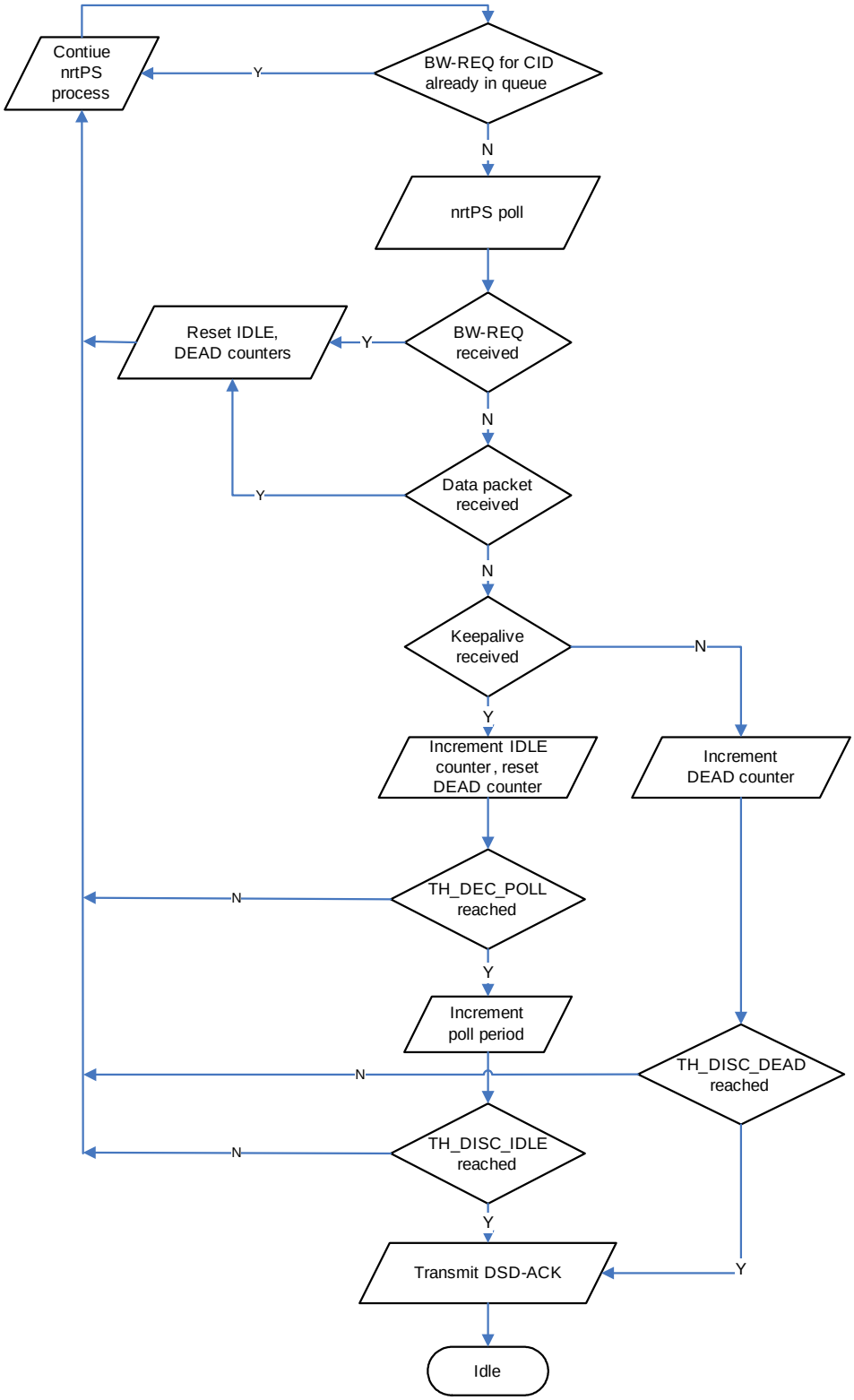

Figure 7-9 The e-nrtPS process flow diagram from the BS's perspective. TH_DEC_POLL, TH_DISC_IDLE and TH_DISC_DEAD are the threshold values used for bandwidth saving algorithm.

The BS which plays the master roll continuously monitors the BW usage by the SS. During an idle period in which the SS has no data to transmit on the UL the SS will transmit a dummy packet. The BS increments an idle counter for every consecutive unused unicast BW-REQ transmit opportunity. When this counter reaches a threshold (given as TH_DEC_POLL) it will begin to retard the poll period (7.19) of that SS or CID. Once the poll period reached a maximum allowed value the retarding shall be stopped and the poll period held at that value.

At any time during the process if a BW-REQ is received by the BS using the poll or using the PM bit of another connection the idle counter in initialized and the poll with the default period is resumed.

7.4.2 Active Management of Polled Connections

During extended idle periods, the BS will transmit a DSD-ACK. This signifies the end of the connection. As shown in Figure 7-9, the BS will take this course of action due to two reasons.

1) The BS may receive dummy packets for an extended period of time. Once the IDLE counter reaches a threshold the connection will transit in to a terminated state. We denote this threshold as TH_DISC_IDLE. This threshold should be set high enough so that there are no unnecessary and frequent connection tear downs. Conversely, the threshold should not be too large so that the system will not make any gain from the additional overhead.
2) The BS may not receive any transmission at all for that connection, which causes the DEAD time counter to be incremented. Once a threshold is reached the connection is terminated. This threshold is denoted by TH_DISC_DEAD. If there is no activity at all on a connection, the BS may assume the SS has lost connectivity. This threshold does not need to be set as high as TH_DISC_IDLE because of this reason.

The next time (after a BS initiated connection termination) the SS needs to send data on that particular connection, it needs to use contention to send a dynamic service addition MAC message (DSA-REQ) to the BS. Upon receiving this message, the BS will reactivate polling with the default period. No extended handshaking is

required. The SS can also piggyback a grant management subheader with the PM bit and the appropriate CID. The BS upon receipt of this will restore the poll for that connection.

Receiving any transmission for the particular CID will reset the DEAD time counter. Receiving a BW-REQ will cause a reset of the IDLE counter as well as the DEAD time counter.

7.4.3 BW-REQ Substitution with Small Packets

In an OFDM based system we have discussed the minimum BW allocation being an integer multiple of OFDM symbols. When a unicast poll is allocated to a SS if the size of the payload to be transmitted would fit into the BW-REQ slot the SS may transmit the payload with a grant management subheader.

A practical example would be downlink (DL) data transfers using FTP for peer-to-peer networking. In this case the UL will consist mostly of TCP acknowledgements which are small packets. If the SS needs to transmit a larger packet it can send a BW-REQ as per normal behaviour. Compare the proposed method with the UGS service class which offers BW grants with minimal delay. This enhancement is a SS side modification which only requires the BS to be able to receive a PDU of unexpected UIUC. This will serve to reduce overheads even further and minimize access delay by bypassing the BW-REQ queue.

7.4.4 BW-REQ Queue Management

The customary T16 timers used for BW-REQ timeout detection will not be used in our implementation. The uncertainty of the BS not receiving the BW-REQ due to poor signal quality is nullified by having the BS provide strict retransmission opportunities in immediately following frames. The SS does not need (nor will it be allowed to use contention) to contend for BW to transmit data, if there exists an active connection.

An nrtPS allocation will be given to a SS only if a BW-REQ for the same CID does not already exist in the BW-REQ queue. This could happen if the SS has in the few preceding frames been serviced and sent up a piggybacked BW-REQ which is

now at a certain position within the queue. When the timer expires for the next poll the BS will find the above said BW-REQ and ignore the poll.

These two conditions prevent the BS receiving duplicate BW-REQs for the same CID, and same payload. Piggybacking will be enabled by default for e-nrtPS. So the queue length at most (and in most conditions) will be equal to the number of active subscribers, n_a, i.e., there will only be one BW-REQ for a given nrtPS SS in the BW-REQ queue.

The proposed request/transmission policy for e-nrtPS is not to allow contention for data transfer but only for management messages. Management messages such as DSX messages may use any BW-REQ transmit opportunity and will be given preferential treatment by being queued in a separate queue, and serviced in a strict priority basis by the BS.

7.5 Comparison with Contention Based Access

We compare e-nrtPS based access with contention based BE access to see how the efficiency, throughput and access delay vary under different traffic conditions in both the UL and DL directions. The three basic scenarios are,

- DL bulk data transfer
- UL bulk data transfer
- DL bursty (HTTP) data transfer.

The simulations are carried out using a fixed WiMAX simulator developed for the simulation package QualNet3.9.5. We use a 256 subcarrier OFDM physical layer and a TDMA frame structure with a frame duration of 4 ms. UL/DL partitioning, the number of contention slots, queue lengths etc can be set as required. Adaptive modulation and coding is enabled as default. The SSs are uniformly distributed within the cell area with the BS in the centre. We have set aside 25% of the frame time (1 ms of the 4 ms frame) for UL data transfer. This includes the contention period and initial as well as periodic ranging regions in all simulations. Unless specifically stated the poll period used is 100 ms.

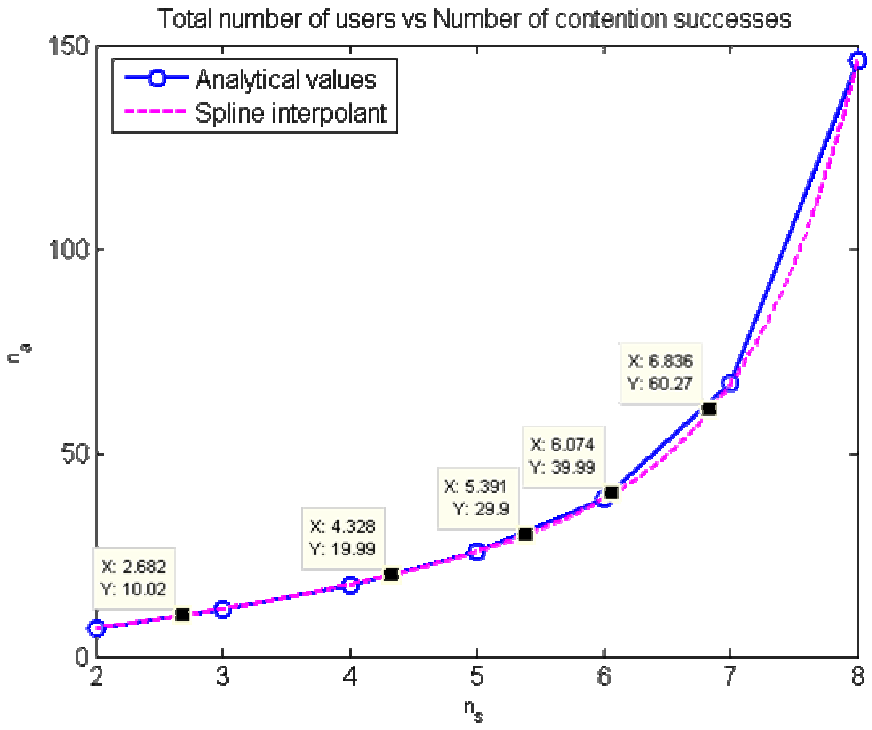

Figure 7-10 Number of active users calculated from (14) for different values of contention success. The data tips show n_s and na for n_a=10, 20, 30, 40 and 60. A cubic spline interpolant is used to estimate the values between the plotted points.

In order to evaluate our analytical and simulation model against the standard BE access method firstly we need to obtain the distributions of the various aspects of the queuing process, namely PDF(a), PDF(b) and PDF(a_p).

Figure 7-10 gives the curve of n_a vs n_s for the parameters specified in the scenario. The found n_s values are used to calculate the analytical UL and DL throughput in sections 7.5.1 and 7.5.3. These values are also used to feed into the analysis to produce the PDF(n_s) which in the general terms of our analysis is given as the arrival process of BW-REQ, a.

Figure 7-11 shows a comparison of the analytical PDF(n_s) with the values obtained from the statistics of a simulation run with 120 active SSs for a period of 5 minutes using 20 contention slots. The two distributions are extremely well matched. As explained previously a saturated system is assumed where the PDF(a_p) is identical to the PDF(a), i.e., the PDF(n_s).

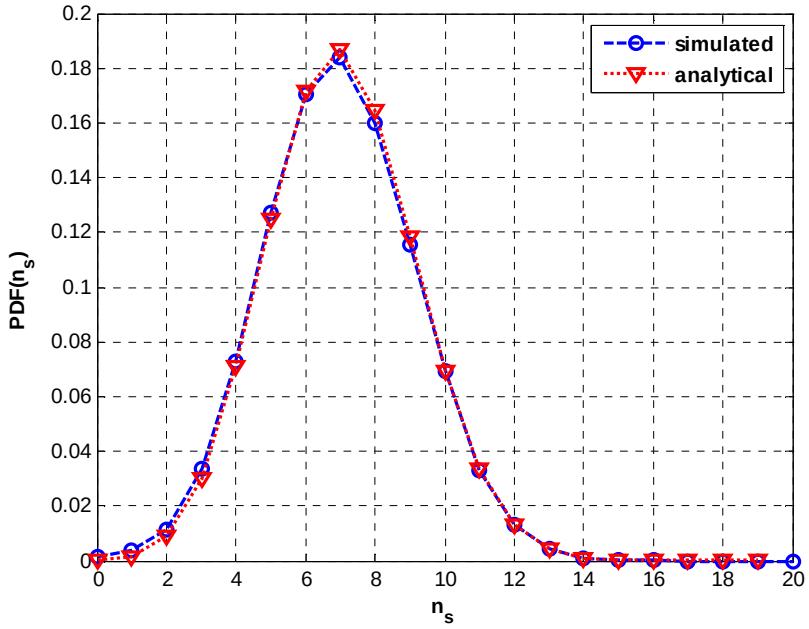

Figure 7-11 Probability distribution of the number of successes, analytical compared with simulated.

In this scenario there will be two packet sizes. 46 Bytes for TCP acknowledgements and 1030 Bytes for data, inclusive of the Mac layer header. For ease of comparison the cell considered is "small" with all SSs using the highest four burst profiles. Equation (7.4) is used to calculate the mean service rate and service distribution for each simulation scenario.

7.5.1 Bulk Data Transfer – Down Link

We investigate the performance of the e-nrtPS scheme and compare it to the contention based access method for DL FTP transfers. The analysis matches well with the DL throughput for both schemes, as seen in Figure 7-12. The e-nrtPS scales down well and can allow high per-user data rates even for a small number of users. This contrasts with the data rate achieve by the contention based scheme with a low number of users. As the system load increases, the e-nrtPS scheme suffers slightly while the contention based scheme improves. The reason for this small drop in throughput is due to the IP queue at the BS being overwhelmed by the incoming traffic. This in turn has a slowing down effect on TCP which isn't taken into account in the analysis. A certain amount of BW is also needed for the polling process which reduces the amount available for the DL.

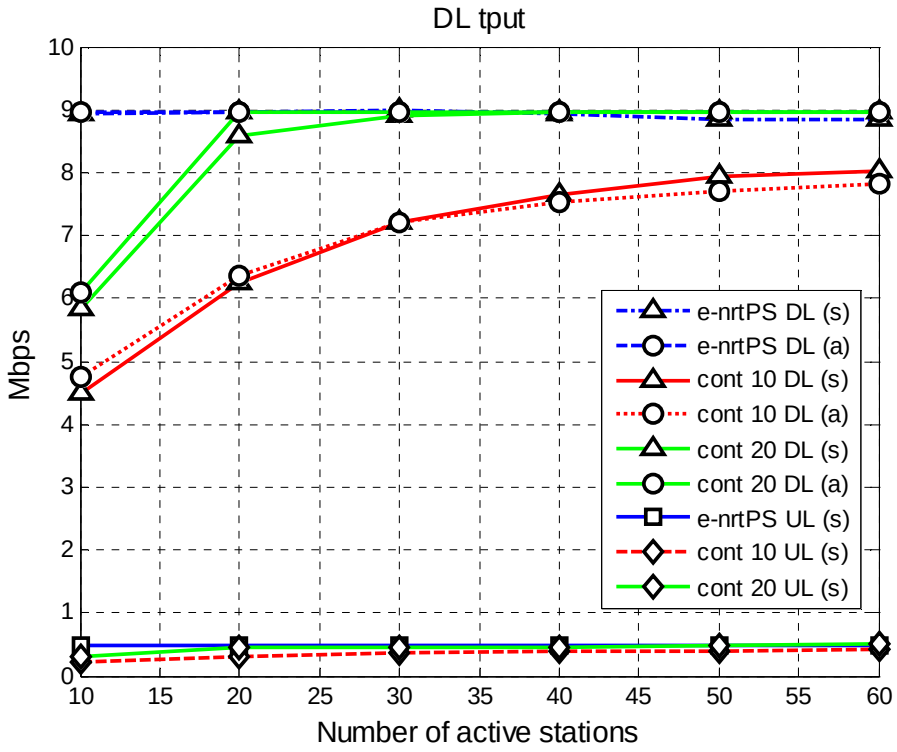

Figure 7-12 Comparison of simulation and analytical results for DL throughput under increasing loads using FTP as the transfer protocol.

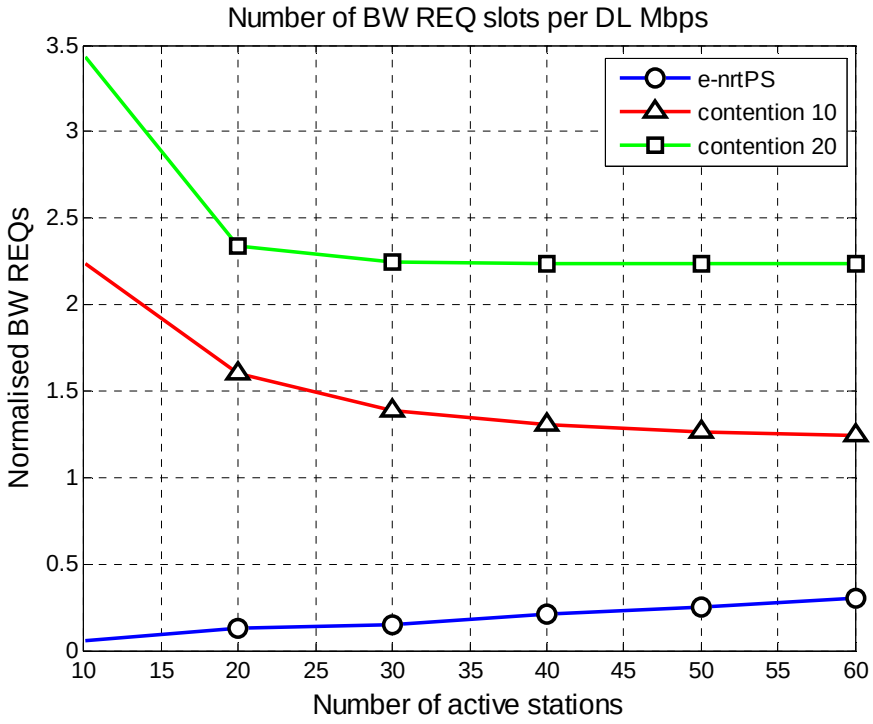

Figure 7-13 Normalised OH (BW REQ slots/frame/Mbps) for e-nrtPS and contention for increasing load using FTP as the transfer protocol.

With 10 contention slots the bottle neck is the number of successes, which in turn is implicitly the number of TCP Acknowledgements (ACKs) that the system can accept. Our model produces results which match well with the simulation. However, in the 20 slot case the bottle neck is no longer the contention process. Instead it is the maximum number of packets that can be served on the DL. Under the simulation scenario the maximum number of DL packets that can be served is 8. So for any UL packet rate more than 4 packets per frame, we would expect the DL to saturate and be the point of the bottleneck.

In order to compare the OHs of the two schemes we normalize the number of BW-REQ slots by the DL throughput. This gives a metric of 'slots per Mbps'. The curves given in Figure 7-13 for 10 slots and 20 slots, show a significantly higher OH when compared to the e-nrtPS scheme.

7.5.2 Bulk Data Transfer – Up Link

Next we look at the throughput for UL bulk data transfers. For the UL we do not consider HTTP flows as it would be uncommon for SSs to host HTTP servers.

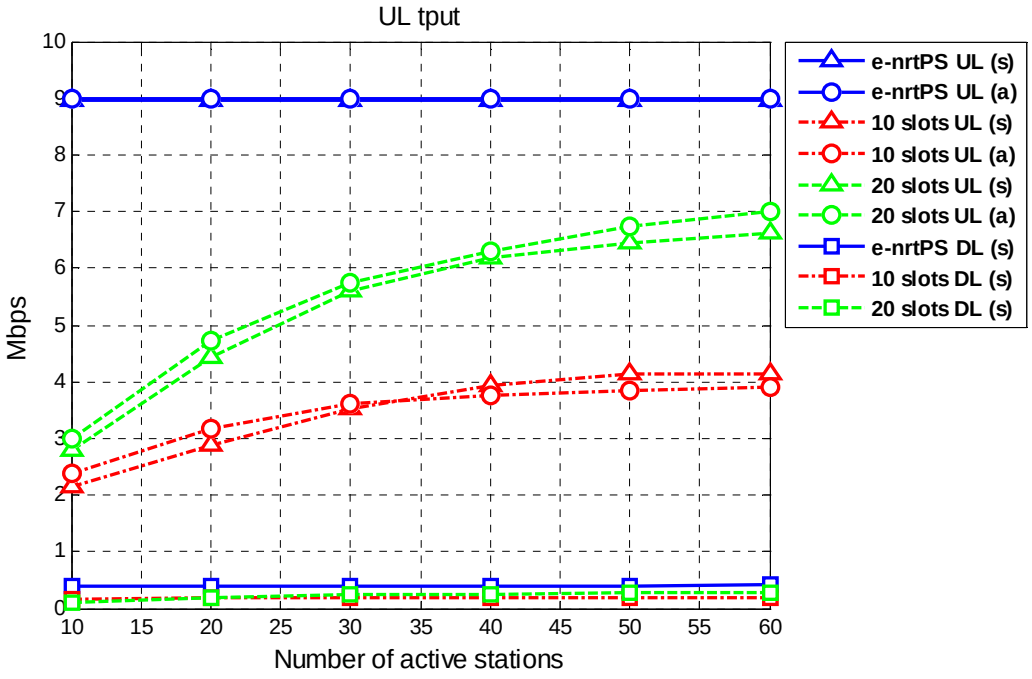

Figure 7-14 Comparison of simulation and analytical results for UL throughput under increasing loads using FTP as the transfer protocol.

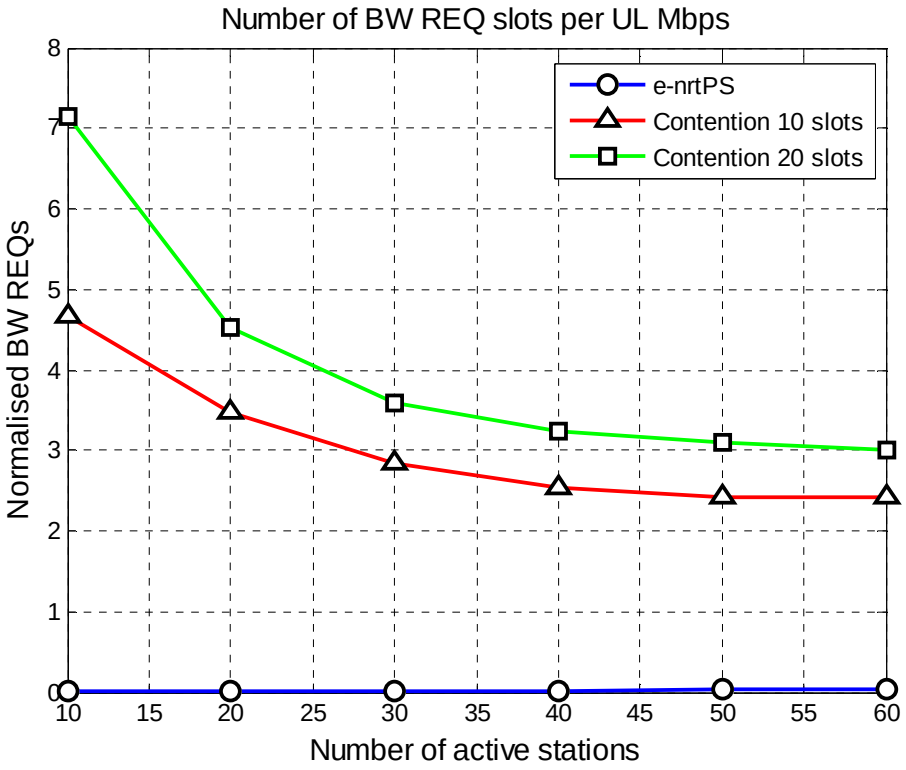

Figure 7-15 Normalised OH (BW REQ slots/frame/Mbps) for e-nrtPS and contention. Simulated increasing load using FTP as the transfer protocol.

Figure 7-14 shows the increased throughput with the e-nrtPS scheme. The per-Mbps overhead is shown in Figure 7-15. An increase of throughput is achieved in the 20 slot case over the 10 slot case. However the overheads have proportionally increased as well. In terms of frame time, the overhead is between 20%~40% of the time allocated for UL in our scenario. Polling based access can utilize all available bandwidth and shows better scalability than contention based access.

7.5.3 Bursty Traffic – Down Link

As stated previously we only consider HTTP based DL bursty traffic flows. The simulations shown in following section are from the perspective of a single SS as well as of a cell as a whole using standard nrtPS as defined by the standard and e-nrtPS as per our proposed modification.

As the poll period is increased the average throughput for a single station drops, Figure 7-16. The simulation is run for a duration of 30 minutes with the same sequence of HTTP flows, varying the poll period between 20 ms and 500 ms. The

BW saving enhancement gradually drops the poll period to 500 ms during inactive times.

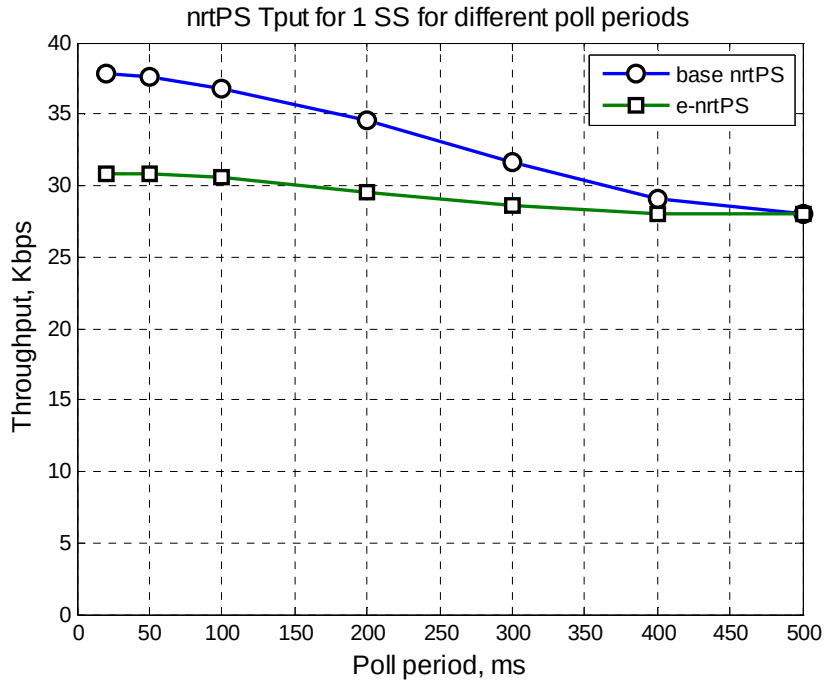

Figure 7-16 Throughput of a single SS downloading HTTP traffic. The same sequence of pages and sizes is used for every simulation run.

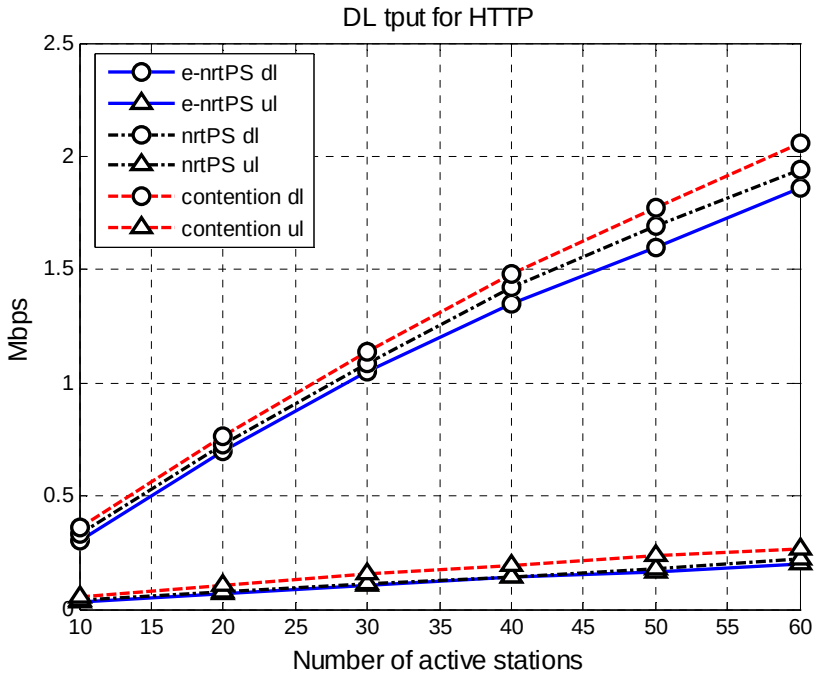

Figure 7-17 Comparison of HTTP UL and DL throughput for e-nrtPS, nrtPS and contention based access. Contention uses 20 BW-REQ slots per frame.

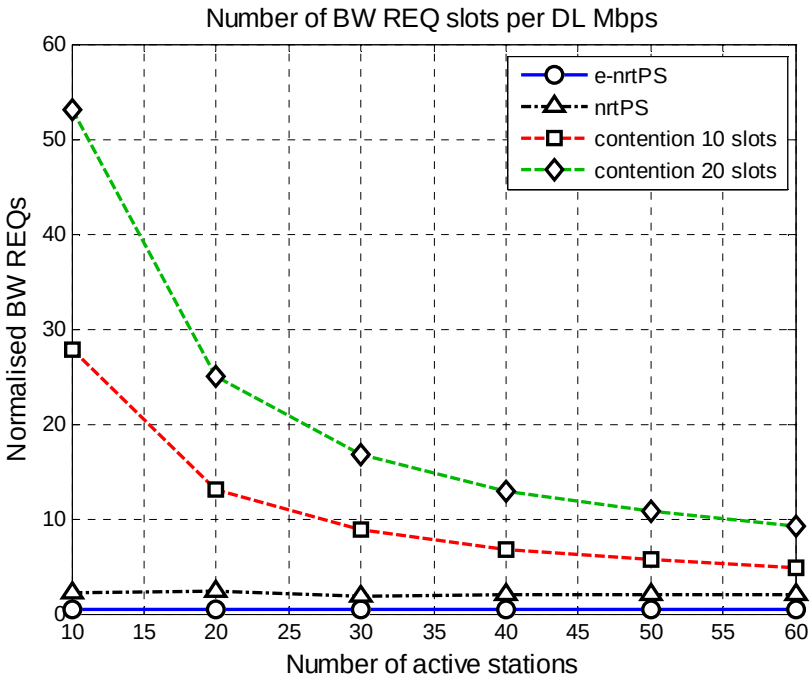

Figure 7-18 Normalised OH (BW REQ slots/frame/Mbps) for e-nrtPS, nrtPS and contention with 10 and 20 slots. Traffic type is DL HTTP.

When the application has been idle for a long enough time period for the BW saving enhancements to come into effect the latency for the first packet of the flow could be as much as the maximum poll period. While this feature is the main reason for the low overheads and high BW saving in e-nrtPS it is also the only draw back. This increase in "starting latency" is the cause for the drop in throughput seen in Figure 7-16. When cell aggregate throughput is compared the contention based scheme with 20 slots is slightly ahead. This higher performance comes at the expense of almost 20 times more overheads, Figure 7-18. These types of flows are the lowest priority for any ISP which makes it extremely important to reduce overheads.

The main consideration in this section is DL bursty throughput. However we have included the UL throughput in Figure 7-17. The reason for this is that any TCP based flow being bidirectional, depends on both the UL as well as the DL. The e-nrtPS enhancements directly affect the UL flow which implicitly affects the DL flow. This explanation also applies to all preceding TCP based traffic throughput plots.

7.5.4 Mix of Bulk and Bursty Traffic

In order to gauge the performance of e-nrtPS with both types of traffic we simulate a cell with 60 SSs communicating simultaneously. 30 of those are HTTP clients downloading data from 30 servers on an external subnet. 20 are downloading data from 20 external FTP servers and the other 10 are uploading FTP data to external clients. 2 ms (50% of the frame time) has been set aside for all these flows with a minimum of 1 ms set aside for the UL subframe. A default polling period of 100 ms is used for e-nrtPS while 20 contention slots are used for the contention based method. The results are given in Table 7-2.

Table 7-2 Simulation results for aggregate throughput for mixed traffic type case.

Metric	Throughput Contention (Mbps)	Throughput e-nrtPS (Mbps)	Throughput % increase
Aggregate Values			
Total UL T'put	2.178	8.072	270 %
Total DL T'put	7.601	8.468	11.4 %
Per SS Values			
UL FTP T'put	0.163	0.681	318 %
DL FTP T'put	0.251	0.319	27 %
DL HTTP T'put	0.0375	0.0358	-4.5 %

Except for the HTTP client throughput (which shows a drop of less than 4.5%) all other metrics show a much better performance, especially the UL throughput. The reason for a lower HTTP throughput is due to the BW saving enhancement made to the e-nrtPS scheme. Note that the UL FTP throughput aggregate increase for the cell, is less than the sum of throughput increase for all SSs combined due to some redistribution of OHs.

7.6 Conclusion

Best effort traffic has always been serviced using the lowest priority service classes in most MAC protocols. In previous chapters, we have done detailed analysis on the performance of the BE service class, in terms of the underlying contention resolution technique employed by fixed WiMAX. In this chapter the queuing and de-queuing processes have also been taken into account and an analytical model created to form a

more complete picture of the BE class. While there are minimal QoS guarantees for BE traffic, the amount of overhead is very high. The goal has been to enhance the polling service to be able to cope with the dynamic resource requirements of BE traffic while at the same time reduce the amount of overheads so as to be a viable replacement for contention based access. We have repeated the analysis of the arrival and departure process for nrtPS based access. We take into account the behaviour (service rates, waiting times, dropping rates) of the bandwidth request queues at the base station to estimate throughput. Using representative parameters for fixed WiMAX, analytical as well as numerical results have been generated which we have compared with simulation results.

Several modifications have been proposed to increase efficiency of the nrtPS polling scheme. We employ dynamic retarding of the poll period by the BS when idle periods are detected. We have also suggested a threshold based scheme for disconnection of connections and a dummy/keep-alive transmission scheme. The simulation model used includes all proposed enhancements which have been built on top of the IEEE 802.16d standard.

Simulations for bulk data transfer show that e-nrtPS has a clear advantage over contention for BE traffic. More so for UL data transfer with almost a 30% increase in throughput with much lower overheads and finite access delays. For DL bulk data transfer the improvement is in the reduction in overheads, which is significant. For bursty traffic the contention based access method performs slightly better than nrtPS but at the cost of wasted frame time for contention slots. The analysis done agrees well with the simulation results. It is clear that contention based access can produce high throughput, but at the expense of substantial BW wasted on overheads. Comparatively e-nrtPS can produce very high utilization (higher that contention in most cases) of BW but with minimal overheads.

Chapter 8

Conclusion

8.1 Thesis Contributions

The relationship between VoIP packet size and bandwidth efficiency was addressed in Chapter 4. VoIP applications use codecs which sample, compress and packetize, analogue voice into a VoIP stream. WiMAX uses an OFDM PHY, in which a MAC PDU uses an integer number of OFDM symbols on the airlink. Virtually an entire OFDM symbol could be wasted, if the packet size is such that it requires only a few bits of the last symbol. Hence the efficiency of bandwidth usage is directly affected by the packet size. This is more pronounced when the packet size is relatively small, such as, in VoIP applications. The packetization interval determines the packet size. The effects of the packetization interval on system resource usage and QoS of the flow, have been analysed. It was shown that by careful selection of the packetization intervals for VoIP, the number of users can be increased, by minimizing bandwidth wastage on overheads, and padding of OFDM symbols.

A method of selecting the packetization interval based on, packet loss rate, bandwidth usage and latency, was introduced. An index called "**Usability Factor**", ***K***, was defined. This is a measure of how suited a given packetisation interval is, to the prevalent conditions, and the QoS requirements. A new flexible retransmission

strategy for UGS flows was also introduced, in order to facilitate fast recovery of lost UGS packets.

Modifications were proposed to the MAC layer operation, to be able to change the packetization interval during UGS service setup and also during periodic ranging or ranging on demand. This modification can be accommodated in the existing ranging process, and Dynamic Service Addition/Change handshaking process. The Usability Factor can be quantized, and stored as a lookup table, which gives the BS and the SSs a simple way of selecting an optimal interval. A proof of concept was also given based on a simulation scenario, which shows positive results in terms of increased efficiency.

The importance of VoIP as one of the main service types over WiMAX, led to development of an efficient ARQ feedback scheme, as detailed in Chapter 5. This specifically caters for small packets in the DL direction. Through analysis of the standard based ARQ feedback scheme, we have shown that, irrespective of the packet error rate of the link, a substantial proportion of the bandwidth is used for the feedback messages. Transmitting many small packets makes the feedback process very bandwidth hungry and inefficient. However without ARQ, we cannot guarantee any level of QoS. To improve on this without sacrificing performance, we have proposed a contention based negative acknowledgement ARQ scheme. The defining difference between this scheme and many others, as well as the standard scheme implemented in WiMAX, is the feedback mechanism. Subscribers who received erroneous packets contend to send feedback to the base station. We have analytically proven the viability of this scheme in terms of overhead bandwidth usage, and rate of successful packet delivery. The analysis has been validated with simulations, which show a very good performance improvement. It has been shown that for packet error rates lower than 10%, our scheme is more efficient.

The performance of the Best Effort service class using contention based access was analysed in Chapter 6. Best effort traffic may be of low priority in terms of QoS, but it is the most important component of a transport network in terms of quantity. Traditionally, most PMP MAC protocols service BE traffic using contention based access. This provides a means of sharing limited bandwidth, among a large number of stations in a stochastically fair method. An analytical model for the mandatory contention mechanism of Fixed WiMAX was created using discrete time Markov

analysis. The two-dimensional Markov chain includes all states a subscriber station goes through in its request/backoff procedure. Any state in the chain represents a stage in the backoff process. We have made allowance for 'idle' states when a subscriber waits for the base station to grant it bandwidth. This analysis differs from previous work in that there is no explicit acknowledgement to the subscriber of failure.

From the Markov chain model we have derived expressions for access delay and access rate. From theory of occupancy we have derived analytical expression for the number contenders, the number of successes for a given number of contention slots and active stations. Optimal values of these parameters have been derived based on increasing contention success for a fixed number of contention slots. Our simulation scenario has shown that the analysis is accurate in terms of collision rates and access delay. The analytical expression need to be numerically solved to produce values which are then compared with corresponding output from our WiMAX simulator.

In addition to a general scenario we have investigated the ability to predict throughput of TCP based downlink traffic flows using the model. Our TCP model is only accurate enough to model steady state TCP flow using the delayed ACK scheme. The comparison of analytical and simulation results shows much promise. An attempt has been made to regulate TCP based FTP traffic by dynamically adjusting the number of contention opportunities in a frame.

In Chapter 7, we investigate using nrtPS as an alternative to contention based access for Best Effort traffic. Analysis of the contention resolution scheme for Best Effort traffic shows that, while bandwidth can be provided to SSs on demand, the system overheads are high. The goal has been to enhance the polling service, to be able to cope with the dynamic resource requirements of BE traffic, while at the same time reduce the amount of overheads. The analysis of the arrival and departure process for nrtPS was repeated, taking into account the behaviour of the bandwidth request queues at the base station. Using representative system parameters for fixed WiMAX, analytical as well as numerical results have been generated, which we have compared with simulation results.

Several modifications have been proposed to increase efficiency of the nrtPS polling scheme. We employ dynamic retarding of the poll period by the BS when idle

periods are detected. We have also suggested a threshold based scheme for disconnection of connections, and a dummy/keep-alive transmission scheme. Simulations for bulk data transfer show that, the proposed e-nrtPS scheme has a clear advantage over contention. More so for UL data transfer, with almost a 30% increase in throughput, with much lower overheads and finite access delays. For DL bulk data transfer, the improvement is in the reduction in overheads which is significant. For bursty traffic, the contention based access method performs slightly better than e-nrtPS, but at the cost of wasted frame time for contention slots. The analysis done agrees well with the simulation results. It is clear that contention based access can produce high aggregate throughput, but at the expense of substantial overheads. Comparatively, e-nrtPS can produce very high utilization, with minimal overheads.

8.2 Future Work

Considering the work covered in this thesis, and the movement of the relevant standards, it would be useful to highlight some future areas of investigation. For any QoS service class to fulfil its potential, robust policing, shaping and scheduling is needed. Methods of cross-layer aware scheduling should be researched. Optimisation of TCP over WiMAX would be extremely valuable. One also can identify issues regarding Mobile WiMAX or IEEE 802.16e which will be the next logical step. With a complex OFDMA PHY, subchannel allocation capabilities and finer sectorisation of cells, it is important to have efficient schemes to make the most of these capabilities. Scheduling becomes three dimensional with spatial diversity. Issues with seamless handover become relevant in the mobile scenario. Interoperability between OFDM based next generation cellular networks, as well as Wi-Fi networks is another area with potential for research. Due to the increasing popularity of WiMAX, time spent in enhancing it will no doubt be time well spent.

References

Amjad, M. K. and A. Shami (2006). Improving the Throughput Performance of IEEE 802.11 Distributed Coordination Function. 23rd Biennial Symposium on Communications 2006

Baines, R. (2006) "The Roadmap to Mobile Wimax." IEE, Communications Engineer, Volume 3, Issue 4, Aug.-Sept. 2005 Page(s):30 - 34:

Balanis, C. A. (1977). Antenna Theory: Analysis and Design (2nd edition), John Wiley and Sons Inc.

Ben-Jye, C. and C. Chien-Ming (2006). Adaptive Polling Algorithm for Reducing Polling Delay and Increasing Utilization for High Density Subscribers in WiMAX Wireless Networks. 10th IEEE International Conference on Communication systems, 2006. ICCS 2006 Singapore . .

Ben-Jye, C., C. Yan-Ling, et al. (2007). Adaptive Hierarchical Polling and Cost-Based Call Admission Control in IEEE 802.16 WiMAX Networks. IEEE Wireless Communications and Networking Conference (WCNC), 2007.

Bianchi, G. (2000). "Performance analysis of the IEEE 802.11 distributed coordination function." IEEE Journal on Selected Areas in Communications, 18(3): 535-547.

Bianchi, G., L. Fratta, et al. (1996). Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs. Seventh IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, .

Bianchi, G. and I. Tinnirello (2005). "Remarks on IEEE 802.11 DCF performance analysis." IEEE Communications Letters 9(8): 765-767.

Bruno, R., M. Conti, et al. (2005). Throughput Analysis of UDP and TCP Flows in IEEE 802.11b WLANs: A Simple Model and Its Validation.

Cable Television Laboratories Inc (2007). Data-Over-Cable Service Interface Specifications, DOCSIS 3.0. CM-SP-MULPIv3.0-I05-070803 MAC and Upper Layer Protocols Interface Specification**:** 1-743.

Carl Eklund, Roger B. Marks, et al. (2007). WirelessMAN: Inside the IEEE 802.16 Standard for Wireless Metropolitan Area Networks, Standards Information Network, IEEE press.

Chang Wook, A., K. Chung Gu, et al. (1999). Hybrid ARQ protocol for real-time ATM services in broadband radio access networks. Proceedings of the IEEE Region 10 Conference TENCONN '99.

Chia-Hui, W., I. C. Ray, et al. (2003). Rate-sensitive ARQ for real-time video streaming. IEEEGlobal Telecommunications Conference, 2003. GLOBECOM '03.

Ching-Ling, H. and L. Wanjiun (2007). "Throughput and delay performance of IEEE 802.11e enhanced distributed channel access (EDCA) under saturation condition." IEEE Transactions on Wireless Communications 6(1): 136-145.

Chris Hellberg, Dylan Greene, et al. (2007). Broadband Network Architectures: Designing and Deploying Triple-Play Services, Prentice Hall.

Chuan Heng, F. and M. Zukerman (2001). Performance evaluation of IEEE 802.11. Vehicular Technology Conference, 2001. VTC 2001 Spring. IEEE VTS 53rd.

ETSI. (2007). "Metropolitan Area Networks." 2007, from http://www.etsi.org/etsisite/website/technologies/hiperman.aspx.

Fall, K. and K. Varadhan. (2006, April 6, 2006). "The ns Manual." 2006.

Ganesh Babu, T. V. J., T. Le-Ngoc, et al. (2001). "Performance of a priority-based dynamic capacity allocation scheme for wireless ATM systems." IEEE Journal on Selected Areas in Communications, 19(2): 355-369.

Goldman, J. (2005). "Moving WiMax Into VoIP." Insights, 2007, from http://www.wi-fiplanet.com/columns/article.php/3500176.

Goldman, J. (2005). "WiMax Certification Proceeds at Cetecom Labs." News, 2006, from http://www.wi-fiplanet.com/news/article.php/3526516.

Griffith, E. (2006). "802.11n Draft Approved." News, 2007, from http://www.wi-fiplanet.com/news/article.php/3578886.

GSM Association. (2008). "GSM Facts and Figures." 2007, from http://www.gsmworld.com/news/statistics/index.shtml.

Gurbuz, O. and E. Ayanoglu (2004). A transparent ARQ scheme for broadband wireless access. IEEE Wireless Communications and Networking Conference, 2004. WCNC 2004.

Gyung-Ho, H. and C. Dong-Ho (2004). Fast retransmission mechanism for VoIP in IEEE 802.11e wireless LANs. IEEE 60th Vehicular Technology Conference, 2004. VTC2004-Fall 2004

H. S. Chhaya, S. G. (1997). "Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol." Wireless Networks 3: 217–234.

Hardasmal, F. E. and T. Sanders. (2007). "A Conversation with Cetecom Laboratories." 2007, from http://www.wimax.com/commentary/spotlight/spotlight-cetecom.

Hattingh, C. and T. Szigeti (2004). End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs, Cisco Press.

Haykin, S. S. (1994). Communication systems. New York, Wiley.

Heyaime-Duverge, C. and V. K. Prabhu (2002). Traffic-based bandwidth allocation for DOCSIS cable networks. Proceedings of Eleventh International Conference on Computer Communications and Networks, 2002. .

Hong Shen, W. and N. Moayeri (1995). "Finite-state Markov channel-a useful model for radio communication channels." IEEE Transactions on Vehicular Technology, 44(1): 163-171.

Hoymann, C. (2005). "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16." Computer Networks 49(3): 341-363.

Hu, F., G. Zhu, et al. (2001). Enhanced ARQ-based packet loss recovery for real-time communication. International Conferences on Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001.

Hyogon, K., Y. Sangki, et al. (2005). "Resolving 802.11 performance anomalies through QoS differentiation." IEEE Communications Letters, 9(7): 655-657.

IEEE 802.16 WG (2004). IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001)**:** 0_1-857.

IEEE (2005). IEEE 802.16. IEEE Standard for Local and metropolitan area networks. Part 16: Air Interface For Fixed Broadband Wireless Access Systems. 3 Park Avenue, New York, NY 10016-5997, USA.

IEEE Computer Society (2004). IEEE Standard for Information technology- Telecommunications and information exchange between systems- Local and metropolitan area networks- Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 6: Medium Access Control (MAC) Security Enhancements. IEEE Std 802.11i-2004**:** 0_1-175.

IEEE Computer Society (2005). IEEE Standard for Information technology — Telecommunications and information exchange between systems — Local and metropolitan area networks — Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements. IEEE Std 802.11-1997, IEEE**:** i-189.

Jeffrey G. Andrews, Arunabha Ghosh, et al. (2007). Fundamentals of WiMAX: Understanding Broadband Wireless Networking, Prentice Hall.

Jianxin, W. and J. Speidel (2003). "Packet acquisition in upstream transmission of the DOCSIS standard." Broadcasting, IEEE Transactions on 49(1): 26-31.

Johnson, N. L. and S. Kotz (1977). Urn models and their application : an approach to modern discrete probability theory. New York, Wiley.

Johnson, N. L. and S. Kotz (1977). Urn models and their application : an approach to modern discrete probability theory. New York, Wiley.

Kai-Chien, C. and L. Wanjiun (2007). "The Contention Behavior of DOCSIS in CATV Networks." IEEE Transactions on Broadcasting, 53(3): 660-669.

Kai-Chuang, H. and C. Kwang-Cheng (1995). Interference analysis of nonpersistent CSMA with hidden terminals in multicell wireless data networks. Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1995. PIMRC'95.

Kanjanavapastit, A. and B. Landfeldt (2003). An analysis of a modified point coordination function in IEEE 802.11. 14th IEEE Conference on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. .

Khanna, V. K., H. M. Gupta, et al. (2005). A contention-free 802-11 protocol for multi-WLAN configurations. IEEE International Conference on Personal Wireless Communications, 2005. ICPWC 2005.

Lidong, L., J. Weijia, et al. (2007). Performance Analysis of IEEE 802.16 Multicast and Broadcast Polling based Bandwidth Request. IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007. .

Maaroufi, S., W. Ajib, et al. (2007). Performance Evaluation of New MAC Mechanisms for IEEE 802.11n. First International Global Information Infrastructure Symposium, 2007. GIIS 2007. .

Marks, R., D. P. Satapathy, et al. (2001) "New IEEE WirelessHUMAN™ Project Developing Standards for Fixed Wireless Access in License-Exempt Bands." DOI:

Miki, N., H. Atarashi, et al. (2003). Experimental evaluation on effect of hybrid ARQ with packet combining in forward link for VSF-OFCDM broadband wireless access. 14th IEEE Conference on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. .

Min, C. and W. Gang (2003). A novel hybrid ARQ algorithm for real-time video transport over wireless LAN. 14th IEEE Conference on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. .

Muscariello, L., M. Meillia, et al. (2004). An MMPP-based hierarchical model of Internet traffic. IEEE International Conference on Communications, 2004. ICC 2004

Nair, G., J. Chou, et al. (2004). "IEEE 802.16 Medium Access Control and Service Provisioning." Intel Technology Journal 08(03).

Niyato, D. and E. Hossain (2005). Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks. IEEE Global Telecommunications Conference, 2005. GLOBECOM '05. .

Niyato, D. and E. Hossain (2006). "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks." Mobile Computing, IEEE Transactions on 5(6): 668-679.

Ohrtman, F. (2005). WiMAX Handbook : Building 802-16 Wireless Networks. Emeryville, Calif., McGraw-Hill/Osborne.

Pareek, D. (2006). The business of WiMAX. Chichester, England ; Hoboken, NJ, John Wiley.

Perera, S. and H. Sirisena (2006). Contention Based Negative Feedback ARQ for VoIP Services in IEEE 802.16 Networks. 14th IEEE International Conference on Networks, 2006. ICON '06. .

Po-Chin, H., Z. Zhi-Li, et al. (2000). Channel condition ARQ rate control for real-time wireless video under buffer constraints. International Conference on Image Processing, 2000. .

Ramos, N., D. Panigrahi, et al. (2005). "Quality of service provisioning in 802.11e networks: challenges, approaches, and future directions." Network, IEEE 19(4): 14-20.

Scalable Network Technologies Inc. (2005, 2006). "Programmers Guide to Qualnet 3.8, 3.9.5." from www.snt.com.

Seferoglu, H., Y. Altunbasak, et al. (2005). Rate distortion optimized joint ARQ-FEC scheme for real-time wireless multimedia. IEEE International Conference on Communications, 2005. ICC 2005.

Seung-Eun, H., K. Oh-Hyeong, et al. (2006). Performance Analysis of Single- and Multi-Channel Contention Resolution Algorithm for the DOCSIS MAC Protocol. IEEE International Conference on Communications, 2006. ICC '06. .

Shepard, S. (2006). WiMax crash course. New York ; London, McGraw-Hill.

Sik, C., H. Gyung-Ho, et al. (2005). Fast handover scheme for real-time downlink services in IEEE 802.16e BWA system. IEEE 61st Vehicular Technology Conference, 2005. VTC 2005-Spring.

Stallings, W. (2005). Wireless Communications & Networks, Pearson Prentice Hall.

Stevens, W. R. (1994). TCP/IP illustrated. Reading, Mass., Addison-Wesley Pub. Co.

Stewart, W. J. (1994). Introduction to the numerical solution of Markov chains. Princeton, N.J., Princeton University Press.

Suitor, K. (2004). What WiMAX Forum Certified™ products will bring to Wi-Fi. Business White Paper, Broadband Wireless Access, WiMAX Forum, Redline Communications.

Sung-Min, O. and K. Jae-Hyun (2005). The analysis of the optimal contention period for broadband wireless access network. Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops.

TTA. (2005). "Standards." WiBro, 2006, from http://www.wibro.or.kr/new/standards01.jsp.

Uhlemann, E., T. Aulin, et al. (2002). Concatenated hybrid ARQ - a flexible scheme for wireless real-time communication. Eighth IEEE Real-Time and Embedded Technology and Applications Symposium, 2002. .

Wanang, X. and S. Yin (2005). A two-step backoff scheme for improving the performance of the IEEE 802.11 distributed coordination function. IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005. .

Wei-Tsong, L., C. Kun-Chen, et al. (2006). "DOCSIS performance analysis under high traffic conditions in the HFC networks." IEEE Transactions on Broadcasting 52(1): 21-30.

WiMAX Forum. (06-12-2006). "Welcome to the WiMAX Forum." from www.wimaxforum.org/home/.

WiMAX Forum. (2005). "WiMAX Forum Certification of Broadband Wireless Systems." 2006, from http://www.wimaxforum.org/technology/downloads/Certification_FAQ_final.pdf.

Wreless Design & Development Asia. (2007). "WiMAX OFDMA Technology Based on Runcom's Contribution Approved by the ITU as a 3G/4G Standard." Business News & Technology News, 25 Oct 2007, 2007, from http://www.wirelessdesignasia.com/article-7150-wimaxofdmatechnologybasedonruncomscontributionapprovedbytheituasa3g4gstandard-Asia.html.

Yaghoobi, H. (2004). "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN." Intel Technology Journal 8(3): 201-212.

Yang, X. (2003). Enhanced DCF of IEEE 802.11e to support QoS. IEEE Wireless Communications and Networking, 2003. WCNC 2003.

Zhang, H., Y. Li, et al. (2006). A New Extended rtPS Scheduling Mechanism Based on Multi-Polling for VoIP Service in IEEE 802.16e System. International Conference on Communication Technology, 2006. ICCT '06. .

Zhen-ning, K., D. H. K. Tsang, et al. (2004). "Performance analysis of IEEE 802.11e contention-based channel access." IEEE Journal on Selected Areas in Communications 22(10): 2095-2106.

Zhi, Q. and C. Jong-Moon (2004). Analysis of packet loss for real-time traffic in wireless mobile networks with ARQ feedback. IEEE Wireless Communications and Networking, 2004. WCNC 2004.

Zorzi, M. and R. R. Rao (1997). "On the statistics of block errors in bursty channels." IEEE Transactions on Communications 45(6): 660-667.

Bibliography

Andrews, J. G., A. Ghosh, et al. (2007). Fundamentals of WiMAX : understanding broadband wireless networking. Upper Saddle River, NJ, Prentice Hall.

Balamurali, N. and D. Jalihal (2004). An efficient algorithm for joint carrier frequency offset and channel estimation in IEEE 802.16 OFDM systems. 1st International Symposium on Wireless Communication Systems, 2004.

Carsenat, D. and N. Murad (2005). A threshold profile map definition for improved management of a 802.16 network. The 7th International Conference on Advanced Communication Technology, 2005, ICACT 2005.

Casilari, E., F. J. Gonzblez, et al. (2001). "Modeling of HTTP traffic." IEEE Communications Letters, 5(6): 272-274.

Chen, D. T. (2007). On the Analysis of Using 802.16e WiMAX for Point-to-Point Wireless Backhaul. IEEE Radio and Wireless Symposium, 2007.

Chen, J., J. Chen, et al. (2007). Traffic-Variation-Aware Connection Admission Control Mechanism for Polling Services in IEEE 802.16 Systems. IFIP International Conference on Wireless and Optical Communications Networks, 2007. WOCN '07.

Chen, J., W. Jiao, et al. (2005). A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode. IEEE International Conference on Communications, 2005. ICC 2005.

Chia-Hui, W., I. C. Ray, et al. (2003). Rate-sensitive ARQ for real-time video streaming. IEEEGlobal Telecommunications Conference, 2003. GLOBECOM '03.

Chingyao, H., J. Hung-Hui, et al. (2007). "Radio resource management of heterogeneous services in mobile WiMAX systems [Radio Resource Management and Protocol Engineering for IEEE 802.16]." Wireless Communications, IEEE [see also IEEE Personal Communications] 14(1): 20-26.

Cho, D.-H., J.-H. Song, et al. (2005). Performance analysis of the IEEE 802.16 wireless metropolitan area network. First International Conference on Distributed Frameworks for Multimedia Applications, 2005. DFMA '05.

Cooklev, T. (2004). Wireless communication standards : a study of IEEE 802.11, 802.15, and 802.16. New York, Standards Information Network IEEE Press.

Curwen, P. J. (2002). The future of mobile communications: awaiting the third generation. New York, Palgrave.

Dapeng, W. and R. Negi (2003). "Effective capacity: a wireless link model for support of quality of service." IEEE Transactions on Wireless Communications, 2(4): 630-643.

Dhawan, S. (2007). Analogy of Promising Wireless Technologies on Different Frequencies: Bluetooth, WiFi, and WiMAX. The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007.

Eklund, C. (2006). WirelessMAN : inside the IEEE 802.16 standard for wireless metropolitan area networks. New York, IEEE Press.

Eklund, C., R. B. Marks, et al. (2002). "IEEE standard 802.16: a technical overview of the WirelessMAN air interface for broadband wireless access." IEEE Communications Magazine, 40(6): 98-107.

Gal, D. (2005). "IEEE 802.20 Evaluation Criteria - Traffic Mix and QoS Simulation." IEEE 802.20 Working Group on Mobile Broadband Wireless Access, from http://grouper.ieee.org/groups/802/20.

Gentle, J. E. (2002). Elements of computational statistics. New York, Springer.

Gentle, J. E. (2003). Random number generation and Monte Carlo methods. New York, Springer-Verlag.

Guan, Y. and A. Hu (2006). Bandwidth Allocation Algorithm of VoIP Based on the Adaptive Linear Prediction in the IEEE 802.16 System. ITS 6th International Conference on Telecommunications Proceedings, 2006.

Haitao, W., C. Shiduan, et al. (2002). IEEE 802.11 distributed coordination function (DCF): analysis and enhancement. IEEE International Conference on Communications, ICC 2003.

Hasan, M. A. (2007). Performance Evaluation of WiMAX/IEEE 802.16 OFDM Physical Layer. Department of Electrical and Communications Engineering, Communications Laboratory. Helsinki, Helsinki University of Technology

Hayes, J. (1968). "Adaptive Feedback Communications" IEEE Transactions on Communications, 16(1): 29-34.

Heidemann, J., K. Obraczka, et al. (1997). "Modeling the performance of HTTP over several transport protocols." IEEE/ACM Transactions on Networking, 5(5): 616-630.

Heyaime-Duverge, C. and V. K. Prabhu (2002). Traffic-based bandwidth allocation for DOCSIS cable networks. Proceedings of Eleventh International Conference on Computer Communications and Networks, 2002. .

Hong Shen, W. and N. Moayeri (1995). "Finite-state Markov channel-a useful model for radio communication channels." IEEE Transactions on Vehicular Technology, 44(1): 163-171.

Howon, L., K. Taesoo, et al. (2006). Extended-rtPS Algorithm for VoIP Services in IEEE 802.16 systems. IEEE International Conference on Communications, 2006. ICC '06.

Howon, L., K. Taesoo, et al. (2006). Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems. IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring.

IEEE (2005). IEEE 802.16. IEEE Standard for Local and metropolitan area networks. Part 16: Air Interface For Fixed Broadband Wireless Access Systems. 3 Park Avenue, New York, NY 10016-5997, USA.

IEEE (2006). IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1. IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004)**:** 0_1-822.

Kai-Chien, C. and L. Wanjiun (2007). "The Contention Behavior of DOCSIS in CATV Networks." IEEE Transactions on Broadcasting, 53(3): 660-669.

Koffman, I. and V. Roman (2002). "Broadband wireless access solutions based on OFDM access in IEEE 802.16." IEEE Communications Magazine, 40(4): 96-103.

Lee, H., T. Kwon, et al. (2004). An efficient uplink scheduling algorithm for VoIP services in IEEE 802.16 BWA systems. Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th.

Lee, H., T. Kwon, et al. (2005). "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system." Communications Letters, IEEE 9(8): 691-693.

Nair, G., J. Chou, et al. (2004). "IEEE 802.16 Medium Access Control and Service Provisioning." Intel Technology Journal 8(3): 213-219.

Nie, J., J. Wen, et al. (2005). A seamless handoff in IEEE 802.16a and IEEE 802.11n hybrid networks. International Conference on Communications, Circuits and Systems, 2005.

Nuaymi, L. (2007). WiMAX : technology for broadband wireless access. Chichester, England ; Hoboken, NJ, John Wiley.

Parvez, N. and L. Hossain (2004). Improving TCP performance in wired-wireless networks by using a novel adaptive bandwidth estimation mechanism.

Ramachandran, S., C. W. Bostian, et al. (2005). A link adaptation algorithm for IEEE 802.16. Wireless Communications and Networking Conference, 2005 IEEE. WNC 2005.

Redana, S., M. Lott, et al. (2004). Performance evaluation of point-to-multi-point (PMP) and mesh air-interface in IEEE standard 802.16a. IEEE 60th Vehicular Technology Conference, 2004. VTC2004-Fall.

Roberts, R., R. Hoshyar, et al. (2004). Choice of interleavers for space-diversity codes in HIPERMAN and 802.16a broadband wireless systems. IEEE 15th International Symposium on Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004.

Sayenko, A., O. Alanen, et al. (2007). Adaptive Contention Resolution for VoIP Services in the IEEE 802.16 Networks. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007.

Sengupta, S., M. Chatterjee, et al. (2006). Improving R-Score of VoIP Streams over WiMax. IEEE International Conference on Communications, 2006. ICC '06.

Seung-Eun, H. and K. Oh-Hyeong (2006). Considerations for VoIP Services in IEEE 802.16 Broadband Wireless Access Systems. IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring.

Sung-Min, O. and K. Jae-Hyun (2005). The analysis of the optimal contention period for broadband wireless access network. Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops.

Tarchi, D., R. Fantacci, et al. (2006). Quality of Service Management in IEEE 802.16 Wireless Metropolitan Area Networks. IEEE International Conference on Communications, 2006. ICC '06.

Tornasin, S. and N. Benvenuto (2004). Performance comparison of frequency domain equalizers for the IEEE 802.16a WMAN standard. International Conference on Information and Communication Technologies: From Theory to Applications, 2004. 2004.

Wanjiun, L. and J. Huei-Jiun (2004). "Adaptive slot allocation in DOCSIS-based CATV networks." IEEE Transactions on Multimedia, 6(3): 479-488.

Wei-Tsong, L., C. Kun-Chen, et al. (2006). "DOCSIS performance analysis under high traffic conditions in the HFC networks." IEEE Transactions on Broadcasting 52(1): 21-30.

Wen-Kuang, K., S. Kumar, et al. (2003). "Improved priority access, bandwidth allocation and traffic scheduling for DOCSIS cable networks." IEEE Transactions on Broadcasting, 49(4): 371-382.

You, J., K. Kim, et al. (2005). Capacity evaluation of the OFDMA-CDMA ranging subsystem in IEEE 802.16-2004. IEEE International Conference on Wireless And Mobile Computing, Networking And Communications, 2005. (WiMob'2005),

Zhi, Q. and C. Jong-Moon (2004). Analysis of packet loss for real-time traffic in wireless mobile networks with ARQ feedback. IEEE Wireless Communications and Networking, 2004. WCNC 2004.