Quality-Driven Multi-User Resource Allocation and Scheduling Over LTE for Delay Sensitive Multimedia Applications

Nabeel Khan

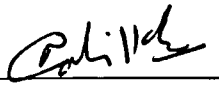
A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science, Faculty of Science, Engineering and Computing, Kingston University

2014

Declaration of Authorship

I, Nabeel khan, declare that this thesis titled, 'Quality-Driven Multi-User Resource Allocation and Scheduling Over LTE for Delay Sensitive Multimedia Applications' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 28/01/2014

Acknowledgements

This dissertation was written during my time as a PhD candidate in the Wireless Multimedia and Networking Research Group at the School of Computing and Information Systems (Faculty of Science, Engineering and Computing of Kingston University). The outcome of this work could not be accomplished without the helps and supports for many people in many ways.

First, I would like to express my sincere appreciation to my PhD advisor, Dr. Maria G. Martini, for accepting me as one of her PhD candidate at her research group. I am grateful to Dr. Maria for her suggestions, guidances and many helpful and stimulating discussions during various phases of the dissertation. In particular, I thank Dr. Maria for providing me all facilities for conducting my research in a conducive environment. During the period of her supervision, I learned a lot from her on how to proceed step-by-step till the end of my PhD thesis and her attitudes of writing a good conference/journal paper.

I would like to give a special thank to my project manager at DOCOMO Euro-Labs, Dr. Dirk Staehle, for his valuable comments, fruitful discussions and motivation during the final year of the project.

I am greatly indebted to all the colleagues of my research group and all PhD candidates for sharing a good time and providing a friendly and pleasant atmosphere during the entire period of my PhD.

Nabeel Khan

Abstract

The expectation from a future generation cellular network is to provide multiplay applications of VoIP, video and data to a continuously growing number of cellular users. The scarcity of the available radio spectrum coupled with the unique traffic handling and Quality of Experience (QoE) requirements of the converged services poses a huge challenge to the network operators. The solution of over-provisioning the network by increasing the amount of bandwidth is not economical. Therefore, efficient partition of network resources becomes mandatory. Scheduling plays an important in determining the overall efficiency of a wireless system. This thesis focuses on quality driven scheduling for efficient resource allocation in multi-user downlink LTE systems. Video traffic contributes a major proportion of network traffic. Therefore, one of the main goals of this work is to design scheduling strategies which consider information about video traffic with the aim of improving the service quality perceived by the user. Various scheduling strategies are proposed taking into account different criteria such as packet delay and importance of a video packet. This thesis presents a novel cross-layer resource allocation architecture which reduces the need for cross-layer signaling and frequent end-to-end link probing (for video rate adaptation) required by other cross-layer approaches. Apart from the novel cross-layer architecture, the thesis applies the concepts of game theory and fuzzy logic frameworks in radio resource management and proposes a composite scheduling rule which considers the service needs of different traffic types such as video, VoIP and data. Results show that the proposed scheduling schemes lead to an efficient partition of radio resources while achieving a significant improvement in the perceived quality as compared to state-of-the-art scheduling rules.

Contents

List of Figures

List of Tables

Abbreviations

AM	Acknowledged Mode
AMC	adaptive modulation and coding
AWGN	Additive White Gaussian Noise
ARQ	Automatic Repeat Request
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
CBR	Constant Bit Rate
CGS	Coarse Grain Scalability
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CRC	Cyclic Redundancy Check
DL-SCH	Downlink Shared Channel
DPS	Delay Prioritized Scheduler
DRC	Dynamic Resource Controller
eNodeB	LTE base station
EESM	Exponential Effective SINR Mapping
EXP	Exponential
EXP/PF	Exponential Proportional Fair
FC	Frame Copy
FCS	Fuzzy Composite Scheduling
FDD	Frequency Division Duplexing
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIFO	First In, First Out
GMSK	Gaussian Minimum Shift Keying

GOP	Group of Pictures
HARQ	Hybrid Automatic Repeat Request
HoL	Head of Line
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
IFFT	Inverse Fast Fourier Transform
ISI	Inter Symbol Interference
LIP	Linear Integer Programming
LTE	Long-Term Evolution
MAC	Medium Access Layer
MBMS	Multimedia Broadcast and Multicast Service
MGS	Medium Grain Scalability
MIESM	Mutual Information Effective SNR Mapping
MIMO	Multiple Input and Multiple Output
M-LWDF	Modified Largest Weighted Delay First
M-LWDFQ	Queue-aware M-LWDF rule
MOS	Mean Opinion Score
NAS	Non access stratum
NBS	Nash bargaining solution
NPS	Nash Product Scheduling
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPF	Opportunistic Proportional Fair
OPLF	Opportunistic Packet Loss Fair
OS	Opportunistic Scheduling
PAPR	Peak-to-Average Power Ratio
PCCH	Paging Control Channel
PDCP	Packet Data Convergence Protocol
PDUs	Protocol Data Units
PF	Proportional Fair
P-GW	PDN Gateway
PLF	Packet Loss Fair
PLR	Packet Loss Rate

PRF	Priority Function
PMI	Precoding Matrix Indicator
PPS	Packet Priority Scheduler
PRB	Physical Resource Block
PSK	Phase Shift Keying
PSNR	Peak Signal to Noise Ratio
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase-Shift Keying
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RFPA	Radio Frequency Power Amplifier
RI	Rank Indicator
RLC	Radio Link Control
RNC	Radio Network Controller
ROHC	Robust Header compression
RRM	Radio Resource Management
RUI	Resource Utilization Index
SC-FDMA	Single Carrier-Frequency Domain Multiple Access
SINR	Signal to Interference Noise Ratio
SISO	Single Input Single Output
SNR	Signal to Noise Ratio
SVC	Scalable Video Coding
TB	Transport Block
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
TU	Typical Urban
UL-SCH	Uplink Shared Channel
UM	Unacknowledged Mode
UMTS	Universal Mobile Telecommunication System

Dedicated to my parents

Chapter 1

Introduction

The back bone of any business's prosperity is to realize the forces that will drive the business in the future. This also applies in the field of mobile-communication. A rapid increase in the number of subscribers has attracted several new vendors and operators. There is an ever increasing competition between new and existing vendors and operators to provide better services to the subscriber in a cost efficient manner. In order to achieve this, adopting, understanding, and efficiently using new technologies and standards becomes mandatory. Technical advancements in mobile devices with features such as mega-pixel cameras and high definition video recorders have extended the use of mobile phones from simple voice communication devices to multi-purpose highly interactive end devices running many services. In the near future, voice communication (VoIP) over future wireless networks will be a minority as shown in Figure 1.1. End applications have been extending from simple voice communication to highly interactive, bandwidth hungry and low delay requirement applications. These advancements in end applications have resulted in new communication technologies and standards so that demands in services can be met and at the same time more and more subscribers can be served with an acceptable quality of service. Therefore operators must ensure not only delivering services to some of the end users but different service needs (delay and data rate requirements) of all the active subscribers must be met in an efficient and cost effective manner.

Currently the cellular technological advancements proposed by the Third Generation Partnership Project (3GPP) are the most widely deployed. According to research forecasts, mobile data subscribers will reach 1.8 billion in 2014 [1]. 3GPP's recent proposals called Long Term Evolution (LTE) and its subsequent modification called LTE-Advanced are the standards that will take the cellular technology in the 2020s. Scheduling becomes extremely important in determining the overall performance of an LTE system. It is

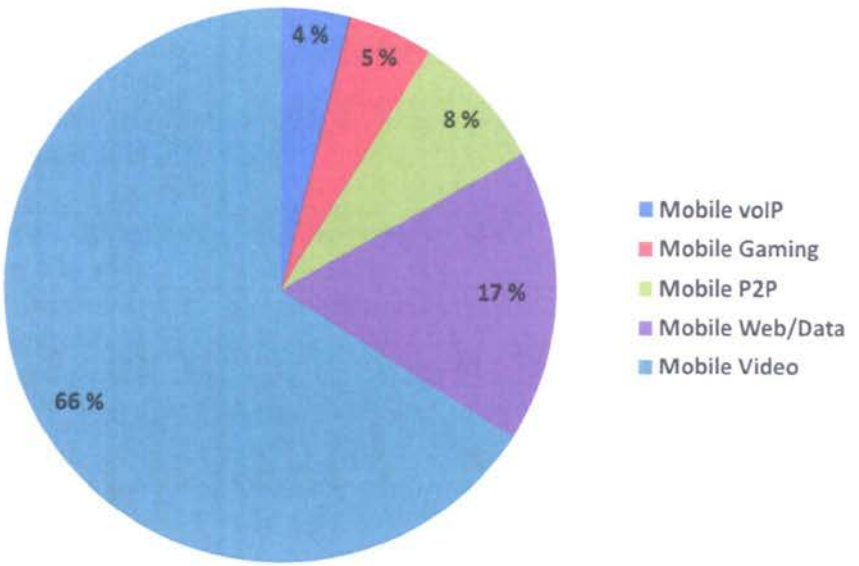


FIGURE 1.1: Mobile data traffic growth forecast in 2014 [1] [2].

the main component which determines which users should be served or assigns resources at any scheduling epoch. Link level (base station to mobile terminal) efficiency largely depends upon the design of the scheduler. An efficient scheduler design must ensure fairness (according to the service needs of each user) in the system by optimally utilizing the amount of radio resources available. The main motivations of the operators would be the end user satisfaction, not merely the amount of system throughput. Scheduling in the LTE standard is more challenging than in earlier standards, mainly because earlier standards were based on single carrier system, (mainly time resources were divided among the users) in contrast to LTE which is a multicarrier system where system resources are shared among users both in terms of time and frequency. Hence not necessarily any scheduling algorithm efficiently designed for a single carrier system performs well in LTE.

1.1 Motivation and Objective

The evolution of mobile networks to high speed, IP-based infrastructure has put onus on the network operators to provide quality services to users running differentiated services on mobile devices. In addition to adding network capacity, the network operators must devise a policy of efficiently partitioning the network resources. The diverse Quality of Service (QoS) requirement of different traffic types is shown in Table 1.1 [5]. According to

TABLE 1.1: Comparison of QoE Expectations and Performance Requirements by Service Type [5].

Services	QoE Expectations	Performance Attributes
Internet	Low – best-effort	Variable bandwidth consumption, Latency and loss tolerant
Enterprise/Business Services	High - critical data	High bandwidth consumption, Highly sensitive to latency, High security
Peer-To-Peer	Low - best effort	Very-high bandwidth consumption, Latency and loss tolerant
Voice	High - Low latency and jitter	Low bandwidth per call, Highly sensitive to latency
Video	High - low jitter and extremely-low packet loss	Very-high bandwidth consumption, Very sensitive to packet loss
Gaming and Interactive	High - low packet loss	Variable bandwidth consumption, Very sensitive to packet loss, Highly sensitive to latency

the table, each service has different QoS requirements in terms of packet delay bound, packet loss rate threshold and bit rate requirements. Furthermore, each traffic type exhibits different bit rate characteristics. For instance, video traffic exhibits variable bitrate characteristic along with the the intensive bandwidth requirements. On the other hand, VoIP traffic exhibits constant bitrate characteristic with low bandwidth needs. Similarly the traffic characteristics of web browsing, FTP, gaming are different. The heterogeneity in different traffic's bitrate and strict QoS requirements has to be dealt with a dynamic policy management. The policy management depends upon the network ability to respond to congestions by gracefully degrading the service quality so that a minimum level of service quality is maintained by the network. It is important to note that adding capacity to the wireless network by increasing the spectrum is an important step which must be taken by the operators to handle the continuous growth of mobile data. However, capacity enhancement by adding more resources is not the answer to this complex problem as the resources are limited and expensive. Any increase in system capacity by increasing the bandwidth would eventually be consumed by bandwidth hungry applications such as video and peer to peer traffic. Furthermore, if an operator satisfies the service needs of the users by adding more bandwidth, then it will be a competitive disadvantage for such an operator against the providers providing the same services by minimizing cost/quality rather than cost/bit. Advanced Radio Resource Management (RRM) procedures are one of the key features of 4G wireless systems which must be exploited by the operators. For instance, one of the features is the maximization of the system throughput by allocating the resources to the most

appropriate users (users with the best channel quality). On the other hand, meeting the service quality requirements of each and every user is also critical. These two objectives are conflicting and there is a risk in achieving one at the expense of the other. Therefore, the trade-off between these objectives has to be addressed according to the policy rules of the operators.

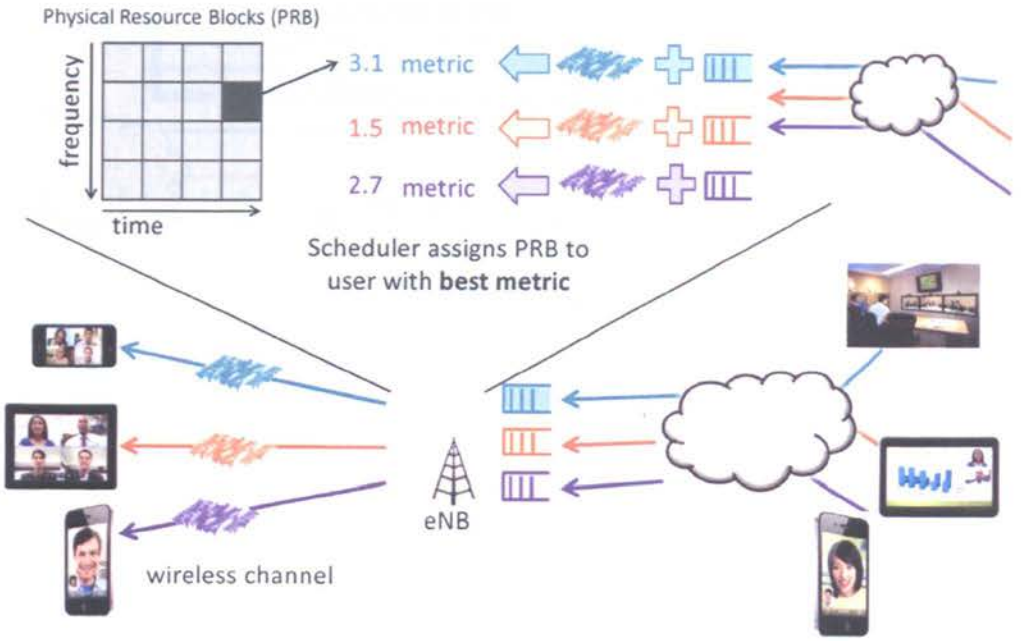


FIGURE 1.2: Scheduler overview. A network with live video conversation and video conferencing flows. The scheduling metric is based on the tight delay bound limit of each packet.

Packet scheduling is one of the most important functions of RRM and plays a key role in distributing radio resources among different users with different service needs. It is one of the key elements in implementing operators' policies. LTE is a multicarrier system where resources are divided into time and frequency domains. The basic time frequency resource unit is called a Physical Resource Block (PRB). Defining a radio frequency allocation (scheduling strategy) on a per-PRB basis, as shown in Figure 1.2, is simpler and easier to implement according to the operator defined policy rules.

An operator's policy can be the prioritization of a particular traffic type w.r.t other traffic types. For instance, Figure 1.2 shows video conference users competing for resources. Operators can reserve a bandwidth at the eNodeB for live video conferencing flows and make sure that there is always enough capacity to support live video conversation. The admission control blocks other traffic types under congestion. On the other hand,

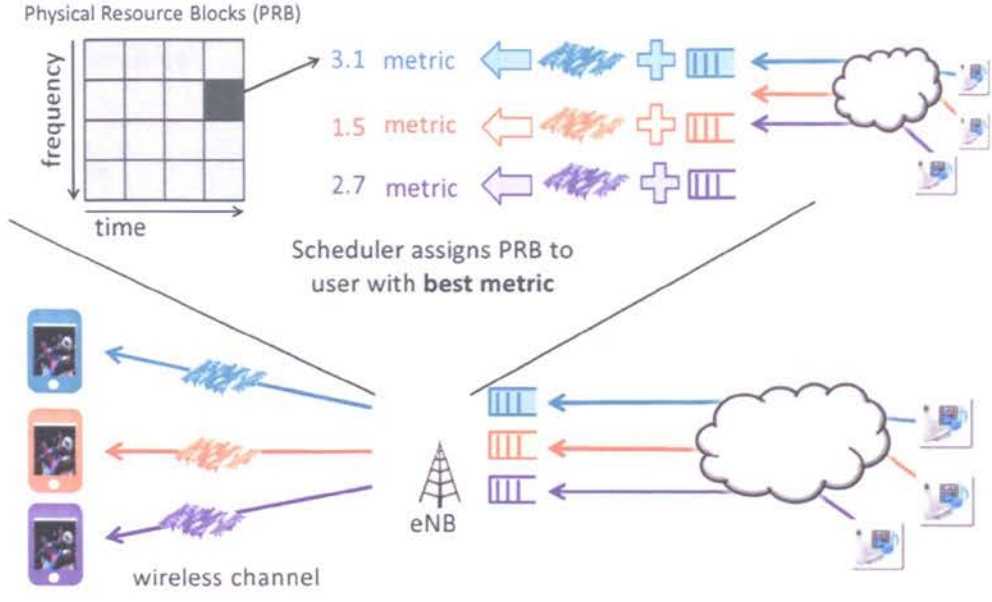


FIGURE 1.3: Scheduler overview. A network with video streaming flows. The scheduling metric is based on the video quality function.

operators can place a packet scheduling algorithm at the MAC layer of eNodeB that guarantees a minimum reserved bandwidth and a maximum delay bound for live video streaming packets. The scheduling metric considers the strict delay bound of each packet. In the event of congestion, the packet scheduler always favors live video conversation flows thus reducing the resource allocation for other traffic types.

In other cases, the operator's policy rule could be a prioritization of video streaming users as shown in Figure 1.3. Video streaming applications are bandwidth demanding, but their data rates can be adapted. Therefore, operators willing to offer quality video streaming to the users must adopt a graceful video quality degradation (adapting the data rate) mechanism to handle the network congestion. Operators must ensure a minimum perceived video quality satisfaction which must be guaranteed even when the network is heavily congested.

It is important to note that best-effort traffic constitutes a significant proportion of mobile traffic as shown in Figure 1.1. Proactive and forward thinking operators would ensure that even under a congested network, some minimum resource reservation must be guaranteed. Traffic classes such as video streaming have an important property of flexible rate adaptation. On the other hand, packets of best-effort traffic have high delay tolerance, *i.e.*, they can reside in the buffer for a longer period of time as compared to

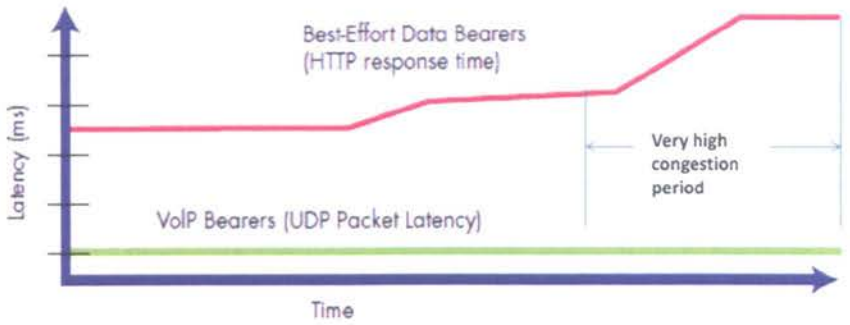


FIGURE 1.4: Packet latency of different traffic types to ensure satisfactory QoE.

other applications such as VoIP, video conferencing and video streaming. The delay tolerant property of best-effort traffic can be exploited by designing a packet scheduling rule which prioritizes packets of real-time traffic while maintaining a minimum resource allocation probability for the best-effort traffic. For instance, Figure 1.4 shows that the packet scheduling function under congestion increases the packet latency of the best-effort traffic while the packet latency of the VoIP traffic remains constant.

1.2 Areas of Contribution

In this thesis, the aforementioned challenges are addressed. Different solutions to this complex and complicated problem are provided. Wireless network operators will likely take an approach split into several phases, starting with the basic QoS aware strategies. In the first phase, the main focus would be on the VoIP and best-effort traffic types, then evolving to accommodate other traffic types according to their business model. Therefore, the proposed solutions fall into the following three categories:

- QoS aware strategies.
- QoE aware strategies.
- Joint QoS and QoE aware strategies.

1.2.1 QoS aware strategies

The main goal of a packet scheduler is to avoid QoS violations in terms of packet delay and loss rate. According to Table 1.1, different traffic classes have different QoS parameters. The scheduler must guarantee a limit on packet's delay budget and loss rate limit according to its QCI (Quality Class Indicator). The QCI is the mechanism which determines the QoS requirement of each flow (also called radio bearer). For instance, Figure 1.4 shows QoS differentiation between delay sensitive and delay tolerant traffic types. According to the figure, the increase in network traffic above the system capacity causes the scheduler to increase the packet delay of the best-effort traffic while maintaining the delay of VoIP traffic at a constant level. It is important to note that best-effort traffic can tolerate higher packet delays and its QoE satisfaction level decreases marginally with the increase in latency thus the scheduler maintains satisfactory service even under network congestion. QoS-aware scheduling strategies are interesting, promising and easier in terms of implementation. However, this type of schedulers require strict admission control as fairness is an inherent feature of the QoS-aware schedulers. Any increase of the incoming traffic above the system capacity would violate the QoS performance of flows already in the network. Therefore with QoS-aware scheduler in place, the admission control policy must limit the incoming traffic according to the system capacity. Video traffic exhibits a highly variable rate (high peak to average rate ratio) characteristics. Consider a scenario where all the video traffic flows are admitted according to average rate requirements and the admission control blocks further flows from entering the network. Under such situation, the network resources may get under utilized (during lower rate periods of video traffic) or may result in QoS violations during the peak traffic periods.

1.2.2 QoE aware strategies

Video traffic contributes a major proportion of network traffic as shown in Figure 1.1. QoE-aware packet scheduling strategies are specifically designed for video streaming traffic. By considering the content of the video traffic, the scheduler aims to maximize the perceived video quality. The information on the contents of different video traffic is provided through cross-layer signaling. Under congestion, the scheduler exploits the redundancy in the video traffic by either dropping the redundant quality layers present in the video stream or sends a feedback signal to the video server to perform timely rate adaptation. These type of schedulers consider different objective functions in the scheduling decision which are based on the video quality. The main goal of the scheduler is to maximize the video quality of the streaming users. When the network is heavily

congested, the scheduler exploits the quality, temporal and spatial redundancy in the video stream, as shown in Figure 1.5, and drops packets having little contribution towards the overall video quality. Different information can be provided to the scheduler such as the decoding deadline of a video frame, packet's dependency on other packets, packet's contribution to the overall video quality.

Optimizing video traffic by designing a separate scheduling function has been gaining importance mainly because most of the mobile network traffic will be video. There are many content aware scheduling strategies designed for video traffic. However, content aware scheduling strategies pose a limitation from an implementation point of view because of the extensive cross layer signaling requirements. Complex application layer information such as distortion associated with each video packet, video frame size and its dependency level, rate and quality of each video layer are required by the scheduling function. Operators intending to use video quality based scheduling strategies must introduce new network elements which can parse a video packet and provide information related to the video streams.

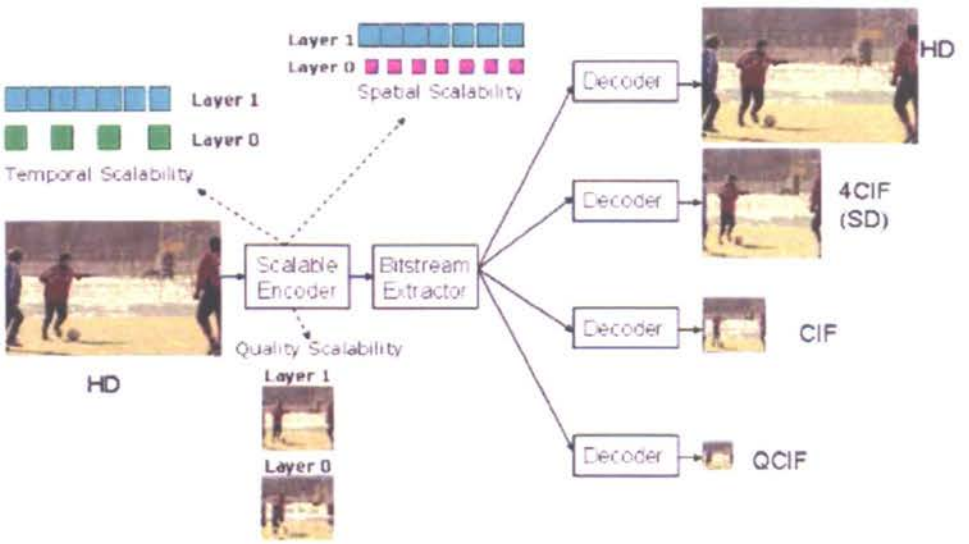


FIGURE 1.5: SVC traffic stream at different scalability level.

1.2.3 Joint QoS and QoE aware strategies

By utilizing the important features of the quality aware strategies discussed above, a novel quality-aware scheduling framework is proposed. In the framework, different traffic types are mapped to different priority classes. The characteristics of Scalable Video

Coding (SVC) traffic, as shown in Figure 1.5, can be exploited by mapping important video layers (video layers giving higher video quality at lower bitrate) to higher priority classes. Similarly, traffic classes with tight delay bound such as video conferencing and VoIP are assigned a higher priority classes. Under congestion, the scheduling function decreases the resource allocation probability of less important priority classes and schedules the most important classes. Packet priority mapping of different traffic classes is based on operators policy rules. The scheduling function depends on packet priority, packets waiting time in the queue and channel quality, thus allowing different traffic types to be served with a common scheduling metric, which is not the case in quality aware strategies discussed above.

The remainder of the chapter presents the thesis structure followed by the description of LTE design goals and protocol architecture.

1.3 Thesis structure

The structure of the thesis is organized as follows:

Chapter 2 presents the performance assessment methodology of scheduling strategies proposed in this thesis. The chapter further discusses state-of-the-art features and their performance over an LTE link layer. The performance is investigated through simulations and complemented by state-of-the-art schedulers performance over an LTE link layer. The chapter also highlights the main issues and important aspects to consider in designing a performance setup for quality aware scheduling over LTE.

The proposed quality-aware scheduling strategies are summarized in Figure 1.6. The remainder of the thesis presents them as follows.

In Chapter 3, an opportunistic packet loss fair strategy for delay sensitive traffic is proposed. The Opportunistic Packet Loss Fair (OPLF) scheduling algorithm is based on a simple dynamic priority function which depends on the Head of Line (HOL) packet delay, the packet loss rate (PLR) and the achievable instantaneous downlink rate of each user. This algorithm overcomes the main limitations of existing algorithms by exploiting multi-user frequency diversity in a novel way thus achieving better performance than state-of-the art algorithms, such as Modified Largest Weighted Delay First (M-LWDF), Proportional Fair (PF), and Packet Loss Fair (PLF), in terms of throughput, PLR and fairness among users.

In Chapter 4, a fair downlink scheduling strategy for different traffic classes is proposed. The goal is to allocate the available resources efficiently, but also to guarantee fairness

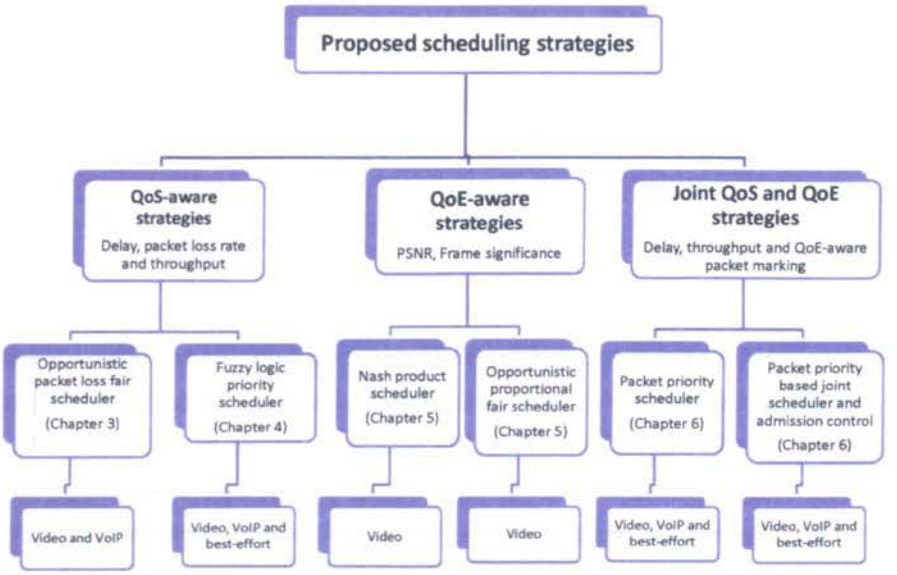


FIGURE 1.6: Tree diagram of the proposed scheduling rules.

among the different users and traffic classes (including real-time and best effort traffic). Existing QoS aware scheduling strategies, such as M-LWDF, Exponential (EXP), Exponential Proportional Fair (EXP-PF) and the Log based scheduling rules, prioritize real-time traffic by considering rules based on the HOL packet delay and the tolerated packet loss rate, whereas they serve best-effort traffic by considering the classical PF rule. These scheduling rules do not prevent resource starvation for non-real-time traffic. On the other side, if both real-time and non real-time traffic are scheduled according to the PF rule, delay sensitive applications suffer from delay-bound violations. In order to fairly distribute the resources among different service classes according to their QoS requirements and channel conditions, the concept of fuzzy logic is employed. Fuzzy logic is ideally suited for problems where a definite mathematical solution is not available. The information about the changes in the radio channel and the traffic rate of each user is uncertain. Fuzzy logic can deal with such situations because of its capability to make approximate reasoning. By employing the fuzzy logic concept, all the traffic classes are served with one priority metric. Simulation results show better intra-class and inter-class fairness than state-of-the-art scheduling rules, without penalizing the system efficiency. The proposed scheduling framework enables to appropriately balance urgency of traffic, system throughput, and fairness.

Chapter 3 and 4 discuss non-content aware strategies, where the service quality of the received video is measured in generic terms of packet delay, packet loss rate or data rate.

In general, these methods exploit the variability of the wireless channel over time and across users, allocating a majority of the available resources to users with good channel quality who can support higher data rates, while maintaining fairness across multiple users. In this context, these strategies utilize a scheduling rule, which is defined as either a function of each user's average throughput, or of each users packet loss rate or delay of the HOL packet.

Video quality, however, is not a simple function of the data rate, delay or data loss but it is rather affected differently by the impact of losses and errors in different segments of the video stream. This is highlighted in an Scalable Video Coding (SVC) bitstream, which consists of one base layer and multiple enhancement layers. As long as the base layer is received, the receiver can decode the video stream. As more enhancement layers are received, the decoded video quality is improved. In multi-user video transmission, this introduces a type of multi-user content diversity that can be exploited by content-aware scheduling policies in optimizing the utilization of the network resource. Therefore by relying on cooperative game theory and in particular on the Nash bargaining solution (NBS), downlink scheduling strategies for scalable video transmission to multiple users is proposed in Chapter 5. In the first part of the Chapter, a novel utility metric based on video quality is introduced which is used in conjunction with a quality-driven scheduler. Results show that the proposed strategy outperforms throughput based strategies and in particular it enables the operator of the mobile system to select the level of fairness for different users in a cell based on its business model. In the second part of the chapter, an Opportunistic Proportional Fair (OPF) scheduling strategy is presented. The priority function on each PRB, considering the time-averaged frame significance and bit throughput, efficiently exploits the multi-user diversity as compared to state-of-the-art strategies. Simulation results confirm the efficient exploitation of time and frequency diversity by showing an improvement in the objective video quality of each user as compared to state-of-the-art strategies.

In Chapter 6, a novel scheduling strategy based on packet priority has been designed. This scheduling strategy targets at offering an appropriate trade-off between efficiency and fairness by considering packet importance in addition to the channel quality of each user, average throughput, and HOL delay.

The proposed strategy reduces the need for cross-layer signaling and frequent end-to-end link probing required by other cross-layer approaches.

The approach is based on the following key features

- Exploitation of application layer packet marking: The emerging SVC standard splits the encoded video stream in different layers of different importance. It is possible to mark packets belonging to different layers with layer information.
- Exploitation of QoE based mapping of SVC layers into priority classes: The mapping can be done at the PDN Gateway (P-GW), thus enabling the mobile operator to perform optimal prioritization, achieving the maximum overall QoE under the constraint of network resources. Video layers achieving maximum video quality (Mean Opinion Score (MOS)) for a given bandwidth are prioritized.
- Opportunistic scheduler: The scheduler exploits QoE-based packet marking and targets at minimizing delay bound violations for the most important priority classes. Exploiting packet marking at the link layer decreases the need for link probing from the base station to the video servers.
- Rate adaptation at the link layer via class-based admission control on the video layers: jointly with the proposed scheduler, this enables appropriate exploitation of the resources when the system is loaded above its capacity.

It should be noted that parts of this thesis have been published in [8] [9] [10] [11] [12].

1.4 LTE design goals and protocol architecture

LTE is a major advancement in terms of physical layer technology in the cellular paradigm with respect to previous releases of 3GPP, with the adoption of OFDM (Orthogonal Frequency Division Multiplexing) as the main new concept in cellular technology; LTE takes benefit from this technology to achieve its design goals, reported in Table 1.2. Some of its salient features include fully switched packet core, only one node in the radio access network, called eNodeB which links the mobile stations, called UE (User Equipment), to the core network which achieves very low latency in user plane as well as control plane. LTE (like previous 3GPP releases) relies heavily on AMC (Adaptive Modulation and Coding) and HARQ (Hybrid Automatic Repeat request) to achieve gains in throughput.

Multimedia applications contribute a major portion of applications in next generation wireless networks. Video streaming is one of the multimedia applications which must be supported efficiently by the LTE standard. QoS (Quality of Service) requirements for video streaming are quite stringent and must be met for all active flows. Hence MAC layer scheduling becomes extremely important in determining the overall performance of an LTE system. Traditional scheduling is divided into channel aware and channel

unaware. Channel aware schedulers take channel information into account and maximize the cell throughput (link efficiency in terms of bits/sec/Hz) [13]. Maximization of cell throughput for real-time applications does not provide a fair system. Efficient scheduling schemes must provide a fair service and all the active flows in the system must get a minimum guaranteed service by efficiently utilizing the spectrum. AMC and HARQ are the basis which provides a framework for an efficient channel aware scheduling strategy. Before highlighting the importance of AMC and HARQ for an efficient scheduling, the characteristics of the different layers in the LTE protocol architecture are shortly summarized below:

The major design goals for 3GPP release 8 are reported in Table 1.2 [6].

TABLE 1.2: LTE design goals [6]

Scalable Bandwidth	1.4, 3, 3.2, 5, 10, 15 and 20 MHz.
Peak Data Rate	Up to 100 Mbps for 20 MHz in downlink for uplink up to 50 Mbps.
Number of transmit and receive antennas for downlink, **Tx * Rx**	4 * 2, 2 * 2, 1 * 2 and 1 * 1.
Number of transmit and receive antennas for uplink, **Tx * Rx**	1 * 2 and 1 * 1.
Spectrum Efficiency	3 to 4 times with respect to High Speed Downlink Packet Access (HSDPA) in Downlink and 2 to 3 times in Uplink.
Latency	Control-plane less than 50-100 ms, User-plane less than 10 ms.
Mobility	High performance at speed up to 120 km/hr and maintain link speed up to 350 km/hr.
Coverage	Optimal performance up to 5 km radius and slight degradation up to 5-30Km, operation up to 100 km may be supported.

1.4.1 Packet Data Convergence Protocol (PDCP)

The main function of this layer is to compress the header in the incoming IP packet [14]; as there is no circuit switching, the entire architecture is packet switched, hence compression (especially for VoIP) becomes very important. Robust Header compression (ROHC) is the protocol used for compression; ROHC can reduce the IP header size from 40 bytes to approximately 1 to 4 bytes. PDCP is also responsible for the security of user plane data, RRC (Radio Resource Control) data and NAS (Non access stratum) data. ROHC is performed before the security operation (encryption, decryption, integrity protection) as ROHC cannot compress encrypted packet.

1.4.2 Radio Link Control (RLC)

In 3GPP release 7, the main RLC functions (segmentation, concatenation, in sequence delivery through Automatic Repeat Request (ARQ)) were performed by a separate node called RNC (Radio network controller). In LTE there is only one node called eNodeB in the radio access architecture, hence RLC is located in eNodeB [15]. Apart from segmenting and concatenating the compressed IP packets, RLC provides reliability through ARQ operation. RLC is operated in two modes: AM (Acknowledged Mode), where RLC requests retransmission of the missing protocol data units (PDUs) from the transmitting entity, this mode is mainly used for TCP-based applications where reliability is more important. For delay sensitive applications (VoIP or video) UM (Unacknowledged Mode) is used in which missing PDUs are not requested for retransmission. It is important to note that almost all the errors due to the dynamic nature of the wireless channel are handled by a much more efficient and fast (ideal for delay sensitive application) retransmission mechanism called HARQ (Hybrid Automatic Repeat Request). Hence a question arises about the retransmission mechanism (ARQ) at RLC. MAC (Medium Access Control) based HARQ can transfer erroneous packets to the RLC in the event of ACK/NAK (HARQ Acknowledge/Non-Acknowledge) corruption, for instance when NAK (due to noise) is interpreted as ACK. The possibility of this event to occur is low, approximately 1% [6], which is however still high when considering that the maximum data rate which LTE can support is 100 Mbps. According to [16] for TCP based applications the probability that a packet may be lost should be less than 10^{-5}. Hence ARQ at RLC becomes extremely important for TCP based traffic. Since RLC and MAC layer resides in one node in LTE, there is a tighter interaction between RLC and HARQ, hence RLC retransmissions are faster in LTE than the previous releases of 3GPP.

1.4.3 MAC (Medium Access Control)

The MAC layer performs HARQ retransmissions, scheduling (both uplink and downlink) [17], and handles control and traffic channels [18]. Traffic channels carry user data. Control channels, also called logical channels, are extremely important for the operation of LTE. From the MAC layer, the RLC layer uses services in the form of logical channels, as reported below.

1.4.3.1 Control channels

Broadcast Control Channel (BCCH): Important system information such as operating bandwidth, cell ID and other important configuration information are transmitted

through this channel. User Equipment (UE) accesses this information before entering the system, eNodeB continuously broadcasts all the configuration information on this channel. When a UE wants to access the system resources, it acquires all the configuration information broadcasted by the eNodeB.

Paging Control Channel (PCCH): In the event of unknown UE location, this control channel is used to page the UE so that cell level location of the UE is known to the network. Paging message is sent to the UE, which returns its cell location to the network.

Dedicated Control Channel: This channel is used to individually configure each UE; information such as handover message is carried over this channel. Hence whenever there is a need for an individual configuration of UE, this control channel is used.

Multicast Control Channel: This channel is used to control multicast information carried over Multicast Traffic channel. Control information related to MBMS (Multimedia Broadcast and Multicast Service) is transmitted over this channel.

Multicast Traffic Channel: All the multicast transmissions (MBMS from eNodeB to UEs) are carried over this channel. MBMS services first introduced in the release 6 of 3GPP which makes it possible to transmit same content across multiple users, thus paving the way for Mobile TV services. MBMS is further divided in to broadcast and multicast services. In broadcast, part of the radio resources in a cell is reserved and all UEs (subscribed to this service) received the transmitted signal. It is important to note that there is no need to track the UEs (as done in unicast transmission) in the RAN (Radio Access Network). A user subscribed to this service simply receives the content without informing the network. On the other hand in multicast service a group is formed called a multicast group, each UE joined this group by notifying the network so that appropriate amount of radio resources can be assigned.

Dedicated Traffic Channel: - Actual user data is transmitted over this channel.

1.4.3.2 Transport channels

In order to offer services to the MAC layer, the physical layer provides transport channels. Transport channels carry user data, and the format and the characteristics with which the data is transmitted over the radio interface is specified by the MAC layer. The MAC (MAC scheduler) layer decides the type of modulation, coding, the size of the transport block (number of used PRBs) and also antenna mapping when MIMO (Multiple Input Multiple Output) mode is used. These characteristics specify the transportation format of the TB (Transport Block) and are called Transport Format. One

TB (two in MIMO mode) is transmitted over the radio interface in each Transmission Time Interval (TTI). For information on TTI in LTE see section 1.5.1.

Broadcast Channel (BCH): It is used to transport information on the BCCH channel. Format with which this information is transmitted is fixed, generally the most robust modulation and coding scheme is used *i.e.* Transport Format is always fixed so that there is a high probability of acquiring the broadcast information over the dynamic channel.

Paging Channel: It is used to transport information on the PCCH channel. In the LTE standard paging time is kept fixed, so that mobile terminals remains in the sleep mode (in the event of no transmission) and wake up only (according to the predefined time) in the event of paging time instant. This feature saves the battery power of the mobile terminal.

Downlink Shared Channel (DL-SCH): Downlink data is transported over this channel. MAC layer specifies the Transport Format according to which modulation scheme, coding rate are specified for each user. All the important features like HARQ, spatial multiplexing and link adaptation are supported by this channel. It is important to note that this channel transmits user data in the unicast mode only, that is broadcast and multicast data is not supported.

Uplink Shared Channel (UL-SCH): Similar to the DL-SCH instead of downlink, all the features are used for uplink.

Multicast Channel: This channel is used to transport multicast and broadcast data within a cell or from one cell to another. All the MBMS data is transported on this channel.

1.4.4 Physical Layer

The main operations of the physical layer include coding, HARQ operation (the physical layer performs the soft combining operation in which redundancy information is increased in each retransmission), modulation and multi antenna processing in case of MIMO (Multiple Input and Multiple Output). The PRB, shown in Figure 1.8, is the basic transmission unit. PRBs are controlled by the MAC layer and these are assigned to different UEs according to certain criteria (such as channel conditions, buffer status of each user). After MAC layer scheduling, a decision is taken about the assignment of PRBs to different UEs. These assigned PRBs are then processed by the physical layer which performs CRC (Cyclic Redundancy Check) operation. The CRC bits are attached to each PRB, then FEC (Forward Error Correction) through Turbo coding

is applied to each PRB. After FEC, modulation is performed. It is important to note that modulation scheme and coding rate is determined by scheduler at the MAC layer. UE receives the transmitted signal, demodulate and decode the signal, and informs the HARQ at the MAC layer about the status of the transmission. In the event of erroneous transmission, the erroneous packet is kept at the HARQ buffer and NAK is sent to the MAC layer scheduler at the eNodeB.

1.5 LTE transmission schemes

Physical layer OFDM and MAC layer OFDMA offer considerable improvements over 3GPP's previous technologies such as UMTS (Universal Mobile Telecommunication System) and High Speed Packet Access (HSPA). The use of OFDM is a novel concept in cellular systems: previously single carrier modulation was employed, which results in a considerable amount of ISI (Inter Symbol Interference). ISI is the result of the delay spread in a signal due to the multipath effect. In the single carrier system if the data rate increases the symbol time decreases. In the event of considerable increase in the data rate one symbol can spill in to the adjacent symbol, thus ISI imposes an upper limit on the data rate in a single carrier system. In the frequency domain, multipath propagation causes different amounts of distortion (phase shift) and the signal arrives at the receiver out of phase which further limits the capacity of the system. In order to combat multipath propagation, channel inversion or rake equalizers are employed, which causes a complex channel equalizer implementation. The complexity exponentially increases above 5 MHz.

Contrary to the single carrier system, OFDM utilizes narrow band subcarriers which results in long symbol duration and thus an increase in the data rate can be achieved by simply increasing the number of subcarriers (increase in the parallel data stream). All the subcarriers are tightly spaced and are orthogonal to adjacent subcarriers; this causes an efficient and flexible use of bandwidth. Another remarkable feature of OFDM is the use of cyclic prefix that further combats the ISI. Cyclic prefix is simply a fixed duration of guard band. The only disadvantage of OFDM is a considerable increase in the signal's peak-to-average power ratio (PAPR). The single carrier system utilizes GMSK (Gaussian minimum shift keying) and PSK (phase shift keying) which produces a constant envelope modulation which results in a linear operation of power amplifiers. In OFDMA larger PAPR causes the power amplifier to be operated in the clipping regions (maximum and minimum amplitude of the signal) which decreases the efficiency of the RFPA (Radio Frequency Power Amplifier). Due to the reduced efficiency, higher RFPA are required which results in more power consumption at the base station. Due to the

limited power capability of the mobile terminal SC-FDMA (Single Carrier-Frequency Domain Multiple Access) is employed. For more information on SC-FDMA refer to [6]. This chapter mainly focuses on LTE downlink, thus the remaining discussion will mainly focus on the LTE downlink transmission.

1.5.1 LTE Frame Architecture

In the LTE standard two frame structures are defined, one for TDD (Time Division Duplexing) referred as frame structure Type 2 and the other used for both TDD and FDD (Frequency Division Duplexing) referred as frame structure Type 1. The total duration of a downlink frame is 10 ms, which comprises of 20 slots of 0.5 ms each, as shown in Figure 1.7. A frame is further divided into 10 subframes; the duration of each subframe is 1 ms, i.e., two slots equal one subframe, which is also called a Transmission Time Interval (TTI). The PRB is a basic time-frequency resource which comprises of a set of 12 contiguous sub-carriers over one slot as shown in Figure 1.8. This is the basic resource unit allocated by the scheduler, i.e., resources are allocated in units of PRB. The number of PRBs available depends upon the system bandwidth, ranging from 6 in 1.4 MHz up to 100 in 20 MHz bandwidth.

There are seven OFDM symbols in one slot, with each symbol is separated by a guard interval called cyclic prefix (normal cyclic prefix); this makes a total of 84 resource elements (7 OFDM symbols $\times$ 12 sub-carriers). When the extended cyclic prefix is used there are six OFDM symbols per slot, which makes a total of 72 resource elements as compared to 84 resource elements with the normal cyclic prefix. Therefore, the extended cyclic prefix has more overheads (large guard intervals) but such mode is more robust in an environment where delay spread is an issue. In the remainder of this thesis, the normal cyclic prefix mode is considered and frame structure Type 1 (FDD duplexing) is used. The remainder of this chapter presents the basic state-of-the-art scheduling strategies.

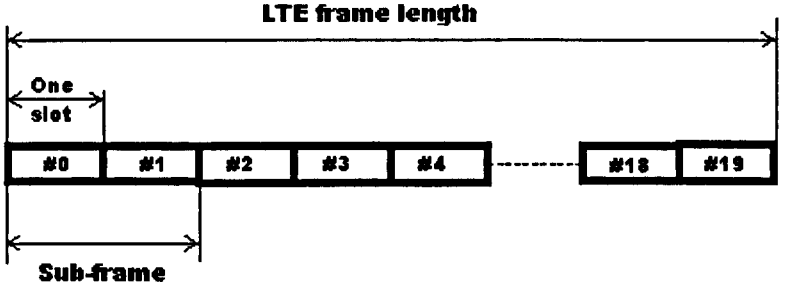


FIGURE 1.7: LTE radio frame length.

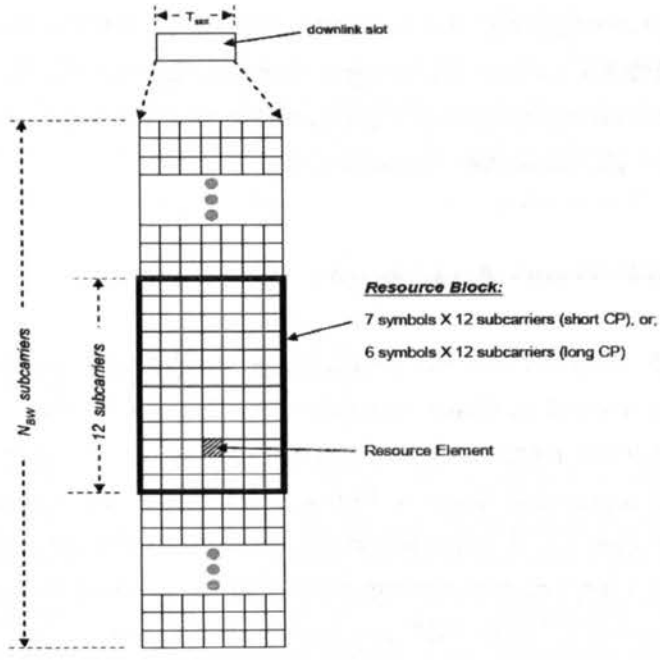


FIGURE 1.8: Basic PRB architecture.

1.6 Channel dependent scheduling in LTE

The time-frequency resource is dynamically assigned among the users in LTE, whereas in HSPA the resources are divided in terms of time and channelization code. In shared channel transmission the LTE scheduler plays a key role in dividing the resources among the users. The scheduler decides which user will get what amount of resources (PRBs). Scheduling decisions are taken at 1 ms intervals. By exploiting multiuser channel diversity, considerable gain can be achieved, hence the throughput achieved by the scheduler can be greatly increased if the scheduler takes channel conditions into account. The same concept was employed in HSPA where the scheduler selects the users whose channel conditions are better, thus maximizing the cell throughput. Because of the OFDMA technology, LTE utilizes not only the time domain variations but also the frequency domain variations. Each UE periodically or aperiodically (depending on the system configuration) sends instantaneous channel quality information to the base station (eNodeB). Based on this channel quality information, the scheduler allocates the appropriate PRBs which face less attenuation for each UE. There are different feedback granularities in the standard, for instance the UE can send just a single Channel Quality Indicator or a separate CQI for each PRB. The scheduler then carries out the averaging, *i.e.*, it maps the signal to interference noise ratio experienced by the PRBs into one (equivalent) Additive White Gaussian Noise (AWGN) SNR. One of the mostly used averaging methods is the MIESM (Mutual Information Effective SNR Mapping) [19] [20]. After resource

allocation, the scheduler uses the MIESM method to calculate a single CQI value to be used for all the allocated PRBs of a particular UE.

After a short description of the LTE protocol architecture, Chapter 2 discusses the performance assessment methodology in analyzing the performance of novel scheduling strategies proposed in Chapters 3, 4, 5 and 6.

Chapter 2

Scheduling for LTE wireless systems - Performance Assessment Methodology

The performance or dependability analysis of modern complex systems is a huge challenge which can be addressed either via analysis or via simulation. With analytical methods, the reader can easily understand and verify the proof of the analytical numeric method employed in solving the research problem. The performance analysis using the analytical method requires a mathematical model. However, the application of the analytical approach becomes inapplicable with the increase in size and complexity of the model. In order to cope with the the dynamic nature and size of the model, approximation methods (approximating an analytical numeric method) are employed. The use of approximation methods becomes inaccurate for large complex systems having complex dynamic properties.

The system model for radio resource allocation includes several complex sub-models which change both in the time and frequency domains, for instance the impact of the channel model, exhibiting fading and multi-path propagation phenomena which affect the radio resources both in the time and frequency domains. The probabilistic nature of the incoming traffic, coupled with the multi-user downlink scenario with dynamic channel quality, pose a huge challenge in the mathematical modeling of such systems. Such models can be approximated by using a restrictive assumptions on the arrival process of the traffic and changes in channel quality over time. It is important to note that the contribution of this thesis includes novel QoS, QoE and joint QoS and QoE aware scheduling algorithms under different scenarios achieving different goals. Therefore, developing analytical numerical models for each of the underlying scenarios with different

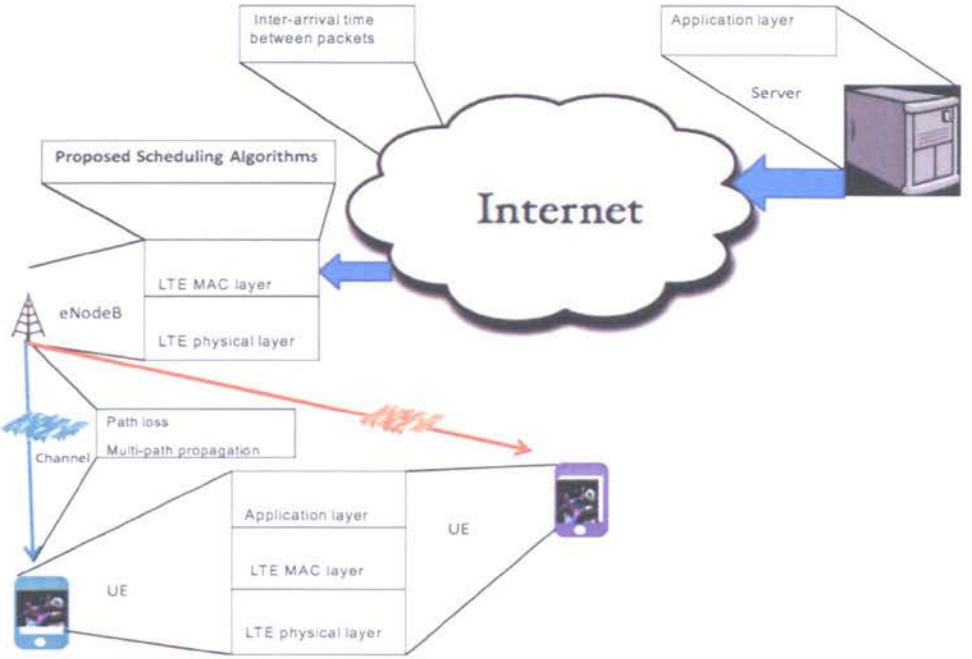


FIGURE 2.1: Performance assessment setup used in the thesis for the analysis of novel scheduling strategies proposed in Chapters 3, 4, 5 and 6.

goals would require restrictive assumptions and may result in inaccuracies. The use of the simulation methodology in analyzing the performance of non-linear complex systems with diverse goals is becoming increasingly useful. In order to analyze the performance of the novel scheduling strategies proposed in this thesis, the simulation methodology is used as the main tool. The proposed performance assessment setup is shown in Figure 2.1. According to the figure, the important entities used in the assessment setup are:

- Wireless channel: The wireless channel is an important entity as all the impairments on radio resources are due to path-loss, multi-path propagation and other fading phenomena.
- LTE physical layer: Wireless simulators evaluating the radio resource management algorithms generally utilize a system level simulator in which the physical layer of the wireless network is used as an abstraction. On the other hand, a link level simulator has a complete set of physical layer procedures implemented. In this work, all the proposed scheduling strategies are implemented on top of the complete physical layer procedure of an LTE system.
- LTE MAC layer: The novel scheduling algorithms proposed in this thesis are developed at the MAC layer (at the eNodeB) of an LTE system. The MAC layer is responsible for assigning radio resources to different users in the network.

- Application layer: The proposed scheduling strategies deal with cross-layer (mainly for video traffic) working of the application, MAC and physical layers. Therefore, the application layer plays an important role in the performance assessment setup. It provides important information about the quality and rate characteristics of the video stream.

The remainder of the chapter discuses in detail the design of the performance assessment setup. Section 2.1 discusses the link layer implementation in a simulation platform utilized in this thesis. Section 2.2 presents some of the basic simulation scenarios from the link layer of the simulator. Section 2.3 presents the complete design of the proposed set up, shown in Figure 2.1, consisting of newly built blocks for the performance analysis of novel quality aware scheduling strategies designed in the subsequent chapters.

2.1 Simulation Platform

There are several simulation platforms available for simulating wireless communication systems, such as OMNET++, OPNET, NS-2 and MATLAB. Considering the LTE physical and MAC layer features, all the built in libraries and features of all the simulators are studied in detail. All the LTE physical layers features such as HARQ soft combining, Turbo coding, AMC and CRC (Cyclic Redundancy Check) are implemented in the C++ based simulator [4] in MATLAB. The simulator provides physical layer features such as FFT (Fast Fourier Transform) and IFFT (Inverse Fast Fourier Transform). In order to investigate the performance of the scheduling algorithms, a link-level simulator built on MATLAB's object oriented features is selected as the simulation platform with all the basic features of an LTE physical layer. In the subsequent sections, the MATLAB based link level simulation environment for LTE is discussed.

2.1.1 Basic building block of the simulator

The basic building blocks of the simulator are given in Figure 2.2. According to the figure, the transmitter (eNodeB) and the receiver (UE) are linked by the channel model. The downlink data from transmitter to the receiver is passed through the channel model whereas the signaling and all feedbacks (Channel Quality Indicator (CQI) and ACK-/NAK) are assumed to be error free. This assumption is realistic as the most robust modulation and coding scheme is employed for the protection of all the signaling and feedbacks. The signaling from eNodeB to the UE includes information such as the type of modulation, coding (FEC), HARQ process id, scheduling and precoding parameters.

The signaling from UE to the eNodeB includes Channel state and ACK/NAK Information. The channel state information is further divided into CQI, Precoding Matrix Indicator (PMI) and Rank Indicator (RI) [21]. The link level simulator allows for an adaptive as well as fixed modulation and coding scheme (MCS).

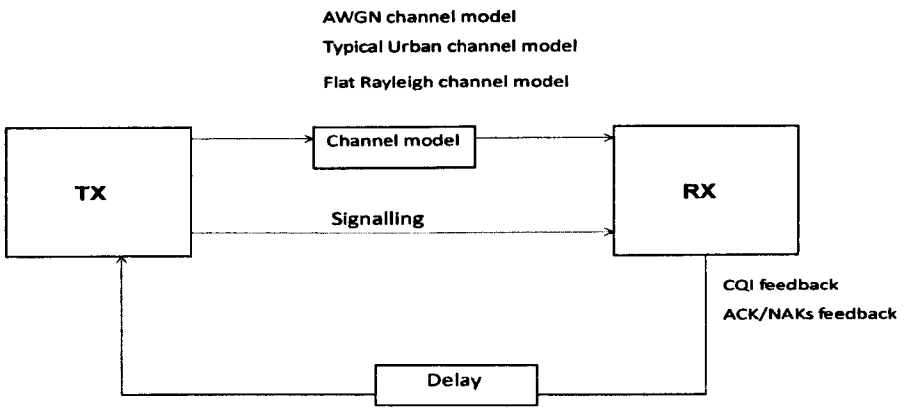


FIGURE 2.2: Basic building blocks of the link level simulator [3].

Table 2.1 reports all the important parameters of the simulator. Different scenarios can be simulated by simply changing the parameter type. For instance, when the parameter related to the number of UEs is set to 1 then the single user downlink scenario is simulated. In order to simulate a multi-user scenario, multiple instances of the UE class is created. The simulator also allows for a static as well as dynamic setting of the modulation and coding scheme. The static MCS scheme can be used to evaluate the throughput performance of a particular MCS scheme over different SNR ranges. On the other hand, dynamic MCS utilizes the modulation and coding scheme depending upon the CQI feedback. The parameters given in the table can be changed to test novel algorithms under different scenarios which is not the case in analytical numerical methods. For instance, a numerical model used for a single user system cannot be directly utilized for a multi-user scenario. Simulation methodology gives flexibility for performance comparison of different algorithms under different scenarios.

TABLE 2.1: Table of important simulation parameters.

Parameters	Typical values
Number of base stations to simulate	1
Number of user equipments to simulate	1
Operating bandwidth (in Hz), allowed values are 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, 20MHz	1.4e6, 3e6, 5e6 or 10e6
1: Single Antenna, 2: Transmit Diversity, 3: Open Loop Spatial Multiplexing 4: Closed Loop SM	1: Single Antenna
Number of transmit antennas at eNodeB	1
Number of receive antennas at UE	1
Channel's fading model. Typical values are fast and Block fading model	BlockFading
Channel model	Typical Urban (TU)
Scheduler type. Typical options are proportional fair and best cqi scheduler	best cqi
Static or Dynamic: whether the scheduler will statically or dynamically assign CQIs.	dynamic
Adaptive modulation coding (AMC): Whether AMC is utilized or not	set
Maximum number of HARQ retransmissions, not including the original transmission. 0, 1, 2 or 3.	3
Normal or Extended cyclic prefix	normal
Delay the uplink channel will introduce (in TTIs)	1

2.1.2 Transmitter structure

The block diagram of the transmitter is shown in Figure 2.3. The design of the transmitter follows the standard proposed in [22][23][24]. PRBs assigned to a UE depend upon the scheduling decisions and the number of bits scheduled depends upon the number of PRBs assigned. CQI feedback plays an important role in determining the number of transmitted data bits to each of the UEs. In the standard, UEs report their instantaneous channel quality by means of a quantized feedback called CQI. There are different feedback granularities in the standard, for instance the user can send just a single CQI for the whole bandwidth or a separate CQI for each PRB. If according to the scheduling decision more than one PRB, with different CQI values on each PRB, are assigned to a user then it is necessary to calculate an average supported CQI value. The scheduler carries out the averaging, *i.e.*, it maps the Signal to Interference Noise Ratio (SINR) experienced by the allocated PRBs into one (equivalent) AWGN Signal to Noise Ratio (SNR). There are two averaging methods supported by the simulator which are Mutual Information Effective SNR Mapping (MIESM) and Exponential Effective SINR Mapping (EESM) [25] [26] [19]. After resource allocation, the scheduler uses either the

MIESM or the EESM (depending up on the status of the LTE_params variable in the Load parameter file) method to calculate the number of data bits allowed for each of the UEs.

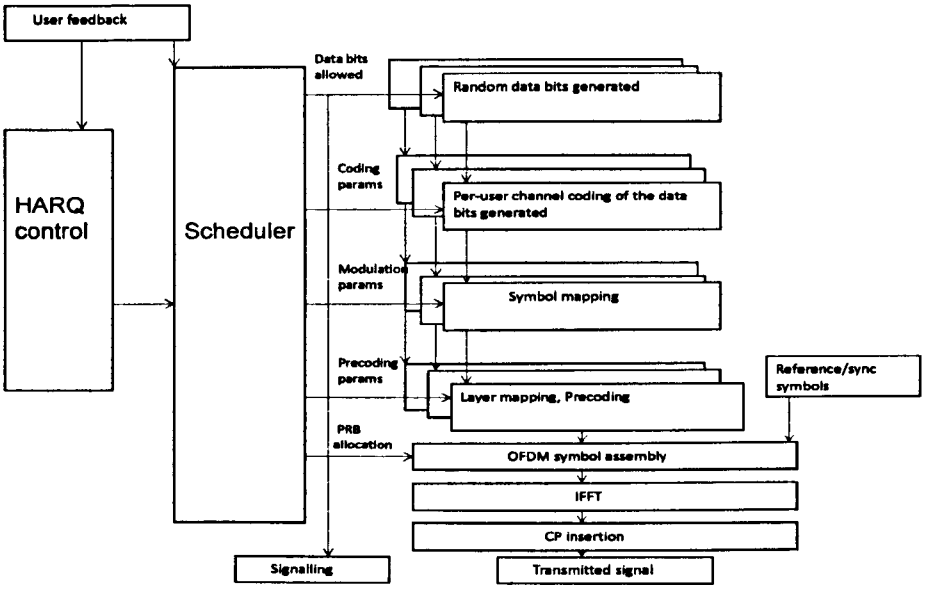


FIGURE 2.3: Transmitter structure of the link level simulator [3].

After a single CQI value to be used for all the allocated PRBs of a scheduled user is calculated, the next step is the evaluation of the modulation and coding scheme associated with the CQI index. Table 2.2 shows all the CQI values used in the LTE standard. Each CQI value corresponds to a particular coding rate and modulation scheme. CQI '1' has the most robust set of coding and modulation used by AMC (Adaptive Modulation and Coding) when channel quality is extremely poor. On the other hand CQI '15', associated to the least robust modulation and coding set, is used when channel conditions are at its best so that the maximum rate can be achieved. The spectral efficiency is 0.1523 bits/s/Hz when the CQI index is one. In order to transmit 78 data bits (for CQI index of 1), the total number of transmitted bits are 1024 as shown in the table (the coding rate is $\frac{78}{1024} = 0.07617$). The associated modulation scheme with the CQI index of 1 is Quadrature Phase-Shift Keying (QPSK) which provides 2 bits per symbol (16-Quadrature Amplitude Modulation (QAM) provides 4 bits per symbol). Therefore, the spectral efficiency for a CQI index of 1 is 0.07617 * 2 = 0.1523. Similarly the spectral efficiency for other CQI values are given in Table 2.2. After the modulation and coding step, the multiple antenna process is applied which comprises a layer mapping and precoding steps. The output signal of the IFFT block is fed to

the Cyclic Prefix (CP) insertion block. The main function of the cyclic insertion is to avoid Inter Symbol Interference. If the channel impulse response is greater than the symbol duration then the effect of the previous symbol can be avoided by the removal of CP at the receiver. According to the standard, the simulator allows for two types of CP. The normal CP consists of 6 OFDM symbols in a 0.5 ms slot. On the other hand extended CP comprises 7 OFDM symbols in a slot, thus resulting in a reduction in the data payload and an increase in the guard interval. The normal mode is useful for small and medium area cell simulations whereas the extended mode is useful for simulating larger areas having extreme time dispersion.

TABLE 2.2: - LTE Channel Quality Indicators.

CQI	MODULATION	Coding rate * 1024	Efficiency
1	QPSK	78	0.1523
2	QPSK	120	0.2344
3	QPSK	193	0.3770
4	QPSK	308	0.6016
5	QPSK	449	0.8770
6	QPSK	602	1.1758
7	16QAM	378	1.4766
8	16QAM	490	1.9141
9	16QAM	616	2.4063
10	64QAM	466	2.7305
11	64QAM	567	3.3223
12	64QAM	666	3.9023
13	64QAM	772	4.5234
14	64QAM	873	5.1152
15	64QAM	948	5.5547

2.1.3 Receiver structure

The design of the UE receiver is shown in Figure 2.4. The information about the PRB allocation and type of MCS to use for disassembling them is signaled by the eNodeB. After disassemble of the PRBs, the detection algorithm is carried out by using either the Zero Forcing or the Linear Minimum Mean Square Error or the Soft Sphere Decoding strategy. Data bits are produced by decoding the detected soft bits. If the decoded data bits results in an error then the NAK signal is generated by the receiver. On the other hand, if the resulted data bits are not in error then the ACK signal is generated by the receiver. It is important to note that the LTE standard requires channel feedback information for the the scheduler at the eNodeB so that the instantaneous channel conditions can be exploited. Therefore, the UE must feedback the CQI, RI and the PMI information to the eNodeB. Therefore, the channel quality feedbacks along with the

ACK/NAK feedback are transmitted by the UE and received by the eNodeB after the specified delay. In order to simulate a multi-user downlink scenario, the object oriented programming features are utilized according to which multiple instances of the receiver are created. Each object of the receiver is characterized by a specific signal-to-noise ratio thus allowing each UE to have its own channel characteristics. Furthermore, the feedback signals from receiver to the transmitter are characterized by a specific delay. The amount of delay is set according to the simulation scenario. For instance, if the uplink delay parameter is set to 1 then CQI and ACK/NAK feedbacks are available after every TTI.

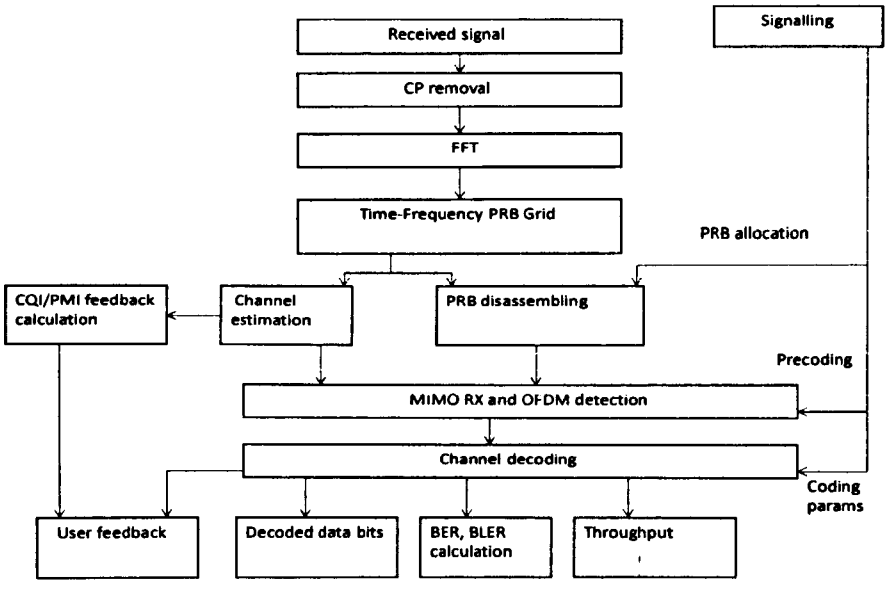


FIGURE 2.4: Receiver structure of the link level simulator [3].

2.1.4 Channel model

The simulator support both the AWGN and the Typical Urban (TU) channel model. AWGN channel model includes linear impairment of white noise which follows a Gaussian distribution of amplitude. It is the simplest form of a channel model which does not consider frequency selective fading, interference and non-linear dispersion due to reflection and refraction. This channel model is more suited to satellite and deep sea communications. Typical urban is also a statistical channel model which takes into account the radio signals fading in a heavily built-up urban environments. This channel model is more suited for tropospheric and ionospheric signal propagation. The channel

impairment in TU model follows a Rayleigh distribution and it is most applicable for environments having no dominant propagation along line of sight between the transmitter and receiver. The channel model requires the average SNR for each of the UEs thus allowing the execution of simulation scenarios having diverse channel quality UEs. The simulator does not provide the path loss impact, therefore the path loss model is incorporated into the simulator by providing output of a newly designed link budget class as the input to the channel model class as shown in Figure 2.5. The HATA model [27] [28] for urban areas is used as the path loss model for radio frequency propagation. Each UE is associated with a particular distance from the eNodeB. The HATA model for urban areas is formulated as:

$$\mathrm{SNR}_U = 69.55 + 26.16 \log_{10} f - 13.82 \log_{10} h_B - C_H + [44.9 - 6.55 \log_{10} h_B] \log_{10} d \quad (2.1)$$

for small or medium sized cities:

$$C_H = 0.8 + (1.1 \log_{10} f - 0.7) h_M - 1.56 \log_{10} f \quad (2.2)$$

and for large cities:

$$C_H = \begin{cases} 8.29\ (\log_{10}(1.54\ h_M))^2 - 1.1, & \text{if } 150 \leq f \leq 200 \\ 3.2\ (\log_{10}(11.75\ h_M))^2 - 4.97, & \text{if } 200 \leq f \leq 1500 \end{cases} \quad (2.3)$$

where

- SNR_U = Signal to noise ratio in dB.
- h_B = Height of an eNodeB antenna, 32 m.
- h_M = Height of the UE antenna, in meters, 1.5 m.
- f = Transmission frequency, 2110 MHz.
- d = Distance between eNodeB and UE.
- C_H = Antenna height correction factor.

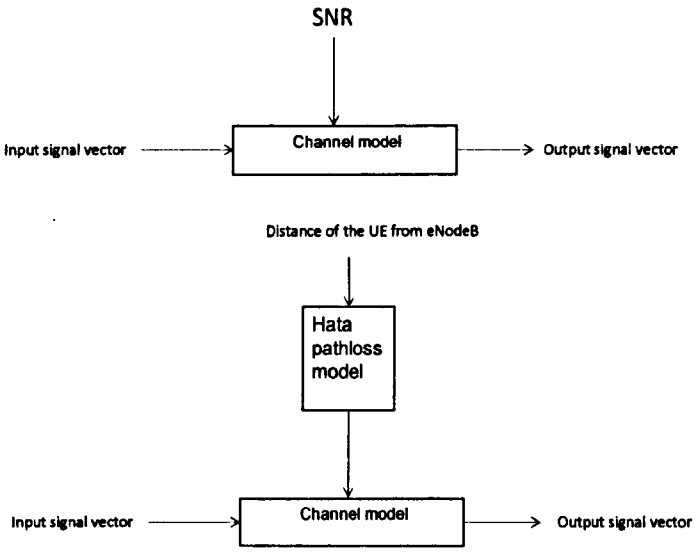


FIGURE 2.5: Channel model of the link level simulator [4]. Integration of path-loss model in the link level simulator.

2.2 Simulations scenarios

By utilizing the object oriented features of the simulator, two types of simulations are utilized in this thesis. Each simulation type results in different computation complexity. The first type is a single downlink setting which covers the link between one eNodeB and one UE. Single downlink set-up allows for the investigation of AMC and feedback optimization, HARQ performance analysis, physical layer modeling for system level simulations [29], channel encoding and decoding modeling [30]. The second type of simulation is a single cell multi-user downlink setting which covers the links between an eNodeB and multiple UEs. This set-up allows for the investigation of multi-user resource allocation strategies [31] [32] which enables the study of multi-user rate regions. The work done in this thesis mainly utilizes the second type of simulation which deals with the multi-user resource allocation problems in a single cell scenario, whereas the first type of simulation studies the performance of HARQ and AMC. It is important to note that the code of the transmitter, receiver and channel model is written using the object oriented features, thus allowing a readable and maintainable code. The code can be easily adapted to the quality aware multi-user resource allocation and scheduling. The subsequent sections presents the basic single and multi-user downlink scenarios.

2.2.1 Simulation of single downlink scenarios

The dynamic nature of the wireless channel causes highly unpredictable variations in the errors. Link adaptation which incorporates Adaptive Modulation and Coding has been effective in improving the link efficiency. However, receiver's noise, unpredictable variations in the interference and fast fading effect cannot be combated. Hence all wireless communication systems incorporate FEC (Forward Error Correction). The increased redundancy in the transmitted signal causes errors to be detected and subset of all errors to be corrected at the receiver. Apart from FEC, another method to combat wireless transmission errors is the use of ARQ (Automatic Repeat Request) technique. Error detecting codes (typically CRC) are applied at the transmitter; the receiver detects the erroneous packets and asks the transmitter to retransmit the erroneous packets. HARQ utilizes both FEC and ARQ. In this section, some basic single downlink scenarios are simulated, thus allowing for the study of HARQ and AMC performance which are very important in designing efficient, in terms of bits/sec/Hz, scheduling strategies.

In order to analyze the performance of HARQ, a single downlink scenario is simulated. The number of UEs to simulate is set to 1 therefore, only a single instance of a UE is simulated. The parameter related to the operating bandwidth is set to 1.4e6. The bandwidth parameter does not play an important role since the main goal of the simulations is to analyze the link performance with and without HARQ. Since it is a single user downlink scenario therefore all the parameters related to multi-user downlink scheduling does not have any impact. Single Input Single Output (SISO) mode (1 Tx and 1 Rx antennas) is utilized in all the simulations. In order to simulate the channel model between the UE and eNodeB, Typical Urban channel is utilized. It is important to note that the channel model requires an SNR value for the link between the UE and eNodeB. In the first set of simulations, HARQ retransmission parameter is set to 0 thus indicating that a packet is dropped when a NAK is received. There are six set of simulations under SNR values of 0, 5, 10, 15, 20 and 25 dB. The second set of simulations are repeated with the HARQ parameter set to 3. Under this scenario when a packet is unsuccessfully decoded at the UE, the packet is not dropped at the eNodeB and is resent. The packet is stored at the buffer until the fifth NAK (unsuccessful decoding of the fourth retransmission) is received. In such an event, the packet is considered as lost if ARQ is not utilized at the RLC.

In HARQ a method called soft combining is used. In soft combining an erroneously received packet is not discarded but retained in the receiver's buffer, and then combined with the next retransmission, hence the error correcting capability is further increased in the next retransmission. LTE HARQ is based on soft combining. Soft combining (handled by physical layer) is further divided into Chase combining and incremental

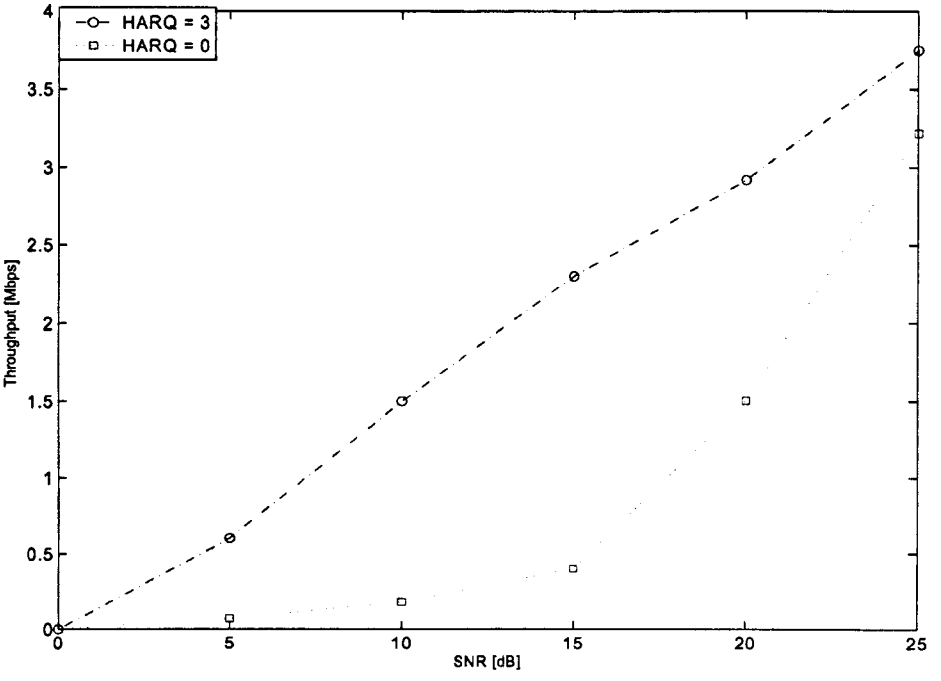


FIGURE 2.6: Throughput (Mbps) achieved with different SNR values for HARQ with maximum = 0 and 3 retransmissions.

redundancy. In Chase combining packets with the same set of coded bits, or in other words the same version of the packet, is retransmitted, whereas in incremental redundancy different sets of coded bits are used in each retransmission. Figure 2.6 shows the gain achieved in throughput without HARQ retransmissions and when HARQ with incremental redundancy (maximum 3 retransmissions) is applied. Only one UE (full buffer condition) with and without HARQ is simulated, and the throughput achieved is compared. According to the figure, a considerable improvement in throughput is achieved when HARQ is employed under all the SNR values. Therefore, HARQ plays an important role in improving the link efficiency in terms of bits/sec/Hz.

AMC and HARQ are the main features which produce an efficient and reliable link level system in LTE. In order to analyze the performance of AMC, three set of single downlink scenarios are simulated. The simulation parameters utilized in the HARQ are kept same with the HARQ retransmission set to 3. In the first set of simulations, parameter related to AMC is set to dynamic over all the SNR values whereas for the other two set of simulations, the parameter utilizes a static CQI values of 10 and 15 corresponding a fixed modulation and coding. Figure 2.7 shows the gain achieved in terms of throughput with AMC, i.e. dynamic CQI (Channel Quality Indicator) assignment, when the SNR of the channel is increased from 1 to 25 dB. The scheduler dynamically assigns the modulation and coding scheme according to the channel quality reports from the UE. Figure

2.7 compares the throughput achieved when the assignment is static (corresponding to CQI=10 and CQI=15) and when CQI is assigned dynamically. Simulations of the single downlink scenarios shows the importance of HARQ and AMC in designing a link efficient multi-user downlink scheduling strategies. Therefore, HARQ and AMC will be the basis upon which multi-user quality aware scheduling strategies are designed.

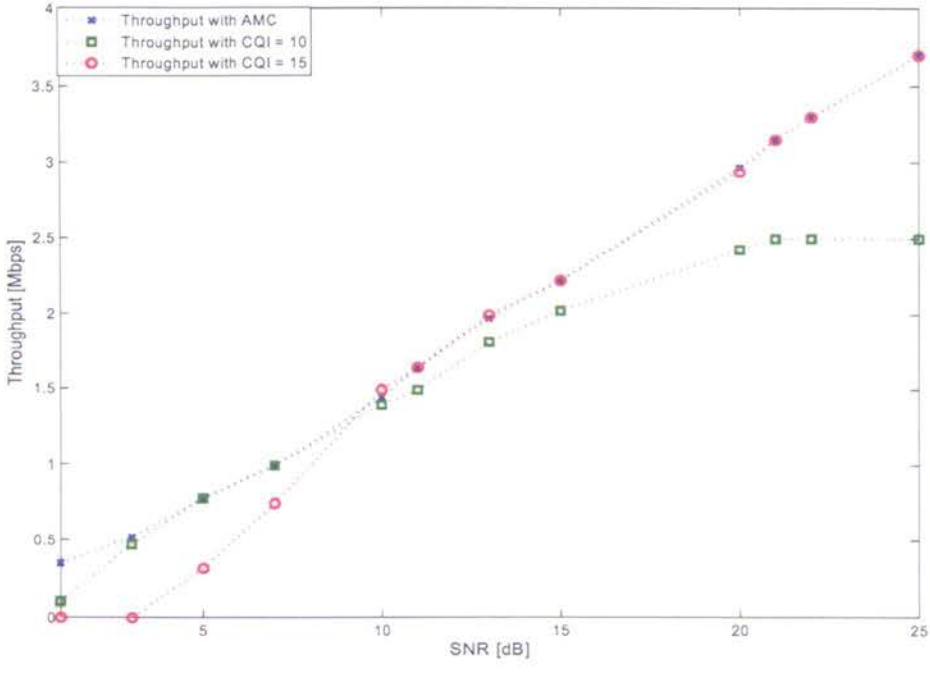


FIGURE 2.7: AMC performance as compared to static CQI assignment.

2.2.2 Simulation of multi-user downlink scenarios

In this section, the performance of some of the basic scheduling strategies are analyzed by designing a multi-user scenario. The parameter related to the number of UEs is set to 6 with each UE having a SNR value. Specifically, the SNR values of 4, 8, 12, 14, 18 and 22 dB are assigned to the UEs. The main goal of the simulation is to analyze the fairness and efficiency performance of the scheduling rules. Different SNR values assigned to each of the UEs correspond different channel conditions, hence the performance of the scheduler for different channel qualities for each UE is studied. The link level simulator includes some of the basic scheduling strategies which are given below:

2.2.2.1 Round Robin scheduler

The Round Robin scheduler [33] is the simplest scheduling algorithm; this scheduler assigns all the system resources to a flow in a round robin fashion. For example if there are 6 PRBs in the system (considering 1.4 MHz bandwidth), all the PRBs are assigned to one UE with appropriate AMC according to the CQI feedback for a period of 1 TTI. Time is shared in equal intervals among all the users. For wireless communication this basic scheduler is ineffective, since it neither provides higher system throughput nor it provides fairness. UE with a low CQI will have an extremely low throughput since the total number of bits transmitted in the corresponding TTI is extremely small because of the necessity to use a robust AMC mode due to the low CQI value. On the other hand UE with better channel quality utilizes the channel only for 1 TTI and then its turn will come after serving all the UEs in the cell.

2.2.2.2 Best CQI scheduler

The most common channel dependent scheduler is the Best CQI scheduler [32]. This scheduler exploits channel variations between users to achieve maximum cell throughput: the greater the difference in the channel quality between users, the greater is the gain in throughput with respect to channel-unaware schedulers. However, users experiencing bad channel quality may not be served. For example users at the cell edges or users experiencing deep fades may not be scheduled compared to users experiencing good channel quality. Thus the user i maximizing the rate $R_{i,\varphi}^{(n)}$ on a PRB (φ) will be assigned the PRB at scheduling epoch n.

2.2.2.3 Proportional fair scheduler

A Proportional Fair Scheduler [34] [35] schedules users based on the following criterion: $i^*(n) = \arg\max \frac{R_{i,\varphi}^{(n)}}{R_{(i,\text{ave})}^{(n)}}$, where $i^*(n)$ represents the user index to be served at the scheduling instant n, $R_{(i,\text{ave})}^{(n)}$ is the average rate achieved over a moving average window. PF compromises cell throughput so that a minimum amount of fairness can be achieved among the different competing flows. UE's having good channel quality can still maximize throughput but not at the expense of depleting users with bad channel quality. Hence there is a trade off between maximum system throughput and fairness among different competing flows.

2.2.2.4 MAXMIN scheduler

The target of a MaxMin scheduler is to maximize the minimum of the users' throughputs [36]. This scheduling strategy is not throughput optimal but maintains a very high level of fairness in terms of throughput: this scheduler is Pareto optimal, i.e., the rate of one UE cannot be increased without decreasing the rate of another UE with a lower rate than the one considered.

2.2.2.5 Resource fair scheduler

The resource fair scheduler evenly distributes the resources among all the competing flows [32]. For instance in 1.4 MHz bandwidth, 6 PRBs are available in each scheduling instant (1 ms), hence if there are 3 UEs competing for accessing radio resources, a resource fair scheduler will assign 2 PRBs to each of the UEs. RF also tries to maximize the throughput by maximizing the sum rate of all the users. This scheduler ensures minimum fairness through fair distribution of PRBs, *i.e.*, $(\frac{N}{I})$ where N is the number of PRBs available and$I$ is the number of UEs. When the number of PRBs is not an integer multiple of the number of UEs some UEs will be assigned $\lfloor \frac{N}{I} \rfloor$ and others $\lceil \frac{N}{I} \rceil$ in order to make up the available resources. In order to ensure fairness a UE assigned $\lfloor \frac{N}{I} \rfloor$ in a current TTI will get $\lceil \frac{N}{I} \rceil$ in the next TTI through uniform randomization.

2.2.2.6 Throughput vs. Fairness

In order to analyze the performance of each scheduler a full buffer scenario is simulated. For the proportional fair scheduler an averaging window of 100 subframes (100 ms) is taken, i.e., an average rate of 100 ms is taken into account. For the Best CQI schedulers if more than one UE have the same channel quality then PRBs are assigned randomly among the users having same CQI. Figure 2.8 reports the throughput for different users, each characterised by a different value of signal-to-noise ratio, for different scheduling strategies. It is clear from Figure 2.8 how efficiently the Best CQI scheduler utilizes channel variations and produces maximum cell (system) throughput, as also shown in Figure 2.9. The MAX-MIN scheduler produces the minimum cell throughput but achieves a fairer system in terms of throughput as shown in Figure 2.8 and in Figure 2.10.

In order to measure fairness of a systems, the Jain's fairness index [37] can be used. It is calculated as shown below:

$$J = \frac{\left(\sum_{i=1}^{I} T(i)\right)^2}{I \sum_{i=1}^{I} \left(T\left(i\right)\right)^2} \tag{2.4}$$

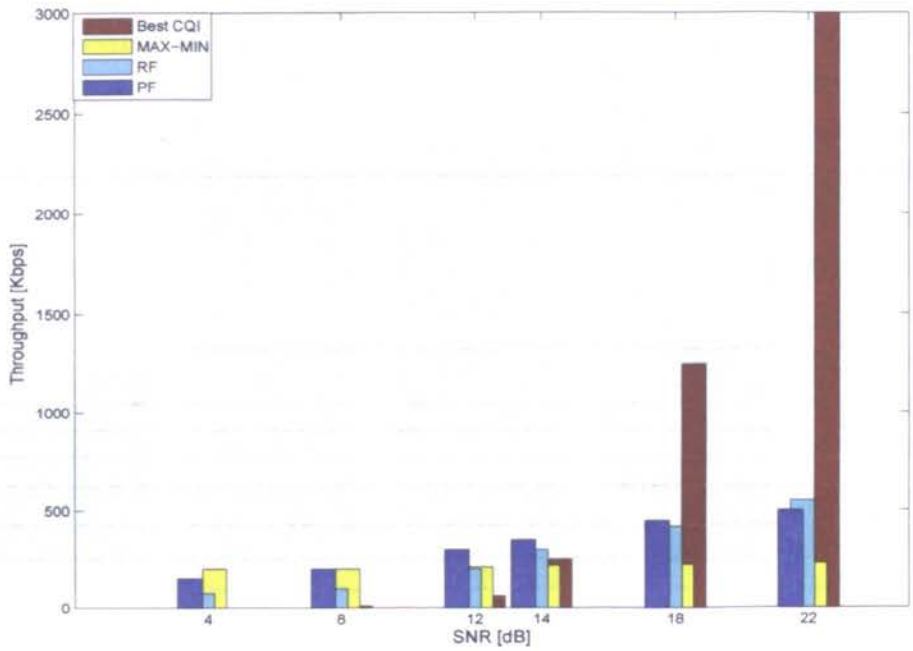


FIGURE 2.8: Throughput (Kbps) achieved for different users with different schedulers.

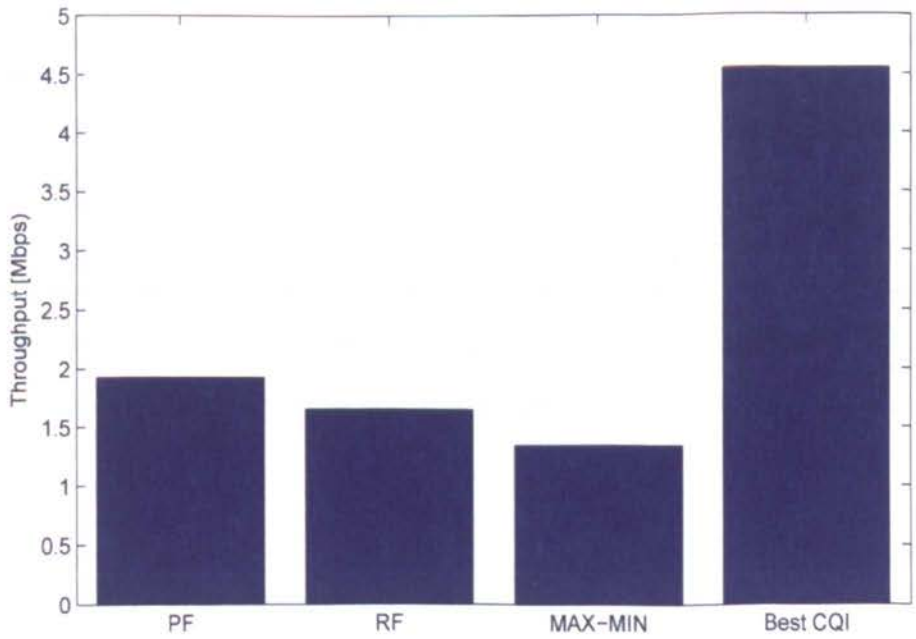


FIGURE 2.9: Total System or Cell throughput (Mbps) for different schedulers.

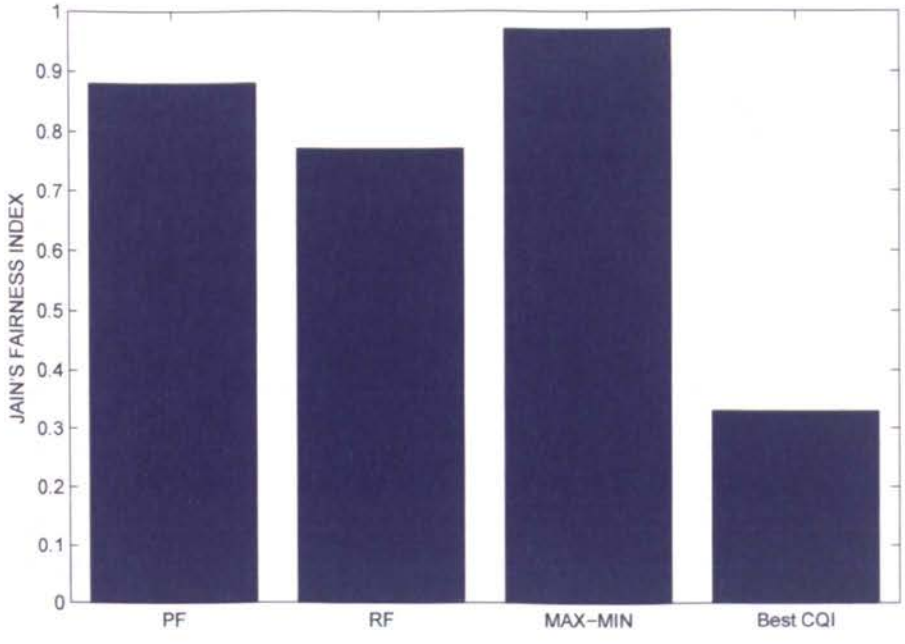


FIGURE 2.10: Jain's fairness index for different schedulers.

where $T(i)$ is the throughput achieved by each user i. If all the UEs have same throughput, the fairness index is 1. Fairness decreases as the differences between throughput increase.

In Figure 2.10, the Jain's fairness index is used to analyze the fairness performance. The performance of the *PF* scheduler is better in terms of fairness as well as system throughput, as shown in Figure 2.9. PF is a better choice than RF when a level of fairness in terms of throughput is required as PF also produces better system (cell) throughput, as shown in Figure 2.9.

2.3 Implementation of the proposed performance assessment setup through the extension of link level simulator.

The basic multi-user simulation scenario discussed in the above section exploits multi-user channel diversity by utilizing the AMC feature. The exploitation of HARQ procedure further increases the system throughput as indicated by the HARQ performance in the single downlink simulation setting. The AMC and HARQ are the most important features in optimum exploitation of the link capacity. The channel aware scheduling

strategies discussed in this chapter form the basis of designing a quality aware scheduling rules. When spectral efficiency is considered as the performance measure, then the Best CQI scheduler is the best scheduling rule as it maximizes the system throughput. Other issues such as fairness and QoS provisioning can be addressed at the expense of a reduced system throughput. The MAX-MIN scheduler achieves fairness (in terms of throughput) at the expense of system throughput. Therefore, an efficient scheduling strategy finds a good trade-off between fairness and efficiency such as the PF scheduler discussed above. This scheduler finds a good trade-off between fairness and efficiency. It is important to note that full buffer scenario, an infinitely large queue size, is considered throughout this chapter. In order to design quality aware strategies, the consideration of parameters such as the packet's waiting time in the queue becomes extremely important. The link level simulator settings discussed in this chapter consider the full buffer multiuser scheduling strategies where the channel quality feedback and time-averaged throughput were considered in the scheduling decisions. However, designing scheduling strategies considering an infinite queue is not realistic mainly for the following reasons:

- Video traffic, constituting a major proportion of traffic, exhibits a variable rate characteristics. The peak to average rate of video traffic is very high, thus the assumption that queues are always buffered with packets is not realistic.
- The amount of time packets reside in the buffer cannot be infinitely large as each packet has a processing deadline before which it should be received. Packets received after the deadline are generally not useful at the receiver. Radio resources utilized in scheduling packets whose decoding deadline has already been elapsed results in an inefficient system. Therefore, packets violating predefined delay budget are dropped.
- In wireless system, multi-user channel diversity allows users with good channel quality to be served well, thus the probability of flows having non-empty queues is low.

Subsequent sections presents the proposed finite buffer performance assessment setup through the extension of link level simulator.

2.3.1 Performance assessment setup for the proposed quality aware scheduling strategies (Chapters 3 to 6)

The overall integration of proposed blocks and the LTE link level simulator is shown in Figure 2.11. According to the figure, the LTE simulator batch file is the Matlab script

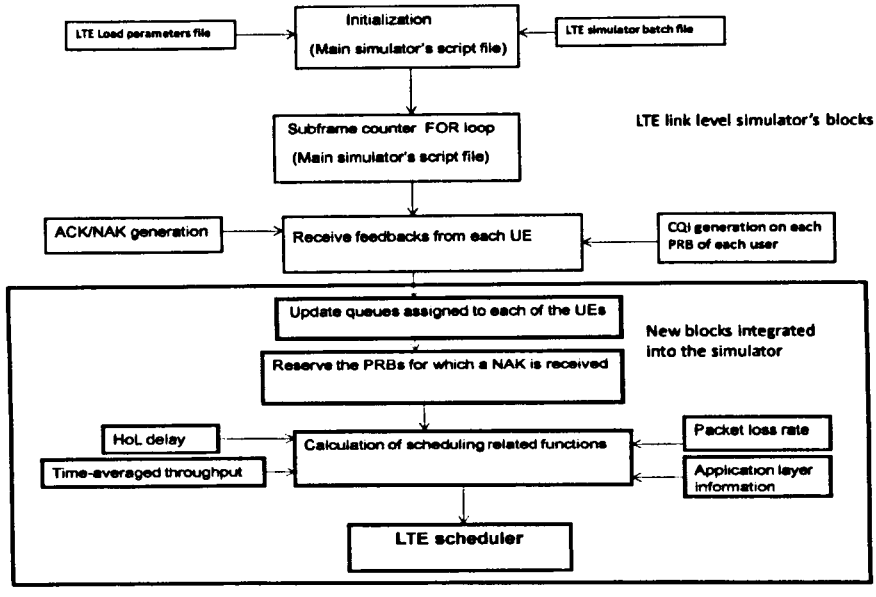


FIGURE 2.11: Integration of proposed blocks with the LTE link level simulator.

which runs the main simulator file. The number of TTIs to simulate is input to the batch file which then passes it to the FOR loop (TTI counter) in the main LTE link level simulator's script file. This script file contains the TTI counter FOR loop which initializes all the relevant classes of the eNodeB, UE, Server and Channel model. All the important parameters shown in Table 2.1 are feed into the simulator by the help of an LTE load parameters file. The main LTE link level simulator's script file passes all the parameters to the relevant class files. Once all the classes are initialized, each entity (UE and eNodeB) functions according to the physical layer procedures discussed in Section 2.1.

Algorithm 1 shows the structure of the main LTE link level simulator's script. According to the pseudo-code, the first TTI initializes and create instances of all the entities by using the class constructors of each entity. Each object of class UE is assigned an identification number (UE_{id}) which serve as the pointer and contains all the information assigned to the UE. When the server parameters are initialized, each UE is assigned a buffer at the eNodeB. However, packets are streamed into the buffer depending upon the start time parameter assigned to the server. After initializing the server parameters, the channel model of each UE is generated. According to Figure 2.5, the channel model requires the average SNR which is produced by the urban path-loss model equation (2.5). The path-loss equation depends upon the distance of the UE from eNodeB. When the object of UE class is created, the constructor passes several important parameters, one

Algorithm 1 pseudo-code of the main simulator file.

```
N_subframe: Total number of TTIs to simulate
N_UE: Total number of UEs to simulate
for TTI = 1 : N_subframe do
  if TTI = 1 then
    for UE_id = 1 : N_UE do
      Create an object of UE class and assign it an id UE_id
      Initialize server parameters for the UE with id UE_id
      Generate channel model for UE with id UE_id
    end for
  end if
  for UE_id = 1 : N_UE do
    Get SNR for UE_id using equation (2.5)
    Get feedback from UE_id
    Update packet buffer matrix assigned to UE_id
    Update performance matrix for UE_id
  end for
  Calculate the number of available PRBs available at the current TTI
  Apply the scheduling algorithm
  for UE_id = 1 : N_UE do
    UE_id channel coding of the generated data bits
    UE_id Symbol mapping
  end for
  OFDM symbol assembly
  IFFT procedure
  CP insertion
  Production of the transmitted signal
  for UE_id = 1 : N_UE do
    Random noise and channel model impact to the transmitted signal according to
    the SNR of UE_id
    Execute receiver process for UE_id
    Execute the decoding for UE_id
  end for
  Calculate all the performance parameters
end for
```

of them is the UE distance from the eNodeB. At each TTI, the average SNR of the UE is produced by the path-loss model. When the UE is served by assigning the resources, the ACK/NAK information is produced depending upon the outcome of decoding at the UE. When the ACK feedback is received then the packet buffer matrix assigned to UE_{id} at the eNodeB is updated. The performance matrix assigned to each UE is updated at every TTI. There can be several performance measures such as the throughput (successfully transmitted packets), average packet delay or video quality related performance metrics such as the Peak Signal to Noise Ratio (PSNR). In the event of NAK, the PRBs assigned to the UEs (UEs responding with NAK) are reserved. The available PRBs at each TTI depends upon the system bandwidth and the reserved PRBs. For instance,

if the operating bandwidth is 5MHz (a total of 25 PRBs available for allocation) and 3 PRBs assigned to a UE results in a NAK then the total number of PRBs available for allocation is 22. The next step is the PRB allocation step which is determined by a scheduling strategy, the packet buffer (matrix containing all the important information such packet's arrival time, delay budget, size) is accessed by the scheduler in evaluating the scheduling metric.

Chapters 3, 4, 5 and 6 discuss novel quality aware scheduling designs. Therefore, the pseudo-code 1 is utilized in each of the chapters with a different scheduling strategy. The output of a scheduling algorithm is the allocation of PRBs to different UEs. After the scheduling step, LTE physical layer's transmission procedure is applied as shown in the pseudo-code. The output of the physical layer's transmission is fed to the random noise generation and channel impairment step which introduces fading and other impairments according to the SNR and the channel model. The impaired signal is fed to each of the UEs. LTE physical layer's reception procedure is applied as shown in the pseudo-code. The output of the physical layer's procedure is the ACK/NAK generation. It is important to note that the pseudo-code does not show the signaling between the UEs and eNodeB. All the signaling between the eNodeB and UEs are signaled without any channel impairment. When the TTI loop is finished, the performance metric over the entire simulation range is computed. It is important to note that the SNR input to the channel (produced by the path-loss model), the fading model of the channel, the packet streaming to the buffers and other parameters utilize the statistical distribution model. Therefore, the results produced by the simulation setup can be easily reproduced by keeping the seed of the random number generation constant which allows the new scheduling strategies to be compared with state-of-the-art algorithms.

Subsequent sections present different server streaming scenarios utilized in the performance assessment for different quality aware scheduling strategies.

2.3.1.1 Proposed performance assessment setup for the QoS aware scheduling strategies (Chapters 3 and 4)

In order to simulate quality aware scheduling strategies, new entities are integrated into the simulator. The most important entity is the server which streams packets into the buffer at eNodeB. Figure 2.12 shows the video streaming process of the server. According to the figure, the server uses either a statistical based video streaming process or the video trace based approach. In statistical model based packet streaming, the inter-arrival time as well as the video packet size are produced using a statistical distribution process. In order to simulate variable rate video traffic streaming, the Truncated Pareto distribution

process is assumed. This statistical distribution process requires the minimum and maximum packet size and inter-arrival time as shown in Figure 2.12. The duration of one video frame and the number of packets in a frame are deterministic, these parameters are set according to the average rate of the video traffic. The LTE link level simulator uses a TTI counter which is incremented by 1 ms. In order to calculate the packet's arrival and discard time in the simulator, the streaming process utilizes the TTI counter. According to Figure 2.12, the TTI number is required by the server class in order to start the streaming. The starting time of the streaming can be generated randomly or deterministically depending upon the simulation scenario. Once the packet is generated, its arrival time is added to the TTI counter. The packet buffer (each field is represented by a column in the matrix) at the eNodeB stores the packets using the First In First Out (FIFO) process. The packet buffer manager function requires the delay budget of each packet. Whenever the packet is streamed into the buffer, the delay budget is added to the TTI counter and stored in the discard time field in the packet buffer matrix.

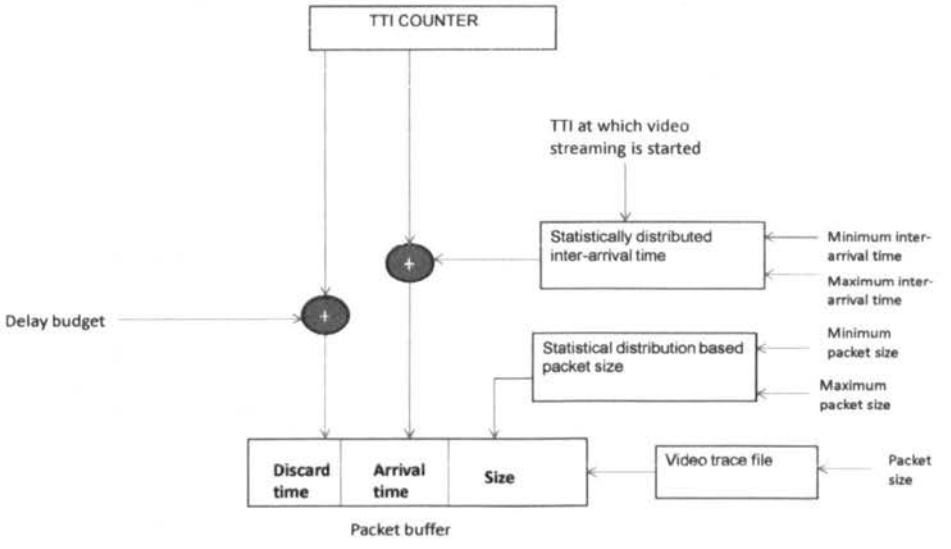


FIGURE 2.12: Proposed packet arrival process at eNodeB's buffer. The process is either based on the statistical distribution or the trace based approach.

Alternatively, the packet size can also be produced by the using a video trace file approach as shown in Figure 2.12. The trace file consists of columns of a video frame size, type of a video frame and encoding and decoding time. The number of packets produced per unit time is calculated by considering the frames per second (fps) and the inter-arrival time. For instance, a video sequence having a frame rate of 20 fps produces

1 frame in 50 ms. Considering the minimum and maximum limit of inter-arrival time of 2 and 8 ms and a video frame size of 1100 bytes, then according to these considered parameters, the Truncated Pareto distribution produces an inter-arrival time sequence of 5ms, 6ms, 8ms, 2ms, 4ms, 7ms..... Now considering that the packet size parameter is set to 200 bytes then the video frame will split into 6 packets with 5 packets having a size of 200 bytes whereas the size of the last packet is 100 bytes. If the value of the TTI counter is 1005 then the packets will be streamed into the buffer at the counter values of 1010, 1016, 1024, 1026, 1030 and the last packet will arrive at a counter value of 1037.

2.3.1.2 Proposed performance assessment setup for the QoE aware scheduling strategies (Chapter 5)

Cross-layer scheduling strategies consider application layer information in their scheduling decisions. In order to simulate cross-layer strategies, the information about the application layer is provided in the packet buffer matrix by adding columns containing information about the video sequence. Chapters 5 and 6 presents scheduling strategies considering the application layer information. Figure 2.13 shows the implementation of GOP based video streaming.

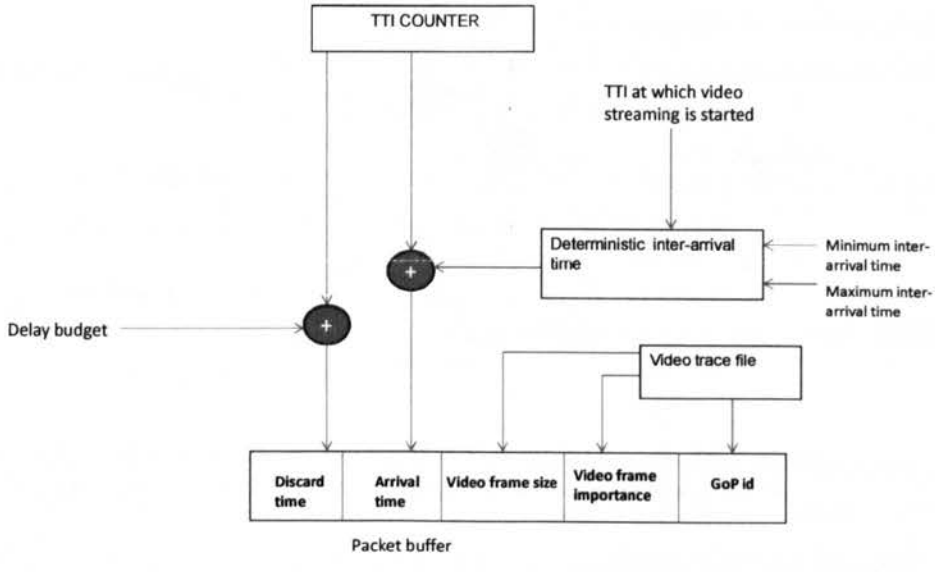


FIGURE 2.13: Proposed GOP based streaming at the eNodeB buffer. The streaming model is used in conjunction with the link level simulator in evaluating the cross-layer scheduling strategies proposed in Chapter 5.

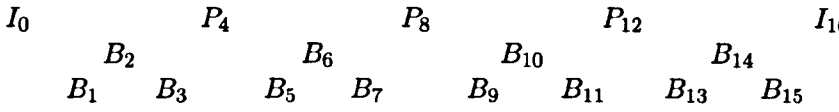


FIGURE 2.14: G16B12, 3 layers temporally scalable Video.

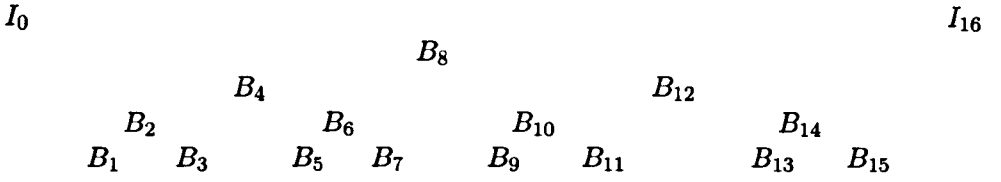


FIGURE 2.15: G16B15 Temporally scalable Video.

It is considered that each video sequence is encoded and organized in independently decodable units called Group of pictures (GOP). A GOP comprises a group of intra and inter frames as shown in Figures 2.14 and 2.15. According to Figure 2.14, I_0 (an intra frame) does not depend on any frame for its decoding. The P and B frames in the GOP are inter frames where a P frame is a forward predictive coded picture and a B frame is a bidirectionally predictive coded picture. Inter frames achieve a higher compression ratio by exploiting the temporal redundancy between neighboring frames. Forward predictive coded picture is predicted from an earlier frame. For instance in Figure 2.14, P_4 is predicted from frame I_0 and P_8 requires frame P_4 for its decoding. On the other hand, bidirectionally predictive coded picture achieves the best compression ratio by exploiting the temporal redundancy between two higher layer neighboring frames as shown in Figures 2.14 and 2.15. Table 2.3 shows the dependency structure of all the bidirectionally predictive coded frames in Figure 2.15. In order to provide information about the GOP structure, size of video frames and quality of each video frame is obtained from video trace files. Trace files can be produced using the video encoder or they can be directly downloaded from an online video trace database in. Table 2.4 shows the trace file sample of the GOP structure of Figure 2.15, the first column comprising the video frame size whereas the second and third columns consist of video frame types and quality respectively. The trace file sample table shows all the video frames in the display order. The trace files can be produced both in the display order as well as in the encoding order from [38]. Figure 2.13 shows the GOP based streaming. The size of all the frames in one GOP (1 Intra frame and 15 Inter frames) are input to the video packet buffer matrix with each frame representing one packet. According to Figure 2.13, the buffer manager time stamps the whole GOP and the scheduler should assign enough resources to schedule the whole GOP before the delay budget expires. For instance, a delay budget of 500 ms corresponds to a scenario where remaining frames in the buffer are dropped and all the 16 frames of the next GOP are input to the buffer matrix after 500 ms cycle. The frames dropped by the scheduler due to deadline violation are

concealed by the Frame Copy (FC) algorithm. In FC the last correctly received frame is displayed in place of the lost frame. Tables 2.5 and 2.6 report the sample frame copy trace files of the GOP structure shown in Figure 2.15. When only the first layer frame of the GOP is scheduled within the delay budget and the remaining frames are dropped then the video quality is evaluated by considering the PSNR of each frame shown in Table 2.5. Similarly when 2 layers are received then Table 2.6 is utilized for the video quality evaluation. The number of trace files for video quality evaluation depends upon the number of temporal layers present in the GOP structure and can be easily obtained from [38] or produced using the SVC encoder. The trace based approach follows the guidelines and procedures given in [39].

TABLE 2.3: Dependency structure of the frames shown in Figure 2.15.

Frames	Dependent frames
I_0	Intra frame does not depend upon any frame.
B_8	I_0 and I_{16}
B_4	I_0 and B_8
B_{12}	B_8 and I_{16}
B_2	I_0 and B_4
B_6	B_4 and B_8
B_{10}	B_8 and B_{12}
B_{14}	B_{12} and I_{16}
B_1	I_0 and B_2
B_3	B_2 and B_4
B_5	B_4 and B_6
B_7	B_6 and B_8
B_9	B_8 and B_{10}
B_{11}	B_{10} and B_{12}
B_{13}	B_{12} and B_{14}
B_{15}	B_{14} and I_{16}
I_{16}	Intra frame does not depend upon any frame.

2.3.1.3 Proposed performance assessment setup for the joint QoS and QoE aware scheduling strategies (Chapter 6)

The video streaming scenario for chapter 6 is shown in Figure 2.16. Instead of transmission of the whole GOP, the video streaming is based on packets as in the previous streaming setup shown in Figure 2.12. However, the streaming setup shown in Figure 2.16 adds packet priority information in the packet buffer matrix. The packet marking

TABLE 2.4: Sample trace file format of the GOP structure shown in Figure 2.15.

Frame size (bytes)	Frame type	Video quality (PSNR)
2196	I_0	41.62594
225	B_1	42.00678
348	B_2	42.04672
225	B_3	42.1076
391	B_4	42.09565
66	B_5	42.17076
240	B_6	42.21184
202	B_7	42.0804
442	B_8	41.9898
66	B_9	42.01629
188	B_{10}	42.01806
147	B_{11}	41.95723
270	B_{12}	42.08406
154	B_{13}	42.0504
201	B_{14}	41.98359
71	B_{15}	41.97584
2261	I_{16}	41.52384

TABLE 2.5: Sample trace file when only one temporal layer in Figure 2.15 is received, 0 indicates that the frame is dropped.

Frame size	**PSNR** [dB]
2196	41.62594
0	38.09945
0	33.8092
0	32.36718
0	29.59872
0	29.58075
0	29.17205
0	28.37831
0	28.42554
0	28.40655
0	28.39925
0	28.48415
0	27.89558
0	27.7797
0	27.27696
0	27.24901
2261	41.52384

TABLE 2.6: Sample trace file when two temporal layers in Figure 2.15 are received, 0 indicates that the frame is dropped.

Frame type	**PSNR** [dB]
2196	41.62594
0	38.09945
0	33.8092
0	32.36718
0	29.59872
0	29.58075
0	29.17205
0	28.37831
442	41.9898
0	41.22557
0	41.11144
0	39.6618
0	36.76608
0	35.56123
0	33.32878
0	33.25925
2261	41.52384

algorithm provides each video layer with the priority index which is then utilized by the scheduling algorithm.

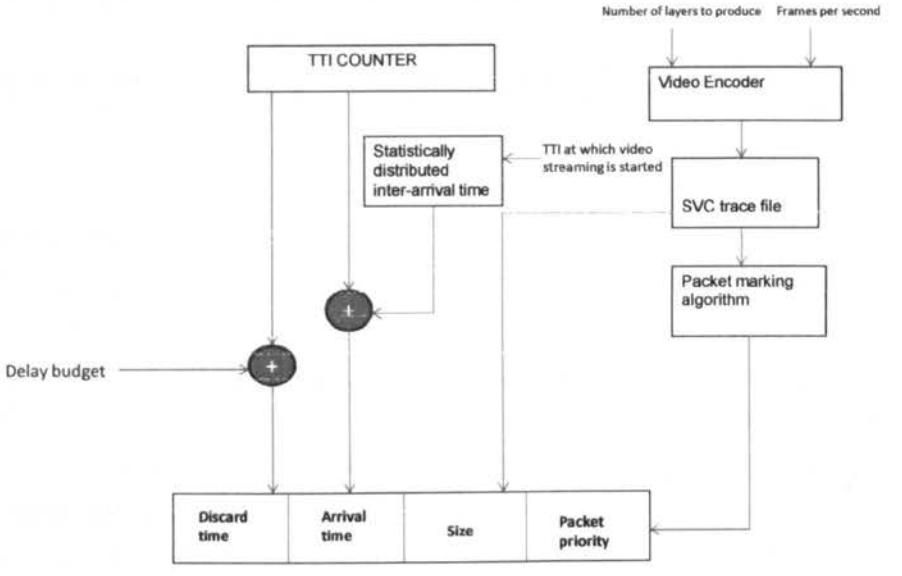


FIGURE 2.16: Proposed prioritized packet streaming setup. The streaming model is used in conjunction with the link level simulator in evaluating the cross-layer scheduling strategies proposed in Chapter 6.

Chapter 3

Opportunistic Packet Loss Fair Scheduling for Delay-Sensitive Applications over LTE Systems

3.1 Introduction

In the previous chapter, the basic scheduling strategies by considering an infinite queue for each UE at the eNodeB was analyzed. It is important to note that designing a scheduling strategy based on full buffer lacks the service quality awareness of different flows with different service needs such as VoIP and video. In this chapter, a QoS aware scheduling strategy aimed at guaranteeing the QoS requirements such as delay bound and packet loss ratio thresholds of different real-time flows is considered. There are many scheduling algorithms designed for single carrier systems to accommodate real time traffic, such as the well known M-LWDF algorithm [40]. This scheduler serves the flow maximizing the product $\gamma_i H_i^{(n)} R_i^{(n)}$, where $H_i^{(n)}$ is the HOL packet delay for queue i, $R_i^{(n)}$ is the rate available at scheduling instant n according to the instantaneous channel condition; γ_j is a constant whose value is adjusted to account for different delay requirements for different flows. There are two main reasons why M-LWDF is used for delay sensitive traffic: one is the fact that it is throughput optimal (analytically), proved in [41], and the second is that it is relatively simple to implement this algorithm, since only time stamping is required for the incoming packets. Due to its simplicity, this algorithm has been widely used in single carrier systems for real time applications. The use of M-LWDF has been adapted for the LTE system (which is based on Orthogonal Frequency Division Multiple Access (OFDMA)) in [42]; a comparison is provided among well known packet scheduling algorithms designed for single carrier systems, such as PF [34],

M-LWDF [40], maximum throughput [13], Exponential Proportional Fair (EXP/PF) [43] [44]. According to this paper M-LWDF, in terms of efficiency and fairness, outperforms the PF, maximum throughput and EXP/PF scheduling algorithms.

Some delay based scheduling algorithms have also been proposed for multicarrier systems. In [45] an Opportunistic Scheduling (OS) algorithm is proposed. This scheduling algorithm is quite complex due to the fact that it performs resource assignment and resource allocation in different steps, resulting in a more complex scheduling algorithm. The algorithm operates in two steps, where the first step consists of allocating subcarriers to users with a good channel by exploiting multiuser diversity; if during the first step some of the subcarriers remain unused, then those subcarriers are allocated to users having higher HOL delay to incorporate fairness. Hence, the second step consists of a subcarrier assignment algorithm which assigns the best subcarriers to users which have suffered from higher HOL delay violations. This scheduling algorithm lacks satisfactory fairness, since there is a high probability that the scheduler allocates all the subcarriers to good channel users, thus inhibiting the fairness step. Moreover, there is a need for a joint scheduling and resource allocation step which takes all the necessary information into account before assigning the resources, in order to allow a tighter control over the resources with lower complexity. Thus the signalling cost is high with such an approach. OS performance over LTE is analyzed in [46], where a simple scheduling algorithm, called Delay Prioritized Scheduler (DPS), is also designed. The DPS algorithm takes only packet delay information into account and assigns the best PRB to the users whose packets have remained in the buffer for a longer period of time. The best PRB for each user is determined by taking instantaneous SNR into account. The DPS algorithm calculates the packet delay of all users and assigns the best PRB to the user with the highest packet delay; in the next iteration the same process is repeated until all the PRBs are assigned. This delay prioritized scheduling algorithm outperforms the OS algorithm in terms of system throughput and also achieves very low PLR even under loaded conditions. However, in the DPS algorithm the only scheduling criterion is the packet delay information, hence users with bad channel conditions will force the PRB allocation towards themselves, thus limiting the system capacity. In [47] a PLF scheduling rule is proposed for OFDMA systems in order to provide QoS for diverse real time traffic; the PLF rule provides short term as well as long term fairness, but this scheduling algorithm lacks the exploitation of statistically independent multiuser frequency selective fading.

A scheduling strategy overcoming some of the limits of the preceding scheduling algorithms is proposed. This is an Opportunistic Packet Loss Fair (OPLF) scheduling algorithm, based on calculating a simple dynamic priority function which depends on HOL packet delay, PLR and achievable instantaneous downlink rate of each user. The

remainder of this chapter is organized as follows. Section 3.2 describes the system model considered, whereas Section 3.3 presents the proposed scheduling strategy. Simulation set-up and results are presented in Sections 3.4 and 3.5 respectively with the concluding remarks appearing in Section 3.6.

TABLE 3.1: Mathematical symbols utilized in Chapter 3.

Symbol	Explanation
i	User/flow index.
φ	PRB index.
n	Current scheduling epoch.
$n_{\text{enter}}(i)$	The time at which a packet of user i enters the buffer at the LTE base station (eNodeB) and is time stamped by the buffer manager.
$\text{PRF}^{(n)}{}_i$	Priority function of user/flow i at scheduling epoch n.
$H_i^{(n)}$	HoL packet delay of user/flow i at scheduling epoch n.
$H_{\max}$	Maximum delay budget of user i's packet.
$\text{plr}^{(n)}{}_i$	Packet loss ratio of user i at scheduling instant n calculated over the moving average transmission window t_w.
$\text{plr}_{\text{thr}_i}$	Maximum tolerated PLR for user i.
$P^{(m)}_{\text{transmit}_i}$ and $P^{(m)}_{\text{drop}_i}$	Number of transmitted and dropped packets of user i over the moving average transmission window t_w.
$R^{(n)}_{i,\varphi}$	Instantaneous rate of user i on PRB φ.
$\overline{R}_i^{(n)}$	Instantaneous rate of user i averaged over all unallocated PRBs.
$\Phi_{\text{URB}}(n,k)$	Set of unallocated PRBs during iteration k at scheduling instant n.
$\lvert\Phi_{\text{URB}}(n,k)\rvert$	Cardinality of $\Phi_{\text{URB}}(n,k)$.
$\Phi_{\text{PRB},i^*}(n,k)$	Set of PRBs allocated to user i^* which maximizes the priority function $PRF_i^{(n)}$ by iteration k at scheduling instant n.
i^*	Index of a user maximizing the priority function $\text{PRF}^{(n)}{}_i$ by iteration K.
$I_{\text{active flows}}$	Number of active flows in the system.
φ^*	User i^*'s least faded PRB among the set of unallocated PRBs, $\varphi^* \in \Phi_{URB}(n,k)$.
γ_i	A constant whose value is adjusted to account for different delay and packet loss rate requirements of different flows.
κ	Truncated pareto distribution parameter for minimum packet size and minimum inter-arrival time.
m	Truncated pareto distribution parameter for maximum packet size and maximum inter-arrival time.
a	shaping factor for the truncated pareto distribution

3.2 System Model

The system model consists of an OFDM single antenna SISO multiuser LTE system with the focus on the downlink. In SISO system a PRB can only be assigned to one

user at any scheduling instant and hence there is no overlapping in PRB allocation. A single cell scenario is considered in which the serving eNodeB is at the center of the cell. The serving eNodeB's MAC scheduler controls all the available PRBs by allocating them to active flows competing for resources. Each user is assigned a buffer at the eNodeB. When a packet reaches the serving eNodeB, the buffer management system time stamps and queues each packet in a FIFO order. At the start of each scheduling instant, *i.e.*, before the multiuser scheduling decision, the HOL packet delay (also called sojourn time) for each user's packet is calculated by subtracting the arrival time of the packet from the current time. If the HOL packet delay is above the considered threshold $D_{\max}$, depending on the QoS requirements for the application, the packet is discarded by the buffer management system.

In the network, users report their instantaneous channel quality by means of a quantized feedback called CQI. There are different feedback granularities in the standard, for instance the user can send just a single CQI for the whole bandwidth or a separate CQI for each PRB. If according to the scheduling decision more than one PRB, with different CQI values on each PRB, are assigned to a user then it is necessary to calculate an average supported CQI value. The scheduler carries out the averaging, *i.e.*, it maps the SINR experienced by the allocated PRBs into one (equivalent) Additive White Gaussian Noise (AWGN) SNR. One of the mostly used averaging methods is the Mutual Information Effective SNR Mapping (MIESM) [19] [25] [26]. After resource allocation, the scheduler uses the Mutual Information Effective SNR Mapping method to calculate a single CQI value to be used for all the allocated PRBs of a particular user. Refer to Chapter 2 (Section 2.3.1) for detailed description on the simulation platform.

3.3 Opportunistic Packet Loss Fair Scheduler

Although delay sensitive applications often tolerate losses, they typically require that the PLR is kept below a threshold. Hence an OPLF scheduling algorithm is proposed which is based on calculating a simple dynamic priority function which depends on the HOL packet delay, the PLR and the achievable instantaneous downlink rate of each user.

The packet scheduler exploits channel information in order to achieve multiuser diversity gain, by assigning PRBs to the users experiencing lower attenuation. However, it is important to note that relying only on the diversity gain can lead to an unfair treatment of users experiencing lower average channel quality. The proposed scheduling strategy addresses this issue by providing fairness and at the same time exploiting multiuser frequency diversity.

The available resource blocks are allocated to users through an iterative process, where the total number of iterations is equal to the total number of PRBs available at each Transmission Time Interval (TTI) and at each iteration only one PRB is allocated to the user which maximizes a priority function.

The proposed priority function, $\mathrm{PRF}^{(n)}{}_i$, of user i (OPLF rule) at scheduling epoch n is:

$$\mathrm{PRF}^{(n)}{}_i = \frac{\overline{R}_i^{(n)}\, H_i^{(n)}\, \mathrm{plr}^{(n)}{}_i}{H_{\max}\, \mathrm{plr}_{\mathrm{thr}_i}} \tag{3.1}$$

where

$$\overline{R}_i^{(n)} = \frac{1}{|\Phi_{\mathrm{URB}}(n,k)|} \sum_{\varphi \in \Phi_{\mathrm{URB}(n,k)}} R_{i,\varphi}^{(n)} \tag{3.2}$$

and

$$H_i^{(n)} = n - n_{\mathrm{enter}}(i) \tag{3.3}$$

$$\mathrm{plr}_i^{(n)} = \frac{\sum_{m=n-t_w}^{n} P_{\mathrm{drop}_i}^{(m)}}{\sum_{m=n-t_w}^{n} \left(P_{\mathrm{transmit}_i}^{(m)} + P_{\mathrm{drop}_i}^{(m)} \right)} \tag{3.4}$$

- $\overline{R}_i^{(n)}$ is the instantaneous rate of user i averaged over all unallocated PRBs.
- $\Phi_{\mathrm{URB}}(n,k)$ denotes the set of unallocated PRBs during iteration k at scheduling instant n and $|\Phi_{\mathrm{URB}}(n,k)|$ its cardinality
- $R_{i,\varphi}^{(n)}$ is the instantaneous rate of user i on PRB φ.
- $H_i^{(n)}$ is the HOL packet delay of user i at current scheduling instant n.
- $n_{\mathrm{enter}}(i)$ is the time at which a packet of user i enters the buffer of the eNodeB and is time stamped by the buffer manager.
- $\mathrm{plr}_i^{(n)}$ is the packet loss ratio of user i at scheduling instant n calculated over the moving average transmission window t_w, $P_{\mathrm{transmit}_i}^{(m)}$ and $P_{\mathrm{drop}_i}^{(m)}$ are the number of transmitted and dropped packets over the moving average transmission window t_w.
- $\mathrm{PLR}_{\mathrm{thr}_i}$ is the maximum PLR tolerated for user i.

- $H_{\max}$ is the maximum HOL delay tolerated for packet of user i.

The priority function above is high when the instantaneous PLR of user i is high with respect to the threshold (hence the user should get more resources to reduce it) and also when the HOL delay is high with respect to the maximum tolerated delay. $R_i^{(n)}$ improves the system efficiency by exploiting the statistically independent multiuser frequency selective fading. The priority function tends to favor the user with the highest HOL delay and the highest PLR; however, if some of the PRBs of such a user are under a deep fade, the factor $R_i^{(n)}$ ensures that those PRBs are not allocated to such user.

The OPLF scheduling strategy is described as a pseudo-code in Algorithm 2.

Algorithm 2 Opportunistic Packet Loss Fair Scheduler

repeat
  Calculate $H_i^{(n)}$, for all i, according to (3.3)
  Calculate $\text{plr}_i^{(n)}$, for all i, according to (3.4)
  while $|\Phi_{\text{URB}}(n,k)| > 0$ **do**
    Calculate $\overline{R}_i^{(n)}$, for all i, according to (3.2)
    for $i = 1$ to $I_{\text{active flows}}$ **do**
      Calculate $\text{PRF}_i^{(n)}$ according to (3.1)
    end for
    $i^* = \arg\ \max \text{PRF}_i^{(n)}$
    $\varphi^* = \arg\ \max R_{i^*,\varphi}^{(n)}$ where $\varphi^* \in \Phi_{URB}(n,k)$
    $\Phi_{\text{PRB},i^*}(n,k+1) = \Phi_{\text{PRB},i^*}(n,k) + \{\varphi^*\}$
    $\Phi_{\text{URB}}(n,k+1) = \Phi_{\text{URB}}(n,k) - \{\varphi^*\}$
    if packet of user i^* is scheduled **then**
      $H_{i^*}^{(n)} = n - n_{\text{enter}}(i^*)$
    end if
    If a user is scheduled on more than one PRB, then MIESM is applied to calculate the average CQI.
  end while
  TTI = TTI + 1
until END OF SIMULATION

$\Phi_{\text{PRB},i^*}(n,k)$ denotes the set of PRBs allocated to user i^* which maximizes the priority function $PRF_i^{(n)}$ by iteration k at scheduling instant n. It is important to note that at each iteration only one PRB is allocated to the user which maximizes the priority function, *i.e.*, the total number of iterations is equal to the total number of PRBs available at each TTI. In order to fully utilize the available resources, if more than one PRB have the same instantaneous CQI, then the PRB which experiences the maximum fading for other users, or in other words has the lowest quality for the other users, is selected. In this way, multiuser frequency diversity is exploited. The priority function is dynamic, since when a best quality PRB is allocated to the user, in the next iteration the factor $\overline{R}_i^{(n)}$ will be changed as user i^*, which maximizes $\text{PRF}_i^{(n)}$, is allocated its best remaining

PRB φ^*. The proposed scheduling algorithm is different from the M-LWDF rule, which also exploits HOL delay but relies on the the proportional fair rule in order to provide fairness. OPLF provides fairness by exploiting multiuser frequency diversity, since in frequency selective fading the same PRB for different users undergoes a statistically independent fading, hence a PRB which is under a deep fade might be the best PRB for another user. The $\text{plr}_i^{(n)}$ factor in the priority function equation provides a degree of fairness for users having poor average channel condition. Simulation results as seen in Section 3.5 confirm that the proposed scheduler provides better fairness and throughput than the M-LWDF and the PLF rule which uses proportional fair characteristics.

The positive features of the proposed algorithm are summarized below.

1. The bandwidth is utilized efficiently, as each user is scheduled on its best remaining PRBs and the scheduler will not assign resources to a user whose channel is under a deep fade, as indicated by the priority function.

2. Real-time traffic with diverse QoS requirements can be accommodated, and only information on the delay threshold $H_{\max}$ and on the packet loss threshold PLR_{thr} is required.

3. QoS parameters are used in the scheduling decisions, which ensures that users will get a minimum proportion of resources even if the average channel condition is low.

4. For real time applications fairness is guaranteed when current packet loss rate is distributed proportionally equal among all the users competing for resources . In this context, the main goal is to ensure fair PLR distribution over the moving average transmission window of size t_w thus achieving short term fairness. According to [48] short term fairness guarantees long term fairness, but not vice versa.

3.4 Simulation Environment

The assumptions considered in the simulations are reported below.

- The channel quality of each user remains constant during the subframe period of 1 ms, although it changes from subframe to subframe.

- CQI feedback from UE to the eNodeB is error free. The error free assumption of the feedback channel is satisfied by using efficient and heavily coded feedback stream as is anyway customary for LTE system.

- It is assumed that equal downlink transmit power is allocated on each PRB.
- It is assumed that, at any time instant, pathloss is fixed on each PRB. Multipath induced fading is represented by a tapped delay line model (Typical Urban). The propagation model used follows the guidelines in [49][50].

Simulation parameters are reported in Table 3.2.

TABLE 3.2: Simulation parameters - Downlink LTE scheduling for real time applications.

PARAMETERS	VALUE
Bandwidth	3 MHz
Carrier frequency	2.1 GHz
No. of PRBs	15
No. of users	Variable (40 to 60)
UE distribution	Uniform
Cell radius	2 km
Application	NRTV
Admission Control	No Admission control
Mode	Tx = 1 and Rx = 1 (SISO mode)
Channel	3GPP-TU (Typical Urban)
Pathloss model	Hata-Cost-231 model (urban pathloss model)
HARQ	Synchronous retransmissions (Up to 3 retransmissions)
Channel Fading	Block Fading (1 ms)
UE speed	15 to 100 km/h (users moving independently at variable speed)
$H_{\max}$	50 ms
$\mathrm{PLR}_{\mathrm{thr_i}}$	1%
t_w	1s

3.4.1 Video traffic model

In order to analyze the performance of channel aware schedulers on real-time traffic, a Near Real Time Video (NRTV) model [51] is used. In order to model the variability in size and inter-arrival time between packets, a Truncated Pareto Distribution is considered, with probability density function (pdf):

$$f_x(x) = \begin{cases} \dfrac{a\kappa^a}{x^{a+1}} & \kappa \leq x < m, \\ (\dfrac{\kappa}{m})^a & x = m \end{cases} \tag{3.5}$$

The values considered for the parameters of the distribution a, m and κ are reported in Table 3.3. The total number of slices in a frame is deterministic and each slice corresponds to one packet. 8 slices (packets) per frame is considered in the simulation scenario. The total duration of one frame is also deterministic, considered here as 50 ms, which results in 20 fps. Based on these parameters, packets are produced and streamed into the user buffer at the eNodeB. Before entering the buffer of each user, packets are time stamped and served in FIFO order. According to the parameters below, the average rate of video streaming is approximately 120 kbps. QoS requirements are assumed the same for each flow. The maximum tolerated delay $H_{\max}$ is set to 50 ms and the packet loss rate threshold $\text{plr}_{\text{thr}_i}$ is set to 1%. When scheduled and transmitted successfully through the air interface, packets are assumed to be played. A packet is assumed to be lost if, due to retransmissions, the deadline of the packet is reached.

TABLE 3.3: Video traffic model parameters

Streaming Information	Distribution	Parameters
Inter-arrival time between frames	Deterministic	20 fps: duration of one frame is 50 ms
Total packets or slices in one frame	Deterministic	8
Inter-arrival time between packets	Truncated Pareto distribution	Min time κ =4 ms Max time m= 8 ms a=1.2
Packet (slice) size	Truncated Pareto distribution	Min size κ =65 bytes Max size m= 150 bytes a=1.2

3.4.2 Benchmark scheduling strategies

The priority functions of the PLF, M-LWDF, and PF rules, considered in the following for comparison, are reported below:

- PLF scheduling rule

$$\text{PRF}_{i,\varphi}^{(n)} = \left(\frac{R_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}}\right) \cdot \left(\frac{\text{plr}_i^{(n)}}{\text{plr}_{\text{thr}} * H_{\max}}\right) \tag{3.6}$$

- M-LWDF scheduling rule

$$\text{PRF}_{i,\varphi}^{(n)} = \gamma_i \frac{R_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} H_i^{(n)} \tag{3.7}$$

- PF scheduling rule

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \frac{R_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} \tag{3.8}$$

In the equations above $R_{i,\mathrm{ave}}^{(n)}$ and γ_i are respectively the moving average of the rate achieved over a transmission window size t_w for user i at time t, and a constant whose value is adjusted to account for different delay requirements of different flows.

It is important to note that all the scheduling rules are implemented in a way that priority function of all active flows are calculated on each PRB, the scheduler allocates the PRB to the user which maximizes the priority function so that all these rules exploit multiuser frequency diversity and a just comparison among all the scheduling rules is achieved.

3.4.3 Performance metrics

The performance of the proposed algorithm is evaluated and compared with the benchmark scheduling strategies above. The comparison is performed in terms of cell PLR and its standard deviation, system throughput, and average system delay.

The cell throughput or system throughput represents the total amount of packets successfully transmitted through the air interface from eNodeB to all the active flows in the time unit. The cell PLR is the ratio of the total number of lost packets to the total number of packets produced. The first two metrics evaluate the efficiency of each algorithm.

In order to analyze the fairness of the different scheduling strategies, the standard deviation of the PLR of each user is also calculated. The lower the standard deviation, the higher the level of fairness among the users.

The average packet delay corresponds to the average amount of time packets reside in the buffer. The characteristics of all the considered schedulers is summarized in Table 3.4.

3.5 Simulation results

The system throughput performance of the proposed algorithm and the reference ones is shown in Figure 3.1. It is clear from the figure that OPLF outperforms the M-LWDF, PF and PLF scheduling rules. When the load is above 45 users, the PLF rule outperforms the M-LWDF rule, mainly because the PLF rule does not incorporate HOL delay

TABLE 3.4: Characteristics of different schedulers

Scheduler	Channel aware	Delay aware	Packet loss aware
PF	yes	no	no
M-LWDF	yes	yes	no
PLF	yes	no	yes
OPLF	yes	yes	yes

information, but uses only PLR and the proportional fair rule, therefore users having higher PLR will get higher priority irrespective of the packet's deadline, whereas the proportional fair term in the PLF rule ensures that bad channel users do not affect the system efficiency. The performance of the PF algorithm is the worst mainly because this rule neither takes HOL delay nor PLR information into account in the scheduling decision. It is important to note that the system throughput of all the schedulers decreases after 55 users as this is an extreme load condition. This means that the QoS of most of the users is affected as more users are added and, since fairness is incorporated in all the algorithms, the effective throughput decreases due to the limited number of resources available, which results in more deadline violations.

The cell PLR performance of all the algorithms is shown in Figure 3.2. It is clear from the figure that OPLF outperforms the M-LWDF, PF and PLF rules also with this metric. Note that the PLR threshold considered for this application is 1%. It is important to note that when the number of users is 45, M-LWDF is slightly better than the PLF scheduling rule, however under high load PLF performs better than the M-LWDF rule which does not exploit PLR information in the priority function.

Figure 3.3 can be considered for evaluating the fairness performance of the algorithms. Fairness is an important criterion as algorithms may provide higher cell throughput at the expense of lower throughput (and hence quality) for users with bad channel conditions. In order to analyze the fairness performance of all the algorithms, standard deviation of the PLR of all the users is considered; a higher standard deviation implies that some users receive good service whereas some receive bad service, hence fairness is low. The proposed scheduler achieves lower standard deviation in comparison with the other considered algorithms. When a user suffers from HOL packet delay violation, such an event is characterized by a higher priority function value in the case of PLF and OPLF schedulers. Therefore more resources are assigned to such a user so that further delay violations are avoided. This is not the case in M-LWDF, delay aware and packet loss blind, and PF, delay and packet loss blind, schedulers. The fairness performance of the OPLF scheduler is better than the PLF scheduler mainly because OPLF is deadline

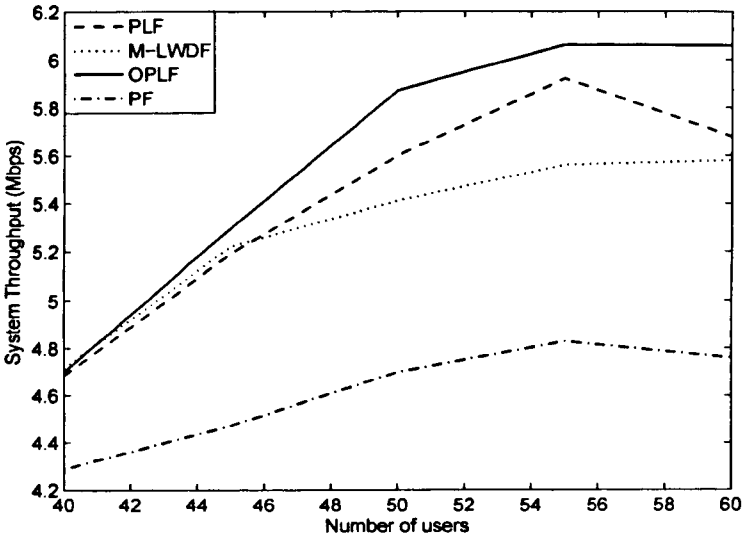


FIGURE 3.1: System throughput vs. number of users.

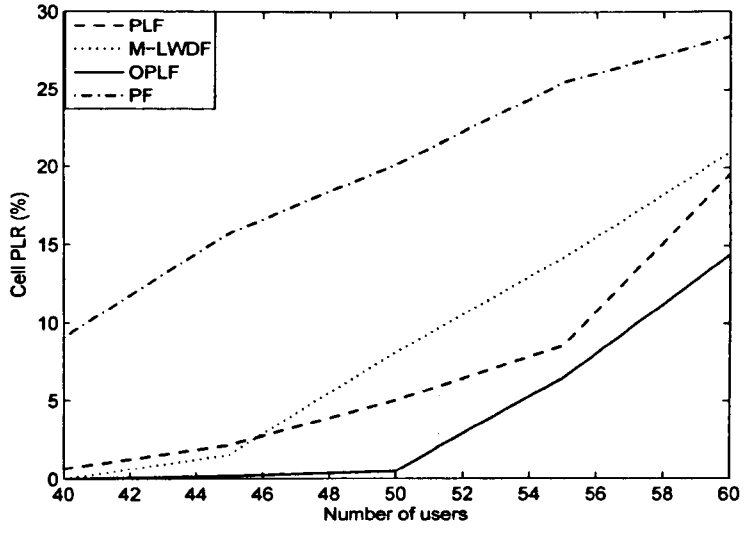


FIGURE 3.2: Cell packet loss ratio (%) vs. number of users.

aware. If users having same PLR violation, OPLF adapts to this situation by giving more priority to the user having a higher HOL packet delay, therefore further delay violations are avoided. It is important to note that, although M-LWDF results in a slightly better system throughput than the PLF scheduling rule when the number of users is 45, PLF outperforms M-LWDF in terms of fairness, as the scheduling decision in PLF is mainly based on the PLR. The fairness performance of the PLF rule is better and close to the OPLF rule under heavy load.

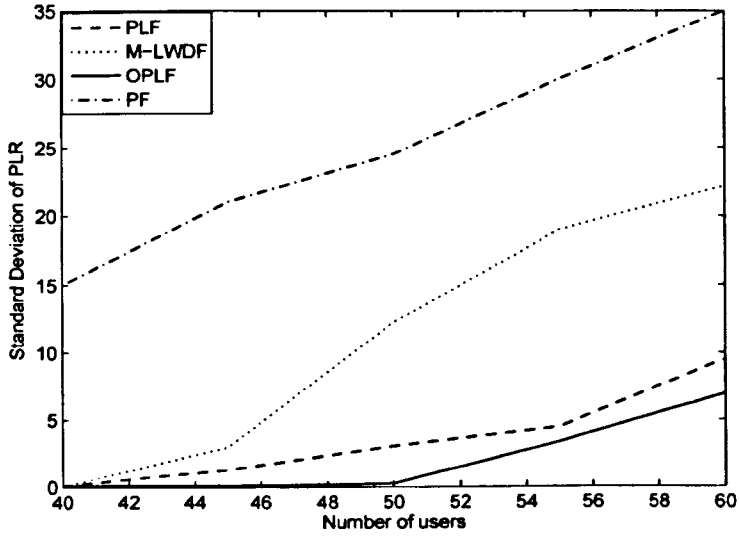


FIGURE 3.3: Standard deviation of PLR vs. number of users.

3.6 Conclusion

An OPLF scheduling algorithm for downlink scheduling at the MAC layer for delay sensitive traffic in wireless systems based on OFDMA is proposed. This algorithm outperforms state-of-the art algorithms, such as M-LWDF, PF, and PLF, in terms of throughput, packet loss rate and fairness, also keeping packet delay below a fixed threshold. With respect to existing algorithms, the proposed algorithm will thus enable the allocation in a cell of a higher number of users served with satisfactory quality.

In this chapter, the proposed scheduling strategy mainly considers the QoS service needs of the delay sensitive flows. The scheduling metric exploited the delay urgency and packet loss threshold limit of the delay sensitive traffic flows. Chapter 4 considers the service needs of the delay sensitive as well as the best-effort flows, thus extending the QoS-aware scheduling strategy to different types of traffic with different service needs.

Chapter 4

QoS-Aware Composite Scheduling using Fuzzy Reactive And Proactive Controllers

4.1 Introduction

QoS aware scheduling solutions available in the literature broadly fall into three classes. Some approaches solve the problem of resource allocation using optimal solutions, in other cases resource allocation and resource assignments are performed in two separate steps; other approaches simply target at adapting schemes originally proposed for Time Division Multiple Access (TDMA) to OFDMA systems. Thus, the scheduling solutions fall into the following three classes:

1. In [52][53] [54][55] resource allocation is modeled as a convex optimization problem. The water-filling algorithm is used to solve the convex optimization problem by considering a continuous objective function. Linear integer programming is also widely used in solving the resource allocation problem by first transforming the nonlinear optimization problem into a linear problem. The main drawback of these strategies is the high computation complexity. Since the TTI in LTE is only 1 ms, these algorithms are not feasible from an implementation point of view.

2. In the second class of approaches, such as in [56] [57] [58], scheduling is performed in two steps. The first step consists of resource allocation, which determines the number of resources allocated to each user. The resource allocation step is followed by the resource assignment step, which determines which resources are assigned to

each user. This class of scheduling algorithms are specifically designed for delay sensitive applications and does not provide a priority differentiation between delay sensitive and best-effort flows.

3. The third approach is the adaptation of TDMA strategies for OFDMA systems. Scheduling rules designed for single carrier systems such as the PF [34], M-LWDF [40] and EXP-PF [43] are adapted for an OFDMA system by calculating these rules on each resource. This adaptation is referred to as an OFDMA/TDMA strategy. These scheduling rules are analyzed by [42] for delay sensitive applications over an LTE system. According to [42], M-LWDF is the best scheduling rule for delay sensitive applications in terms of fairness and efficiency. A very good survey on these scheduling strategies for LTE is provided in [59]. As each of these scheduling rules are based on the proportional fair rule, the calculation of these scheduling metrics on each PRB allows the exploitation of multi-user time and frequency diversities. The complexity of the OFDMA/TDMA approach grows linearly with the number of users and resources. Thus, it can be implemented in real time. However, for delay sensitive traffic these scheduling rules cannot provide fairness for users with relatively low SINR [8].

In this Chapter, the following issues of the third class of strategies are addressed:

- Intra-class fairness issues for delay sensitive traffic: Scheduling rules for delay sensitive traffic consider the ratio of instantaneous channel quality and time-averaged throughput (proportional fair rule) along with either the linear [40], logarithmic [60] or exponential [43][60] function of the Head of Line (HoL) delay [61]. The HoL delay refers to the amount of time packets reside in the buffer and is also known as the sojourn time. It is important to note that video is delay sensitive traffic, hence packets arriving late are generally not useful at the receiver. Therefore, packets are associated with a predefined HoL delay bound and packets violating the delay bound are dropped from the queue. The utilization of HoL delay and the proportional fair rule in the scheduling decisions are not sufficient to avoid delay bound violation of flows having lower channel quality. Video traffic exhibits highly variable bit rate characteristics, *i.e.*, the instantaneous peak rate is higher than the average rate. Lower channel quality video flows exhibiting peak instantaneous rate have high probability of delay bound violation mainly because of the proportional fair rule in the scheduling decisions. In other words, these scheduling rules achieve higher HoL delay for the packets of flows having higher average rate and lower channel quality. On the other hand, flows having good channel quality and lower average rate are scheduled way before their delay bound. The probability of delay

bound violation of the flows exhibiting lower channel quality and higher average rate is very high which results in an unfair system.

- Inter-class fairness issues: In the literature [59], composite scheduling rules serve the best effort traffic by using the classical proportional fair rule, *i.e.*, ratio of instantaneous channel quality to the time-averaged throughput [34] [62] [63] [64]. They prioritize delay sensitive traffic by considering either the logarithmic, exponential or linear function of the HoL delay. However, such composite scheduling strategies result in a higher priority difference between the delay sensitive and best-effort traffic classes. In other words, the higher the difference among the scheduling priorities of traffic classes, the lower will be the multi-user channel diversity exploitation. In LTE, multi-user channel diversity has more significance since it is a multi-carrier system which allows multi-user diversity exploitation in the time and frequency domain.

The aforementioned issues are addressed by using the concept of fuzzy logic priority [65] where flow's delay urgency (ratio of packet's HoL delay and delay bound) is utilized along with the time-averaged channel quality. Instead of exploiting the time-averaged throughput and the linear, logarithmic or exponential function of the HoL delay, a fuzzy function of the HoL delay coupled with time-averaged channel quality is used in the scheduling decisions. The HoL delay along with the time-averaged channel quality is processed by a fuzzy proactive controller. Further, whenever a flow suffers a delay bound violation, the scheduler reacts to this event and increases the priority of that flow. The delay bound violation input is processed by a fuzzy reactive controller. The main goal of the composite scheduling rule is to consider the service needs for delay sensitive as well as the best-effort traffic. In the previous chapter, the scheduling rule considers only the video traffic. In this chapter, the scheduling rule and scenarios are extended to handle more than one delay sensitive traffic types. Furthermore, the main goal of the proposed composite scheduling rule is to balance the probabilities of QoS violation of the delay sensitive as well as the best-effort traffic types.

A block diagram representing the proposed FCS is given in Figure 4.1. The scheduling metric comprises a time-domain priority component based on reactive and proactive controllers and a frequency domain priority based on detailed information on instantaneous CQI feedback per PRB. In order to dynamically adjusts the priority level between best-effort and delay sensitive flows, a fuzzy based DRC (discussed in Section 4.3.3), is introduced as shown in Figure 4.1. Intra-class fairness (fairness in terms of achieved QoS among the flows within each of the traffic classes) is provided by the fuzzy proactive and reactive controllers whereas inter-class fairness (priority differentiation between the delay sensitive and best-effort flows) is provided by the DRC. In fuzzy logic, the *output*

fuzzy set is defined as the range of all possible output values that can be assigned to a fuzzy controller. The output of the controller lies within the *output fuzzy set*. The larger the *output fuzzy set*, the higher the priority of the controller. In the proposed scheduling framework, each traffic class has its own *output fuzzy set*. A fixed *output fuzzy set* is assigned to the delay sensitive traffic class, whereas the *output fuzzy set* of the best-effort traffic is dynamically adjusted. The *output fuzzy set* of the best-effort traffic class is set by the DRC based on the latency (packet's HoL delay) and QoS violation of the delay sensitive flows as shown in the figure. The higher the latency and QoS violations of the delay sensitive flows, the lower the *output fuzzy set* of the best-effort traffic. The final priority on each PRB is a function of the time and frequency domain priority metrics as shown in Figure 4.1.

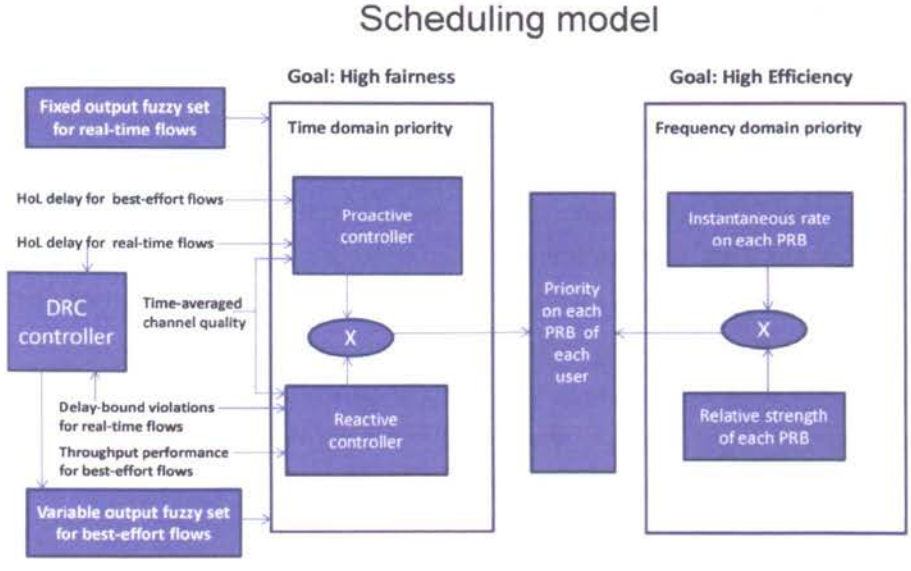


FIGURE 4.1: Time and frequency domain models of the FCS scheduling framework.

The remainder of the chapter is organized as follows. Section 4.2 presents the considered system model and the problem statement. Section 4.3 presents the details of the fuzzy-logic based scheduling strategy. Section 4.4 presents the performance evaluation of the proposed approach. In particular, the solutions considered as benchmark for the assessment of the proposed scheduling algorithm are presented in Section 4.4.1, whereas the simulation set-up is presented in Section 4.4.2; results are presented and discussed in Section 4.4.3. Conclusions are drawn in Section 4.5.

TABLE 4.1: Mathematical symbols utilized in Chapter 4.

Symbol	Explanation
i, φ, n	User/flow index, PRB index, Current scheduling epoch.
$\text{PRF}_{i,\varphi}^{(n)}$	Priority function of user/flow i on PRB φ.
$\chi_{i,\varphi}^{(n)}$	Instantaneous Channel quality on PRB φ also known as the normalized subband spectral efficiency.
$\chi_i^{(n)}$	Average PRB quality, in terms of spectral efficiency, of flow i at scheduling instant n.
$\overline{\chi}_i^{(n)}$	The normalized average wideband channel quality of user i over the moving average window of size n_c epochs.
χ_{max}	A constant, *i.e.*, the spectral efficiency (5.5547 bits/sec/Hz) corresponding to the maximum CQI feedback.
M_{PRB}	The number of PRBs available for allocation at each scheduling epoch.
$H_i^{(n)}$	HoL packet delay of user/flow i at scheduling epoch n.
$\overline{H_i^{(n)}}$	Average HoL packet delay of all the delay sensitive flows at scheduling epoch n.
H_{max}	Maximum HoL delay (target HoL delay) budget of user i's packet.
$\text{plr}^{(n)}{}_i$	Packet loss ratio of user i at scheduling instant n calculated over the moving average transmission window t_w.
$\text{plr}_{\text{thr}_i}$	Maximum tolerated PLR for user i.
$P_{\text{transmit}_i}^{(m)}$ and $P_{\text{drop}_i}^{(m)}$	Number of transmitted and dropped packets of user i over the moving average transmission window t_w.
$R_i^{(n)}$	Throughput achieved by flow i at scheduling instant n.
$R_{t_w}^{(n)}$	System throughput over the moving average transmission window t_w.
$R_{i,\text{ave}}^{(n)}$	Exponential time-averaged throughput (over the window of size n_w) at scheduling instant n.
I	Total number of flows in the system. It is the sum of delay sensitive $I_{\text{delaysensitive}}$ and best-effort $I_{\text{best-effort}}$ flows.
$\mathbb{I}\{\text{plr}_i^{(n)} \leq \text{plr}_{\text{thr}}\}$	An indicator function, equal to 1 if its argument is true.
$A_i^{(n)}$ and $B_i^{(n)}$	Fuzzy logic proactive controller inputs. $A_i^{(n)}$ is a ratio of HoL delay and the target HoL delay. $B_i^{(n)}$ is a weighted sum of $A_i^{(n)}$ and $\overline{\chi}_i^{(n)}$.
μ_p and μ_r	Defuzzified outputs of the fuzzy proactive and reactive controllers.
$V_{i,\text{delaysensitive}}^{(n)}$	QoS violation input for delay sensitive flows in terms of packet loss ratio and tolerated packet loss ratio threshold.
$V_{i,\text{best-effort}}^{(n)}$	QoS violation input for best-effort flows in terms of minimum rate requirements and achieved time-averaged throughput.
$C_i^{(n)}$ and $D_i^{(n)}$	Fuzzy logic reactive controller inputs. $C_i^{(n)}$ is a function of QoS violation inputs and $D_i^{(n)}$ is a function of $C_i^{(n)}$ and $\overline{\chi}_i^{(n)}$.
$\mu_i^{(n)}$	Time domain priority of flow i at scheduling instant n.
$\Gamma_{i,\varphi}^{(n)}$	Frequency domain priority of flow i on PRB φ at scheduling instant n.
$\theta_{i,\varphi}^{(n)}$	Relative strength of user i on PRB φ.
α_t and α_f	Fairness (Exponential parameter of the time domain priority) and Efficiency (Exponential parameter of the frequency domain priority) control parameters.
$\mu_{r_{\text{best-effort}}}$ and $\mu_{p_{\text{best-effort}}}$	Defuzzified outputs of the fuzzy reactive and proactive controllers for the best-effort traffic class.
μ_{max}	Defuzzified output of the DRC controller.
$E^{(n)}$ and $F^{(n)}$	Inputs of the DRC controller. $E^{(n)}$ is a function of normalized HoL delay of all the delay sensitive flows. $F^{(n)}$ is a function of normalized QoS violations of the all the delay sensitive flows.
a_i and b_i	Tunable parameters used in the exponential and Log rule.
γ_i	A constant whose value is adjusted to account for different delay and packet loss rate requirements of different flows. It is utilized in the M-LWDF and the EXP/PF rule.
$N_{Q_i}^{(n)}$	Number of packets residing in the queue of flow i at scheduling instant n.

4.2 System model and problem statement

A multiuser downlink SISO LTE / LTE-A system is considered. The single cell scenario comprises an eNodeB MAC scheduler responsible for allocating PRBs to different users in the cell. Each user i is assigned a buffer at the eNodeB and packets of different traffic classes are streamed into the buffer of the eNodeB. For delay sensitive traffic, video and VoIP traffic (the scheduling framework can accommodate all LTE service classes) are considered, whereas for best-effort traffic Constant Bit Rate (CBR) traffic is considered. The packets of each traffic class entering the buffer are time stamped by the scheduler. Packets of delay sensitive traffic are dropped from the buffer if the HoL packet delay is longer than the target HoL delay bound. The main QoS parameters for video and VoIP flows are the HoL packet delay and the PLR, whereas throughput is the important QoS parameter for the flows of best-effort traffic. It is important to note that the HoL delay as well as the target delay are assigned for best-effort flows. However, since best-effort traffic is delay tolerant, therefore packets violating the target HoL delay are not dropped from the buffer. A CQI feedback mechanism is utilized, where each user feedbacks information about the channel quality on each PRB. Due to the adoption of adaptive modulation and coding (AMC) in Long-Term Evolution (LTE), each CQI value corresponds to a specific value of spectral efficiency for each PRB.

At scheduling epoch n, the normalized time-averaged wideband spectral efficiency, $\overline{\chi}_i^{(n)}$, of user i over the moving average window of size n_c is defined as:

$$\overline{\chi}_i^{(n)} = \frac{1}{\chi_{\max}} \left[\frac{1}{n_c} \sum_{k=n-n_c}^{n} \chi_i^{(k)} \right] \tag{4.1}$$

with

$$\chi_i^{(n)} = \frac{1}{M_{\text{PRB}}} \sum_{\varphi=1}^{M_{\text{PRB}}} \chi_{i,\varphi}^{(n)} \tag{4.2}$$

where $\chi_i^{(n)}$ is the average PRB spectral efficiency of user i at scheduling instant n and $\chi_{i,\varphi}^{(n)}$ is the instantaneous subband spectral efficiency of user i at PRB φ. $\chi_{\max}$ is a constant, *i.e.*, the spectral efficiency (5.5547 bits/sec/Hz) corresponding to the maximum CQI feedback, and M_{PRB} is the number of PRBs available for allocation at each scheduling epoch.

Given the available information about:

- the HoL packet delay for each flow $H_i^{(n)}$;
- the channel quality of each flow on each PRB, hence the resulting spectral efficiency $\chi_{i,\varphi}^{(n)}$;
- the tolerated delay bound $H_{\max}$;
- the QoS performance of the delay sensitive flows in terms of packet loss ratio, $\text{plr}_i^{(n)}$ and of the best-effort flows in terms of time-averaged throughput $R_{i,\text{ave}}^{(n)}$;

the scheduling problem is defined as: *How to allocate to the different users the M_{PRB} PRBs in each scheduling interval in order to fulfill the QoS requirements of each of the flows from different traffic classes so that a good trade-off between fairness and efficiency is achieved.*

In order to mathematically formulate the problem, following parameters are defined:

$R_i^{(n)}$: Throughput achieved by flow i at scheduling instant n.

I: Total number of flows in the system. It is the sum of delay sensitive $I_{\text{delay-sensitive}}$ and best-effort $I_{\text{best-effort}}$ flows.

$\text{plr}_i^{(n)}$: The packet loss ratio of flow i at scheduling instant n calculated over the moving average transmission window t_w:

$$\text{plr}_i^{(n)} = \frac{\sum_{m=n-t_w}^{n} P_{\text{drop}_i}^{(m)}}{\sum_{m=n-t_w}^{n} \left(P_{\text{transmit}_i}^{(m)} + P_{\text{drop}_i}^{(m)} \right)} \tag{4.3}$$

where

$P_{\text{transmit}_i}^{(m)}$: Number of transmitted packets of flow i over the moving average transmission window t_w.

$P_{\text{drop}_i}^{(m)}$: Number of dropped packets of flow i over the moving average transmission window t_w.

The main goal of the scheduler is to maximize the system throughput $R_{t_w}^{(n)}$, subject to the QoS constraints of the delay sensitive flows, over the moving average transmission window t_w:

$$R_{t_w}^{(n)} = \max\left(\sum_{i=1}^{I} \sum_{m=n-t_w}^{n} R_i^{(m)} \right) \tag{4.4}$$

subject to

$$\frac{1}{I_{\text{delay-sensitive}}} \sum_{i=1}^{I_{\text{delay-sensitive}}} \mathbb{I}\{\text{plr}_i^{(n)} \leq \text{plr}_{\text{thr}}\} = 1 \tag{4.5}$$

where

$\mathbb{I}\{\text{plr}_i^{(n)} \leq \text{plr}_{\text{thr}}\}$ is an indicator function equal to 1 if its argument is true, *i.e.*, when the packet loss rate of flow i is lower or equal than the threshold value plr_{thr}. If the packet loss rate exceeds the threshold then the indicator function is 0. It is important to note that fairness for delay sensitive traffic is guaranteed when the PLR over a short moving average window [47], for instance one second, is below the prescribed threshold for each of the delay sensitive flows in the system. As noted in [48], when the scheduler achieves short-term fairness then the long-term fairness is guaranteed.

The optimal solution of the above problem is not possible without restrictive assumptions on the arrival process of the traffic and changes in channel quality. Therefore, a novel scheduling framework relying on fuzzy logic is proposed. Fuzzy logic is ideally suited for problems where a definite mathematical solution is unavailable. The information about the changes in the radio channel and the traffic rate of each user is uncertain. Fuzzy logic can deal with such situations because of its capability to make approximate reasoning. In the proposed scheduling strategy, each PRB is assigned to the user maximizing a defined metric. The proposed metric comprises a PRB independent part and a PRB-specific part. The PRB independent part calculated for a user describes the "urgency" of an assignment as time-domain priority, whereas the PRB-specific part describes the instantaneous channel quality of the PRB and its relative quality versus other PRBs.

4.3 Fuzzy Composite Scheduling Framework

The FCS framework consists of fuzzy proactive, reactive and DRC controllers. It is important to note that the designs of the proactive and reactive controllers are same. The proactive controller processes the HoL delay whereas the reactive controller processes the QoS violation. The design of the three fuzzy controllers are reported in the following:

4.3.1 Proactive controller

The goal of the proactive controller is to avoid delay bound violations. In order to consider the delay urgency in a dynamic wireless environment, a novel concept is proposed

which comprises a utilization of time-averaged channel quality over a small moving window by using the average wideband spectral efficiency $\overline{\chi}_i^{(n)}$, defined in equation (4.1). The proactive controller processes two inputs. One of these is the HoL packet delay $H_i^{(n)}$ normalized to the maximum tolerated HoL delay $H_{\max}$ of each traffic class:

$$A_i^{(n)} = \begin{cases} \frac{H_{i,\text{VoIP}}^{(n)}}{H_{i,\max}}, & \text{if } i \in \text{VoIP} \\ \frac{H_{i,\text{Video}}^{(n)}}{H_{i,\max}}, & \text{if } i \in \text{Video} \\ \frac{H_{i,\text{best-effort}}^{(n)}}{H_{i,\max}}, & \text{if } i \in \text{best} - \text{effort} \end{cases} \tag{4.6}$$

The goal of the controller is to be proactive for any possible delay violations, hence the second input is designed as the weighted sum of the normalized delay and the normalized average channel quality. It is mathematically defined as:

$$B_i^{(n)} = 0.5(1 - A_i^{(n)}) + 0.5(\overline{\chi}_i^{(n)}) \tag{4.7}$$

The rationale behind the proactive controller's inputs is discussed in Section 4.3.1.1.

In fuzzy logic, the input membership function represents the magnitude of the inputs which are mapped to the output membership function through a set of rules [65]. The membership functions can be linear, exponential, bell shaped or any other shape according to the system requirements. According to [42] the M-LWDF scheduling rule, linear function of the HoL packet delay, outperforms the EXP-PF scheduling rule which is an exponential function of HoL packet delay. Therefore, linear membership functions for the proactive and reactive controllers are selected. The graphical representation of the input membership functions is shown in Figure 4.2. The same input membership functions are used for both the inputs. It is important to note that users with better channel quality result in a higher frequency domain priority on each PRB φ, as there will be a higher number of PRBs with better channel quality. Therefore, the time domain priority should be higher for users with higher normalized HoL packet delay and lower normalized channel quality.

Now the flexibility of fuzzy logic is utilized by mapping the input membership functions to the output membership function through a set of rules. Let μ_p be the output of the proactive controller (defuzzified proactive priority value), the fuzzy rules for the proactive controller are:

1. If $A_i^{(n)}$ is *low* AND $B_i^{(n)}$ is *low* THEN μ_p is *medium*
2. If $A_i^{(n)}$ is *low* AND $B_i^{(n)}$ is *high* THEN μ_p is *low*

3. If $A_i^{(n)}$ is *high* AND $B_i^{(n)}$ is *low* THEN μ_p is *high*

4. If $A_i^{(n)}$ is *high* AND $B_i^{(n)}$ is *high* THEN μ_p is *medium*

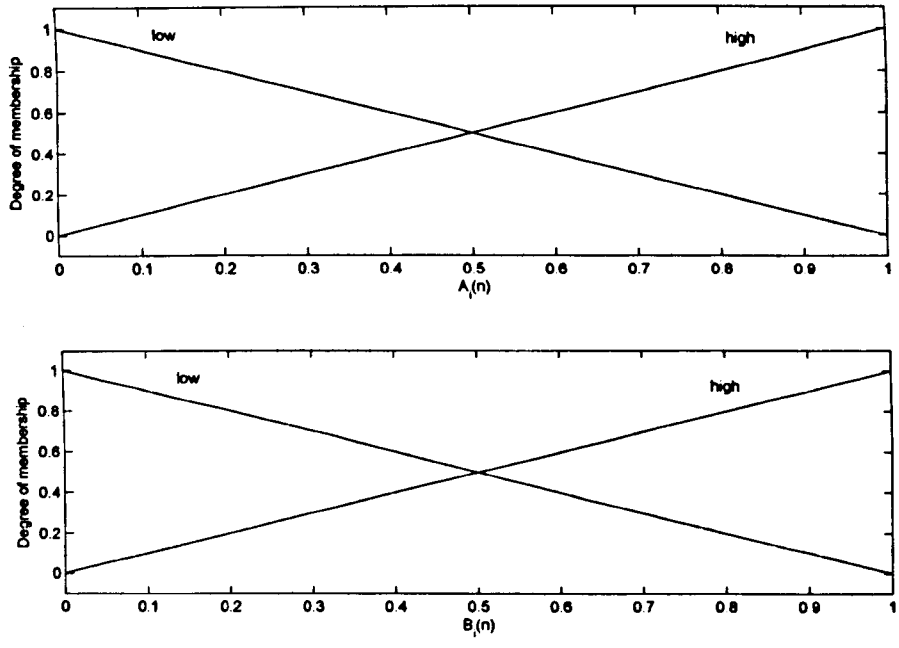


FIGURE 4.2: Input membership functions of the proactive and reactive controllers.

where *low*, *medium* and *high* are the output membership functions as shown in Figure 4.3 and μ_p is the crisp output which along with the reactive controller output quantifies the time domain priority of each user. The main motivation of using the *low*, *medium* and *high* output membership functions is to prioritize flows suffering from lower channel quality and higher HoL delay. The priority of the users with relatively good channel quality increases from low to medium as the HoL delay increases. On the other hand, the priority of users with lower channel quality increases from medium to high. Therefore, fairness is incorporated in the scheduling decisions through the output membership functions and rules of the fuzzy controllers. The main goal of the frequency domain priority is to improve the system efficiency whereas the time domain priority provides fairness through fuzzy proactive and reactive controllers.

The *output fuzzy set* of the membership functions, shown in Figure 4.3, determines the traffic priority of each traffic class. It is important to note that μ_p lies within the *output fuzzy set*. The proactive priority, μ_p, as a function of the inputs $A_i^{(n)}$ and $B_i^{(n)}$ is shown in Figure 4.4.

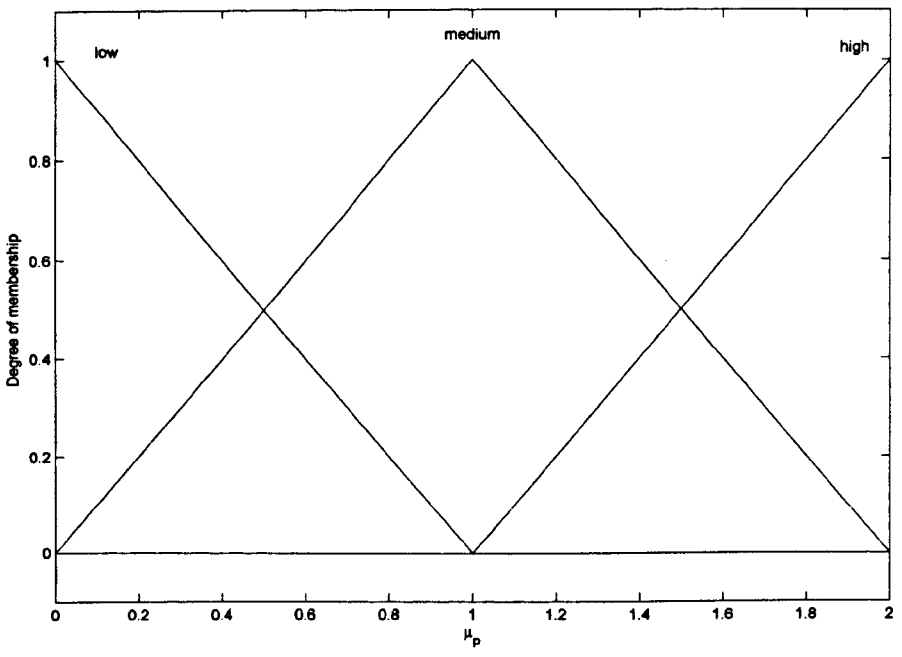


FIGURE 4.3: Output membership functions low, medium and high of the proactive controller. Proactive controller output μ_p lies within the *Output fuzzy set.*

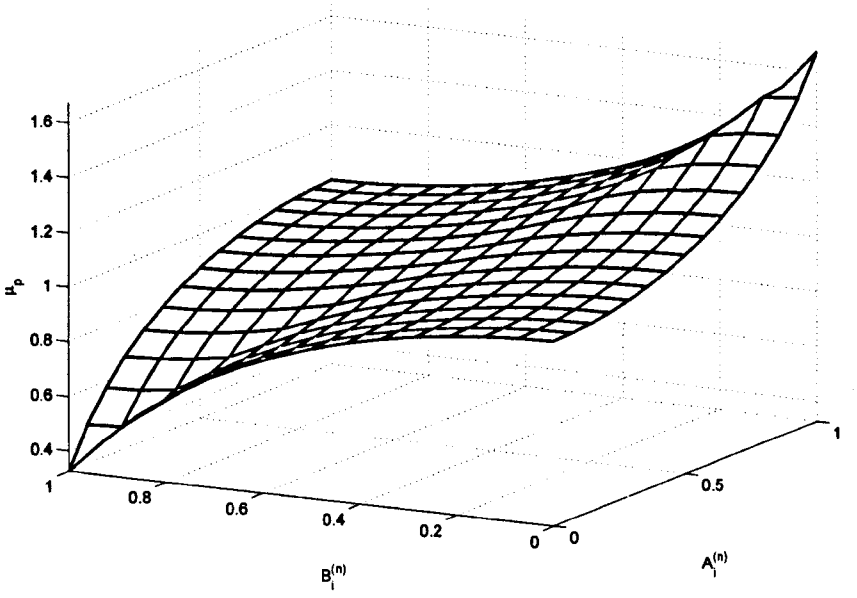


FIGURE 4.4: Proactive controller output, μ_p, w.r.t the inputs.

The steps involved in producing a crisp output in the fuzzy logic system are described below.

1. Fuzzification. This is the process of converting fuzzy input values into a degree of membership for each linguistic term. Each linguistic term, *high*, *medium* and *low*, characterizes a membership function. For instance the proactive controller inputs, $A_i^{(n)} = 0.8$ and $B_i^{(n)} = 0.3$, as shown in Figure 4.5, are fuzzified by the input membership functions *low* and *high*. In the figure, the four rows are the four rules of the proactive controller. Rule one comprises only *low* membership function, therefore input $A_i^{(n)} = 0.8$ and $B_i^{(n)} = 0.3$ are fuzzified by the *low* membership function as shown in the figure.

2. Fuzzy inference. This is the core process of the fuzzy logic system, comprising a mapping from a given input to an output using the membership functions and logical operators in the *if-Then-Else* rules. According to Figure 4.5, the *and* logical operation is performed, according to which the minimum of the two fuzzified inputs is mapped to the output membership function. The result of the fuzzy inference process is the degree of the output membership functions fulfilled by the logical operations between the fuzzified inputs. The result is the truncated output membership functions as shown in the third column of Figure 4.5.

3. "Defuzzification" and production of the final "crisp" output. The crisp proactive priority output μ_p produced is shown in Figure 4.5. The output of each rule is combined to give the final fuzzy set, as shown in the fifth row and third column in Figure 4.5. The defuzzification process is simply the centroid calculation on the final fuzzy set as shown in Figure 4.5.

4.3.1.1 Rationale

The inputs $A_i^{(n)}$ and $B_i^{(n)}$ are a function of the HoL delay, hence the system is made more proactive for any possible delay violations. The second input, weighted sum of the normalized HoL delay and time averaged channel quality, enhances the system fairness. For instance consider two users - user 1 and 2 - having normalized average channel quality of 0.8 and 0.6 respectively, and normalized HoL packet delay of 0.4 and 0.8 respectively. If the second input is selected as a simple function of the average channel quality, *i.e.* $B_i^{(n)} = \overline{\chi}_i^{(n)}$, then the output of the proactive controller, μ_p (Figure 4.4), for user 1 and 2 is 0.844 ($A_i^{(n)} = 0.4$, $B_i^{(n)} = 0.8$) and 1.05 ($A_i^{(n)} = 0.8$, $B_i^{(n)} = 0.6$) respectively. The difference in the proactive priority of the two users is $1.05 - 0.844 = 0.206$. On the other hand, if the weighted sum equation (4.7) is used, then the output for user 1 and

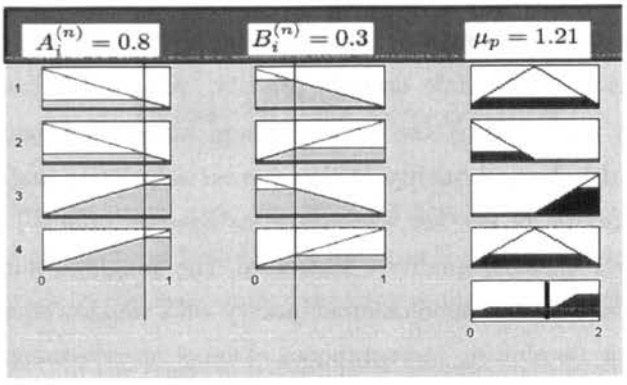


FIGURE 4.5: Scheduling rules of the proactive controller.

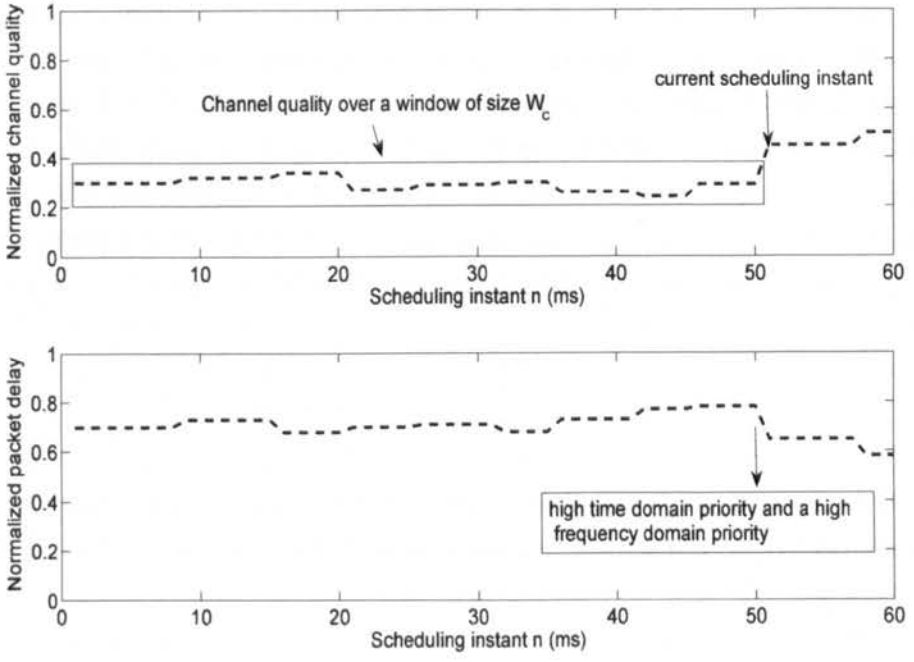


FIGURE 4.6: Rationale of the proactive controller design.

2 is 0.872 ($A_i^{(n)} = 0.4$, $B_i^{(n)} = 0.6$) and 1.14 ($A_i^{(n)} = 0.8$, $B_i^{(n)} = 0.7$), with a difference in proactive priority of 0.268. A higher priority with weighted sum equation quantifies the urgency in the service needs of user 2 having relatively higher packet delay and lower channel quality. Therefore the system is more sensitive to the HoL delay. If the instantaneous channel quality of the user improves, the system exploits it. For instance consider Figure 4.6 where the channel quality increases at the *current scheduling instant*; the result is a higher time domain priority, quantifying lower time-averaged channel quality over a window of size n_c epochs and higher HoL packet delay. Because of the increase in the channel quality at the *current scheduling instant*, the frequency domain priority (function of current instantaneous channel quality) also increases with PRBs having better channel quality. Therefore, the weighted sum of the normalized HoL delay and the time averaged channel quality with weights equal to 0.5 makes the system opportunistic (exploiting instantaneous channel improvements) and delay aware.

4.3.2 Reactive controller

Delay sensitive applications can tolerate packet losses if they are below a given threshold. To provide fairness in multimedia traffic, packet losses should be kept below a given threshold for all users. The goal of the reactive controller is to distribute the packet losses proportionally equal across all the users. In order to define the inputs of the reactive controller, the packet loss ratio $\text{plr}_i^{(n)}$ of user i given in equation (4.3) is utilized. The packet loss ratio can easily be calculated by using the number of dropped and transmitted packets over a small transmission window. The design of the reactive controller is similar to the proactive controller except that the fuzzy inputs are based on the packet loss rate over a moving average transmission window. A window size of 1 s is considered in the simulation study. The amount of QoS violation in terms of packet loss ratio and tolerated packet loss threshold, plr_{thr}, of user i at scheduling instant n is:

$$V_{i,\text{delay-sensitive}}^{(n)} = \frac{\text{plr}_i^{(n)}}{\text{plr}_{i,\text{thr}}} \tag{4.8}$$

The QoS parameter for the delay sensitive traffic is the packet loss ratio whereas for the best-effort flows, the QoS performance parameter is the ratio of minimum rate required to the achieved time-averaged throughput .

$$V_{i,\text{best-effort}}^{(n)} = \frac{R_{\min}}{R_{i,\text{ave}}^{(n)}} \tag{4.9}$$

where $R_{i,\text{ave}}^{(n)}$ is the time-averaged throughput and $R_{\min}$ is the minimum rate requirement. The QoS violation input, $C_i^{(n)}$, for the reactive controller is:

$$C_i^{(n)} = \begin{cases} \frac{V_{i,\text{VoIP}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{if } i,j \in VoIP \\ \frac{V_{i,\text{Video}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{if } i,j \in Video \\ \frac{V_{i,\text{best-effort}}^{(n)}}{V_{j,\max}^{(n)}}, & \text{if } i,j \in \text{best} - \text{effort} \end{cases} \tag{4.10}$$

It is a requirement of the fuzzy logic system that the inputs of the fuzzy controller should lie within the input fuzzy set, *i.e.*, in between 0 and 1. Therefore, the input is normalized with respect to the flow having the maximum QoS violation, $V_{j,\max}^{(n)}$.

The second input, $D_i^{(n)}$, of the reactive controller is designed as the weighted sum of the normalized QoS violations and the normalized average channel quality. Mathematically it is defined as:

$$D_i^{(n)} = 0.5(1 - C_i^{(n)}) + 0.5(\overline{\chi}_i^{(n)}) \tag{4.11}$$

4.3.2.1 Rationale

The rationale behind the design of the reactive controller is same as that of the proactive controller discussed in Section 4.3.1.1. The input and output membership functions and the *Output fuzzy set* is same as that of the proactive controller. It is important to note that all the inputs (the HoL packet delay, the QoS violations and the time averaged channel quality) could have been utilized by designing a fuzzy priority scheme with three inputs. However, this increases the complexity of the system because, with three inputs, 8 rules and more than 3 output membership functions are required. A fuzzy logic system with two inputs is simpler in terms of implementation and processing. Therefore, the same fuzzy module is called for proactive ($A_i^{(n)}$ and $B_i^{(n)}$) and reactive ($C_i^{(n)}$ and $D_i^{(n)}$) inputs by using the same rules and membership functions. The rationale of the fuzzy controller design is shown in Figure 4.7. For instance consider a loaded system, with the normalized average channel quality of the two users shown in Figure 4.7. During peak traffic, the probability of delay violations is higher. The reactive controller output prioritizes the lower channel quality user in the time domain by considering the time-averaged channel quality over a moving window as shown by point "P1" in Figure 4.7. At point "P1", the time domain priority of the lower channel quality user is higher than for the good channel quality user. On the other hand, the frequency domain priority of the good channel quality user is higher (frequency domain priority is a function of

current instantaneous channel quality discussed in Section 4.3.5). When delay violation occurs for the user with lower channel quality, a further increase in the time domain priority reduces its delay violations. Therefore, under high load, fairness among the flows is achieved by distributing the packet loss rate proportionally according to their channel quality.

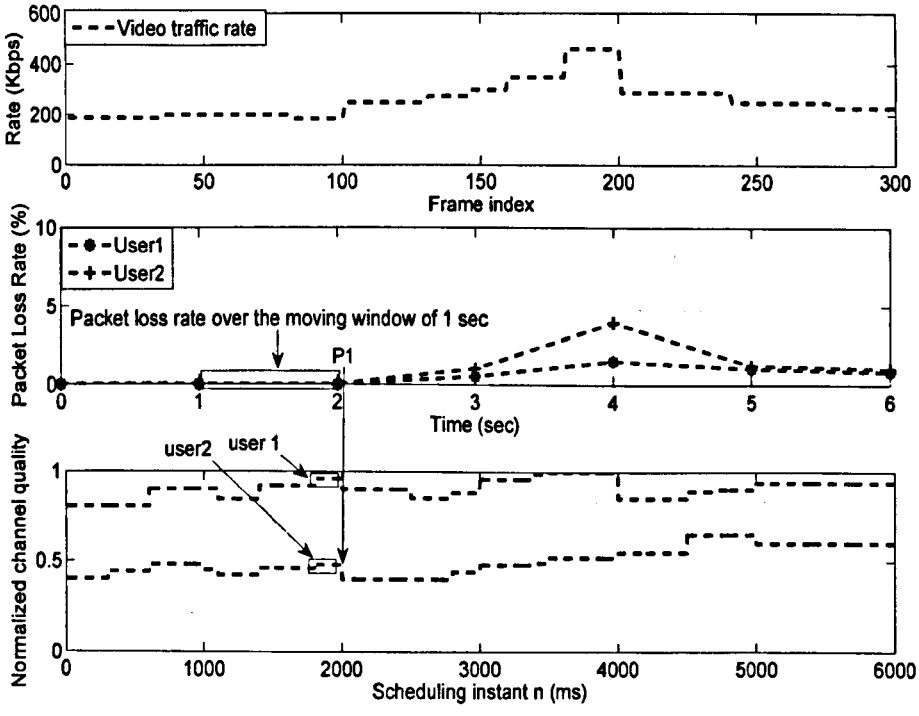


FIGURE 4.7: Rationale of the reactive controller design.

4.3.3 Dynamic Resource Controller

The best-effort traffic class is considered as the lowest priority class. Scheduling rules designed for delay sensitive traffic, such as in [66][67] (see the time utility functions of different traffic classes), give low scheduling priority to the best-effort flows. High priority differentiation between the delay sensitive and best-effort flows causes resource starvation for the best-effort flows [68] [69]. In FCS scheduling framework, inter-class traffic priority differentiation is provided by *output fuzzy set*. The *output fuzzy set* represents the range of all possible output values that can be assigned to the proactive and reactive controllers. The larger the *output fuzzy set*, the higher the priority of the controller. In order to dynamically prioritize flows belonging to best-effort traffic class, the *output fuzzy set* of the best-effort flows is adapted according to the QoS performance of the delay sensitive flows. The *output fuzzy set* of the delay sensitive traffic class is fixed, thus the amount of resource allocations between the delay sensitive and best-effort

flows is adaptable and controlled by the maximum limit of the *output fuzzy set*, μ_{max}, as given in equation (4.12).

$$\mu_{r_{\text{best-effort}}} = \mu_{p_{\text{best-effort}}} \in \{0, \mu_{\text{max}}\} \tag{4.12}$$

where $\mu_{r_{\text{best-effort}}}$ and $\mu_{p_{\text{best-effort}}}$ are the defuzzified outputs of the reactive and proactive controllers. As discussed in Sections 4.3.1 and 4.3.2, the design of both the controllers and the corresponding output fuzzy sets are the same. Flows from each traffic class utilize the same time domain priority by using the reactive and proactive controllers. The average delay and packet loss rate performance of the delay sensitive flows are used to determine the maximum limit of the *output fuzzy set* for the best-effort traffic. Mathematically, the average QoS parameters of the delay sensitive flows are:

$$E^{(n)} = \frac{1}{I_{\text{delay-sensitive}}} \sum_{i=1}^{I_{\text{delay-sensitive}}} \frac{H_i^{(n)}}{H_{\text{max}}} \tag{4.13}$$

$$F^{(n)} = \frac{1}{I_{\text{delay-sensitive}}} \sum_{i=1}^{I_{\text{delay-sensitive}}} V_{i,\text{delay-sensitive}}^{(n)} \tag{4.14}$$

where $I_{\text{delay-sensitive}}$ is the number of delay sensitive users, $E^{(n)}$ is the average normalized delay and $F^{(n)}$ is the average QoS violations of all the delay sensitive users. The input and output membership functions of the DRC controller are shown in Figure 4.8 and 4.9 respectively. The maximum limit, μ_{max} is set according to the following fuzzy rules:

1. If $E^{(n)}$ is *low* AND $F^{(n)}$ is *low* THEN μ_{max} is *high*
2. If $E^{(n)}$ is *high* AND $F^{(n)}$ is *low* THEN μ_{max} is *low*
3. If $E^{(n)}$ is *high* AND $F^{(n)}$ is *high* THEN μ_{max} is *low*
4. If $E^{(n)}$ is *low* AND $B^{(n)}$ is *high* THEN μ_{max} is *medium*

The input degree of membership is determined by the trapezoidal input membership functions. A lower average packet delay and loss rate causes rule 1 to have a higher degree of membership. Therefore, μ_{max} is maximum as given by the centroid of the highest area triangle membership function as shown in Figure 4.9. On the other hand, μ_{max} is set to minimum when a higher average HoL delay and packet loss rate causes the smallest area triangle to be defuzzified through rule 2 and rule 3. If the normalized average delay is lower and average PLR is higher than the medium area triangle is defuzzified as given in rule 4.

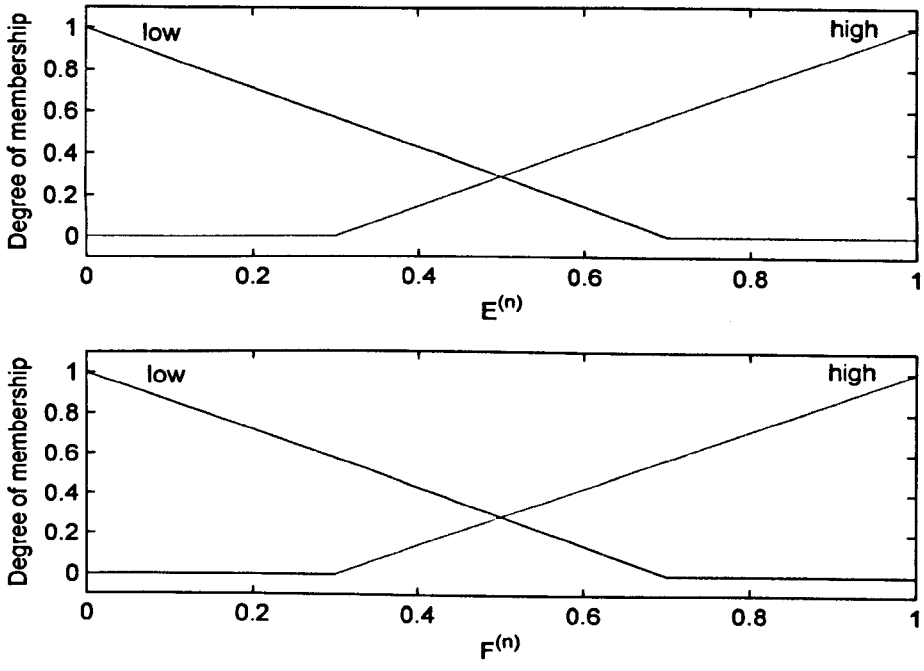


FIGURE 4.8: Input membership functions of the DRC controller.

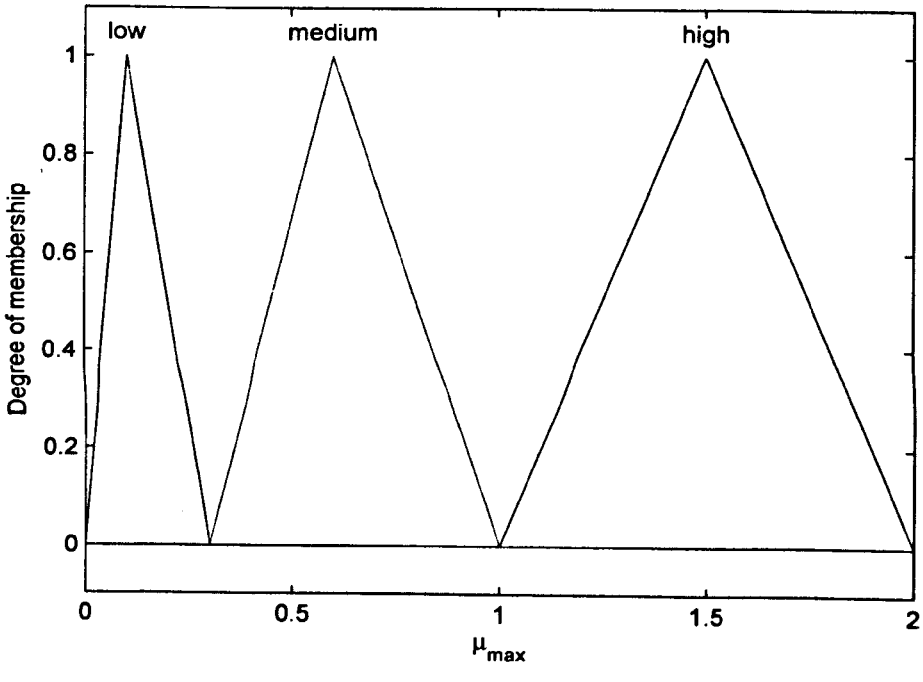


FIGURE 4.9: Output membership functions of the DRC controller

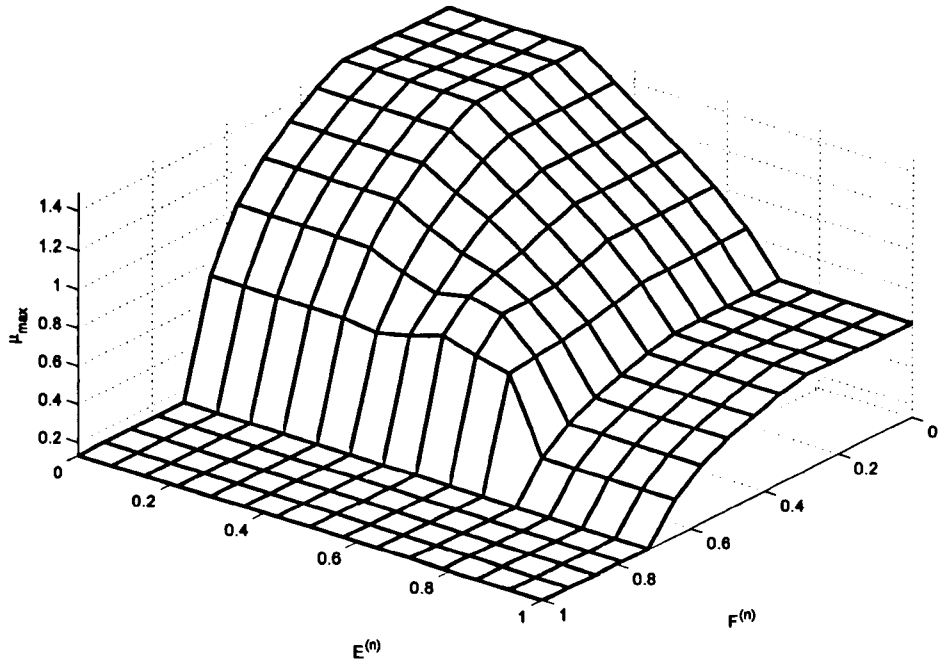


FIGURE 4.10: DRC controller's response to the normalized average delay and average PLR.

4.3.3.1 Rationale

The main rationale of utilizing DRC is to serve the following three goals:

- Utilization of delay tolerant nature of the best-effort traffic: According to the policy guidelines of the QoS architecture in the 3GPP standard, the resource allocation probability of the best-effort traffic class should be minimum in situations where the network becomes congested with delay sensitive traffic. When the traffic load reaches the network capacity, the increase in average packet's latency of the delay sensitive traffic decreases the maximum limit of the *output fuzzy set* for the best-effort flows as shown in Figure 4.10. Since best-effort traffic is delay tolerant, the decreased maximum limit of the *output fuzzy set* ensures delay sensitive traffic gets priority over best-effort traffic.

- Channel diversity exploitation: The main goal of the scheduler is to maximize the system throughput subject to maintaining the deadline violations below the prescribed threshold (equation (4.5)). At lower normalized average packet latency, the priority difference between the delay sensitive and best-effort flows is minimal. Hence, flows from different traffic classes are scheduled based on their QoS performance and channel quality.

- Utilization of same *output fuzzy set* for the DRC, proactive and reactive controllers: The prioritization of the delay sensitive flows *w.r.t* the best-effort traffic can be achieved by using the same *output fuzzy set* for the proactive, reactive and DRC controllers. When the *output fuzzy set* of these controllers are same then the increase in latency of the delay sensitive flows causes a reduction in the *output fuzzy set* of the best-effort traffic as shown in Figure 4.10. When the network becomes heavily congested then delay bound violations occur for the delay sensitive flows. The delay bound violation further reduces the *output fuzzy set* of the best effort traffic as shown in Figure 4.10. Thus, decreasing the resource allocation probability of the best-effort traffic.

4.3.4 Time domain priority

The proactive controller output, μ_p, and the reactive controller output, μ_r, define the time domain priority of the scheduling rule. Let $\mu_i^{(n)}$ be the final time domain priority which is the product of the output of both the controllers given as:

$$\mu_i^{(n)} = (\mu_p.\mu_r)^{\alpha_t} \tag{4.15}$$

where α_t is the time domain fairness parameter which enables the operator of the system to tune the fairness level. The higher the value of α_t, the higher will be the time domain priority of users suffering from relatively poor channel quality, higher HoL delay and higher QoS violations.

4.3.5 Frequency domain priority

The time domain priority, by utilizing past and current CQI feedbacks, considers the channel quality over a small window. The goal of the time domain priority is to control the fairness among the users. On the other hand, the goal of the frequency domain priority is to improve the system efficiency by considering only the current CQI feedback. Due to multipath propagation and interference from the neighboring users, there is a variable amount of fading on the PRBs of each user. Efficiency as well as fairness can be enhanced if this information is utilized. Information on the interference and multipath propagation can be obtained by employing the CQI feedbacks on each of the PRBs.

Hence, a parameter called relative strength of user i on PRB φ is utilized which is given as:

$$\theta_{i,\varphi}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{\chi_i^{(n)}} \tag{4.16}$$

where $\theta_{i,\varphi}^{(n)}$ gives information on the variable amount of fading on the PRBs of each user. If a user is experiencing a high interference on some of the PRBs, this factor assigns a lower weight to such PRBs. On the other hand, the PRBs with the best channel quality for a user will be assigned a higher weight thus fully utilizing the independent multi-user frequency selective fading. The frequency domain priority, $\Gamma_{i,\varphi}^{(n)}$, of user i on PRB φ is the product of channel quality and relative strength:

$$\Gamma_{i,\varphi}^{(n)} = [\chi_{i,\varphi}^{(n)}]^{\alpha_f} \theta_{i,\varphi}^{(n)} \tag{4.17}$$

The parameter α_t, in the time domain priority, controls the system fairness. On the other hand, efficiency is controlled by the parameter α_f. The trade-off between fairness and efficiency is varied by changing these two parameters. It has been shown in [60] that a good trade-off between fairness and efficiency can be achieved by defining a priority function which is the product of the logarithmic function of the time-domain priority and a linear function of the instantaneous rate on each PRB. The time domain priority used in the LOG rule in [60], is a function of the HoL packet delay. In this work, the proposed time-domain priority is derived from fuzzy logic and is a function of the user's HoL packet delay, time-averaged channel quality and packet loss rate. The final priority, $\mathrm{PRF}_{i,\varphi}^{(n)}$, of user i on PRB φ is a function of the logarithm of the time domain priority and it varies linearly with the frequency domain priority as given below:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \log(1 + \mu_i^{(n)})\Gamma_{i,\varphi}^{(n)} \tag{4.18}$$

User i^* is allocated a PRB φ satisfying the following rule:

$$i^* = \mathrm{argmax}\left(\mathrm{PRF}_{i,\varphi}^{(n)}\right) \tag{4.19}$$

It is important to note that state-of-the-art scheduling rules serve best-effort flows with the classical delay insensitive PF rule and prioritize the delay sensitive traffic by considering the HoL delay. The proposed scheduling rule uses the same priority equation (equation (4.18)) for all the traffic classes and prioritize each traffic class by adapting the *output fuzzy set.*

4.4 Performance evaluation

4.4.1 Benchmark scheduling rules

The proposed FCS scheduling rule is assessed and compared with the following state-of-the-art strategies:

- M-LWDF. According to the literature [42], among the existing rules this provides the best trade-off between fairness and efficiency for delay sensitive video traffic. The M-LWDF scheduling rule is given as:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} H_i^{(n)} \tag{4.20}$$

where γ_i is a constant whose value is adjusted to account for different delay requirements for different flows. The FCS scheduling is also compared with the virtual token version of the M-LWDF scheduling rule introduced in [70]. The scheduling rule considers the queue size of each flow instead of the HoL delay. The virtual token based scheduling rule (M-LWDFQ) is given as:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} N_{Q_i}^{(n)} \tag{4.21}$$

where $N_{Q_i}^{(n)}$ is the number of packets residing in the queue of user i's flow at the eNodeB. Both the scheduling rule exploits time diversity by using time-averaged throughput. In order to provide fairness, the M-LWDF rule uses the HoL delay whereas the M-LWDFQ uses the queue size of each flow.

- EXP-PF. This strategy schedules delay sensitive flows according to the following rule:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \gamma_i \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} \exp\left(\frac{\gamma_i H_i^{(n)} - \overline{\gamma_i H_i^{(n)}}}{1 + \sqrt{\overline{\gamma_i H_i^{(n)}}}} \right) \tag{4.22}$$

where $\overline{H_i^{(n)}}$ is the average HoL delay of the delay sensitive flows given as:

$$\overline{H_i^{(n)}} = \frac{1}{I_{\mathrm{delay-sensitive}}} \sum_{i=1}^{I_{\mathrm{delay-sensitive}}} H_i^{(n)} \tag{4.23}$$

The virtual token version of the EXP-PF rule introduced in [70] is also considered. In the modified version, the HoL delay is replaced by the queue size of each of the delay sensitive flows. The EXP-PFQ scheduling rule is given as:

- In order to address the problem of prioritizing different classes of traffic, [60] presents a scheduling strategy based on the log and the EXP rule. Mathematically these strategies are given as:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = b_i \log\left[1.1 + \left(a_i \frac{H_i^{(n)}}{H_{i,\max}^{(n)}}\right)\right] \chi_{i,\varphi}^{(n)} \tag{4.24}$$

$$\mathrm{PRF}_{i,\varphi}^{(n)} = b_i exp\left[\frac{\left(a_i \frac{H_i^{(n)}}{H_{i,\max}^{(n)}}\right)}{1+\sqrt{H_i^{(n)}}}\right] \chi_{i,\varphi}^{(n)} \tag{4.25}$$

where $b_i = \frac{1}{R_{i,\mathrm{ave}}^{(n)}}$ and a_i is a tunable parameter. The higher the value of this parameter, the higher will be the priority of the delay sensitive flows.

All the aforementioned priority rules are for the delay sensitive traffic. These rules calculate the priority for the best-effort traffic according to the PF rule given as:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} \tag{4.26}$$

4.4.2 Simulation Scenario

The performance of the proposed and benchmark scheduling algorithms is investigated by utilizing the simulation platform introduced in Chapter 2 (Section 2.3.1). The fuzzy controllers are designed by utilizing the Matlab's Fuzzy Logic Toolbox. Delay sensitive traffic is characterized by video and VoIP flows. In order to simulate best-effort traffic, CBR flows with the data rate of 400 Kbps are selected. On the other hand, 64 Kbps traffic with the threshold packet loss rate of 1 % and maximum delay budget of 100 ms is selected for VoIP users. These QoS parameters are selected according to LTE QCI (QoS Class Indicators) [72]. Video traffic is generated from a trace file [73] with the average and peak traffic rates of 530 Kbps and 1500 kbps respectively. The maximum delay budget for video packets is 200 ms whereas the threshold packet loss rate is 5 %. For non-critical video applications, 5 % packet loss rate corresponds to a PSNR (Peak Signal to Noise Ratio) of approximately 29 to 30 dB [71]. Therefore, 5% is considered as the threshold packet loss rate for an acceptable video quality. Table 4.2 reports the

TABLE 4.2: Simulation parameters - Downlink LTE scheduling for multi-class traffic.

PARAMETERS	VALUE
Bandwidth, Carrier frequency	5 MHz, 2.1 GHz
UE distribution, Cell radius	Uniform, 1 km
Channel	3GPP-TU (Typical Urban)
Pathloss model	Hata-Cost-231 model
Shadowing model	Log-normal shadow fading
HARQ	Up to 3 synchronous retransmissions
Channel Fading	Block Fading (1 ms)
UE speed	15 to 100 km/h (users moving independently at variable speed)
CQI averaging method	MIESM [19] [20]
$H_{\max}$, PLR_{thr} (video)	200 ms, 5% [71]
$H_{\max}$, PLR_{thr} (VoIP)	100 ms, 1% [72]
$H_{\max}$, $R_{\min}$ (best-effort)	300 ms, 200 Kbps [72]
Number of video, VoIP and best-effort users	18, 27 and 9
Average rate requirements for video, VoIP and best-effort users	530 Kbps, 64 Kbps and 400 Kbps
n_c (Time-averaged channel quality window) and n_w (Time-averaged throughput window)	100 ms

simulation parameters adopted for the LTE system and the wireless channel. 18 video flows (9.54 Mbps), 27 VoIP flows (1.728 Mbps) and 9 best-effort flows (3.6 Mbps) are simulated corresponding to a total average input traffic rate of 14.868 Mbps. The main motivations for such traffic distribution are the following:

- It has been reported in [2] that by 2015 approximately 66 % of mobile's traffic (in terms of petabytes per month) will be video and the proportion of VoIP traffic will be a minority. Therefore, the proportion of traffic in the simulation scenario is dominated by video followed by the best-effort and VoIP traffic. Specifically, a loaded network is simulated with 64 % video, 11 % VoIP and 25 % best-effort traffic (in terms of average input traffic at the eNodeB).

- The proposed scenario corresponds to an average input traffic rate of 14.868 Mbps. Sum rate maximization strategy is simulated to evaluate the channel utilization in terms of average spectral efficiency. The strategy maximizes the system throughput without considering the delay constraints. The average channel quality (in terms of SINR) of the users is set such that the total system throughput, sum throughput of all the flows, produced by the throughput maximization strategy [74] is 13.6 Mbps (2.72 bits/sec/Hz). This corresponds to a heavily loaded system where the input traffic is approximately 110 %, in terms of bits/sec/Hz, of the maximum

system capacity. The main goal is to study the fairness and efficiency performance of the proposed and benchmark scheduling rules when the delay bound and packet loss threshold constraints are considered.

The time-averaged channel quality is considered over the period of 100 TTIs, *i.e.* $n_c = 100ms$. All the benchmark scheduling rules utilize the time-averaged throughput. In order to have a fair comparison, the exponential averaging constant n_w is set to 100 ms. In the literature, the optimum size of the exponential averaging constant is from 100 to 1000, with the 100 being utilized in scenarios yielding high fairness in terms of throughput.

TABLE 4.3: Tunable parameters for FCS strategy.

Strategy	*output fuzzy set* for video	*output fuzzy set* for VoIP	α_t	α_f
FCS1	{0,2}	{0,2}	2	2
FCS2	{0,2}	{0,2.2}	2	2
FCS3	{0,2}	{0,2.2}	3	2

Table 4.3 reports the settings considered in the simulations for the parameters enabling the adjustment of the trade off between efficiency and fairness. The table reports three examples. In the first one (FCS1) the maximum limit of the *output fuzzy set* is the same for both VoIP and Video classes, i.e., video and VoIP traffic flows have the same prioritization. In the second case (FCS2) VoIP is prioritized by increasing the maximum limit of the *output fuzzy set* from 2 to 2.2. The time and frequency domain parameters α_t and α_f are set to 2. Finally, in the third case, the VoIP priority is kept same and the time-domain parameter is increased from 2 to 3. In all cases the maximum limit of the fuzzy set for best-effort flows is variable and changes according to the QoS violation of the delay sensitive flows as discussed in Section 4.3.3 (Figure 4.10).

4.4.3 Results and Discussion

First, the fairness and efficiency performance of the FCS strategy is analyzed according to the settings reported in Table 4.3. Next, the FCS strategy is compared with the benchmark scheduling strategies reported in Section 4.4.1.

Results in terms of packet loss rate (delay sensitive flows) and throughput (best-effort flows) for the proposed rule with different parameters are shown in Figure 4.11 and Figure 4.12, where users are arranged in a decreasing order of the channel quality. The user with lowest index has the best channel quality which then decreases with the increase in user index.

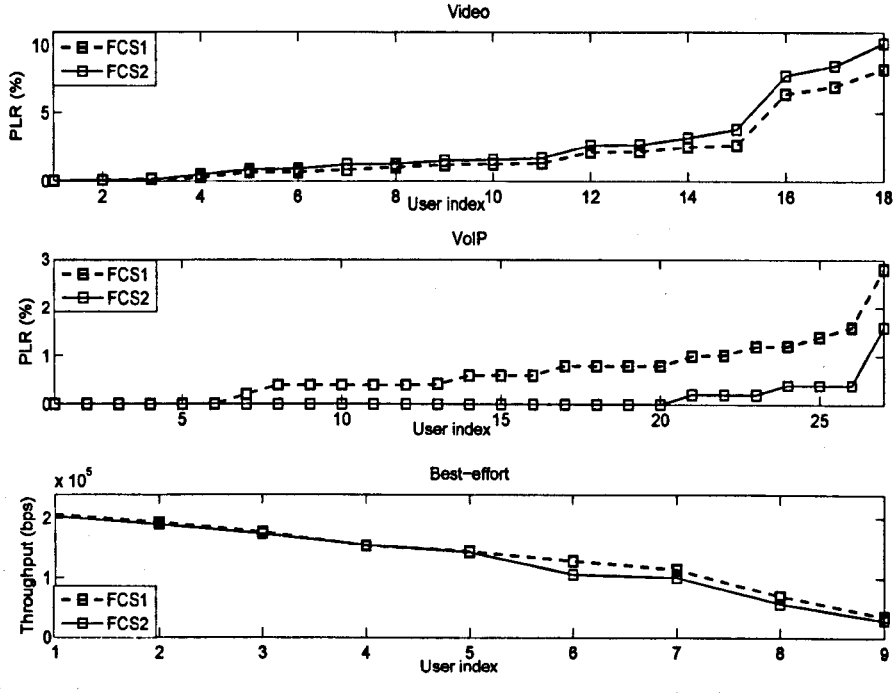


FIGURE 4.11: Performance of each traffic class with different *output fuzzy set.*

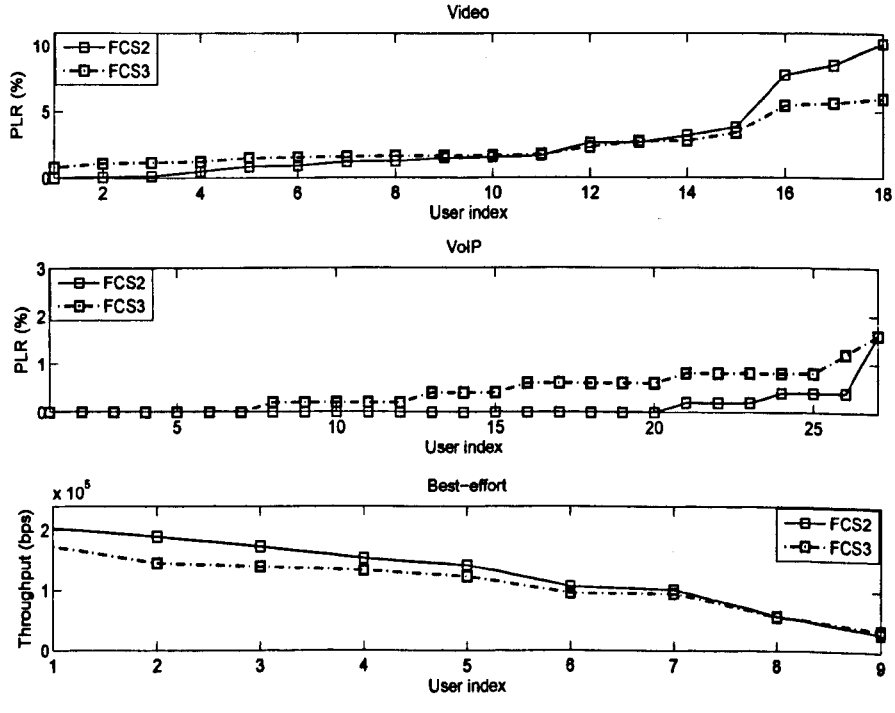


FIGURE 4.12: Performance of each traffic class with different time-domain priority.

Using the same prioritization (*output fuzzy set*) for video and VoIP (FCS1, Figure 4.11) results in a higher QoS violations for the VoIP flows , *i.e.*, 7 VoIP flows are violating the 1% PLR threshold. On the other hand, 3 video flows are violating the 5% PLR threshold.

In the second set of simulations (FCS2, Figure 4.11) there is a significant reduction in the PLR of the VoIP flows, *i.e.*, only 1 VoIP flow has a delay bound violation (PLR) of more than 1%. In FCS2 mode, the impact of the time-domain priority is higher for the VoIP flows. The increase in the HoL delay and PLR prioritized VoIP flows more than the video flows. The result is an increase in the PLR of the video flows as shown in Figure 4.11. There is also a slight reduction in the throughput of the best-effort flows. Higher limit of the fuzzy set, 2.2, for VoIP traffic serves well according to the QoS architecture of LTE as it is the highest traffic priority class. Any further increase in the maximum limit of the fuzzy set for VoIP traffic will penalize the video and best-effort flows.

Next, the impact of the time-domain parameter α_t is analyzed. An increase in the time-domain priority parameter (FCS3, Figure 4.12) allocates relatively more resources to the worst channel channel flows since time-domain priority is a fuzzy function of the HoL packet delay, PLR and time-averaged channel quality. It is important to note that the proportion of video traffic (18 flows with average rate requirements of 540 kbps) is more than the VoIP traffic (27 flows with rate requirements of 64 kbps). Therefore, increase in α_t results in a significant improvement for video flows as shown in Figure 4.12. In other words, the lower channel quality video flows are allocated more resources and as a result their PLR is reduced at the expense of a slight increase in the PLR of VoIP flows. There is also a marginal increase in the PLR of good channel video flows. According to the figure (FCS3, Figure 4.12), the worst served video flow has a PLR of approximately 5.4 % and the worst served VoIP flow suffers from a PLR of 1.6 %. Thus, the FCS3 mode results in an improved fairness performance for the delay sensitive flows. Under high load, the time-domain priority of the delay sensitive flows will always be higher than best-effort flows. Therefore, increase in α_t will further enhance the priority difference and results in a reduction in the throughput of best-effort flows.

Figures 4.13 and 4.14 analyze the performance of state-of-the-art scheduling rules for delay sensitive traffic and compare it with the FCS rule. Although M-LWDF is generally considered as the best scheduling rule for delay sensitive traffic [42], the PLR performance of the M-LWDF scheduling rule for low channel quality video flows is very poor. According to Figure 4.13, the PLR of the worst served user is as high as 20% and approximately 7 flows suffer from the QoS violation, *i.e.*, having PLR above the 5 % threshold. Higher QoS violations for video flows stems from the fact that the M-LWDF

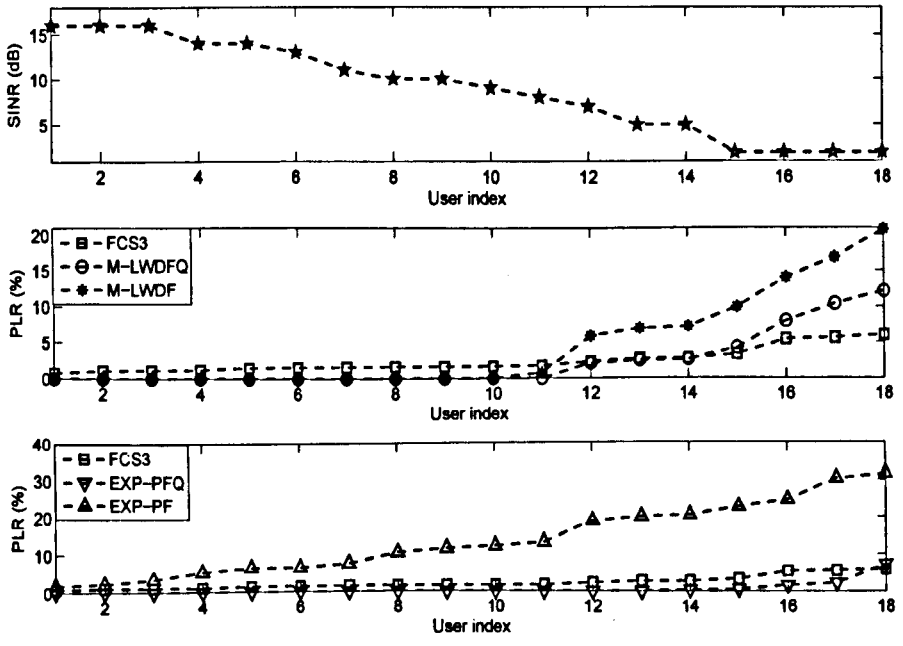


FIGURE 4.13: Performance of the Video flows for different scheduling rules.

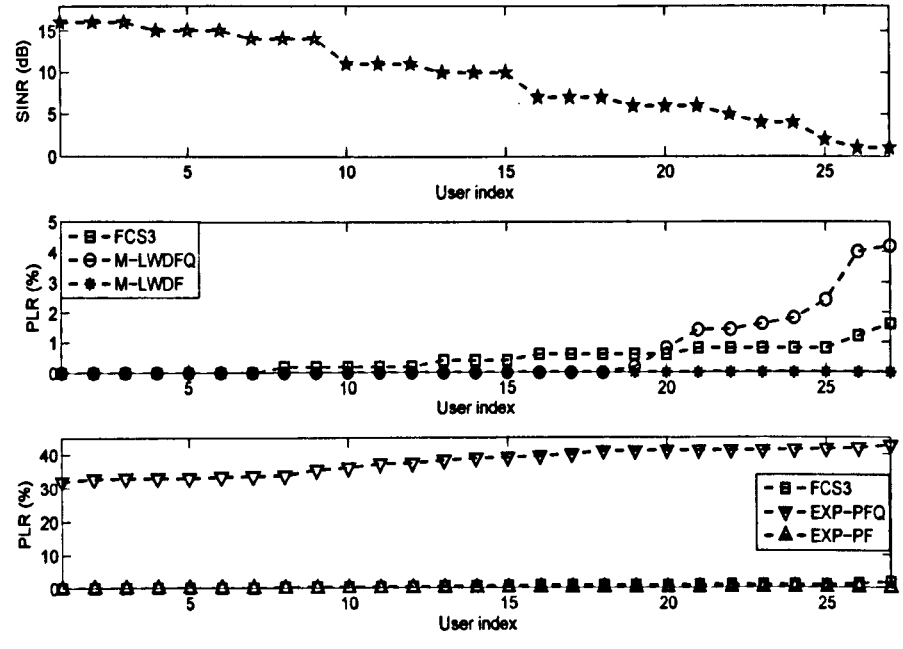


FIGURE 4.14: Performance of the VoIP flows for different scheduling rules.

rule exploits time diversity by considering the time-averaged throughput. The video flows exhibit variable bit rate (higher peak to average rate ratio) characteristics. Therefore, the higher time-averaged throughput in the scheduling decision of the M-LWDF rule increases the probability of delay violations for the video flows having lower channel quality. Hence, high rate delay sensitive flows with lower channel quality suffer from higher HoL delay violation. On the other hand, none of the VoIP flows suffer from delay violations (Figure 4.14) mainly because lower time-averaged throughput prioritizes the VoIP flows irrespective of their channel quality. M-LWDFQ reduces the QoS violations of the video flows by considering the queue size based on virtual token mechanism [70]. The PLR of the worst served user is approximately 12% and there are only 3 flows violating the PLR threshold of 5 %. The improved performance for video flows is mainly due the fact that the M-LWDFQ rule prioritizes high rate flows by considering the number of packets in the queue based on virtual token mechanism, as compared to the M-LWDF rule which relies on the HoL delay. As a result, flows having fewer packets in the queue are penalized if their channel quality is low. Figure 4.14 shows that 7 VoIP flows have PLR of more than 1 %. Therefore, the M-LWDFQ rule increases the QoS violation for the VoIP flows as compared to the M-LWDF scheduling rule.

When compared to the state-of-the-art scheduling rules, the FCS strategy improves the fairness performance for delay sensitive flows mainly due to the fact that this scheduling rule considers the channel quality of a user in a novel way, by taking into account the past and current CQI feedbacks in the time domain priority metric. This allows the users with relatively low channel quality and high HoL delay to be prioritized in the time domain. As a result, the difference in the average waiting time of each flow's packet is low. On the other hand, state-of-the-art scheduling rules favor the good channel quality flows by serving them way before their packet's delay budget. These scheduling rules are highly unfair for the cell edge users as they require a substantial increase in the SINR of the cell edge users so that their packet's delay budget requirements are met. In the FCS scheduling strategy, the PLR over the moving average window is kept below the threshold for each of the delay sensitive flows in the system. Therefore, this rule balance different flows' probabilities of QoS violations. It is important to note that the FCS strategy requires an admission controller to limit the arrival rate of delay sensitive traffic within the achievable rate region. Since fairness is incorporated in the scheduling decisions, an increase in the arrival rate above the system capacity violates the QoS performance of the flows already being served.

The performance of the EXP-PF and EXP-PFQ scheduling rules for video and VoIP traffic classes is shown in Figure 4.13 and 4.14 respectively. For video flows, the EXP-PF scheduling rule increases the QoS violations significantly, *i.e.*, approximately all the flows have delay violations of more than 5 %. The performance of the EXP-PF rule

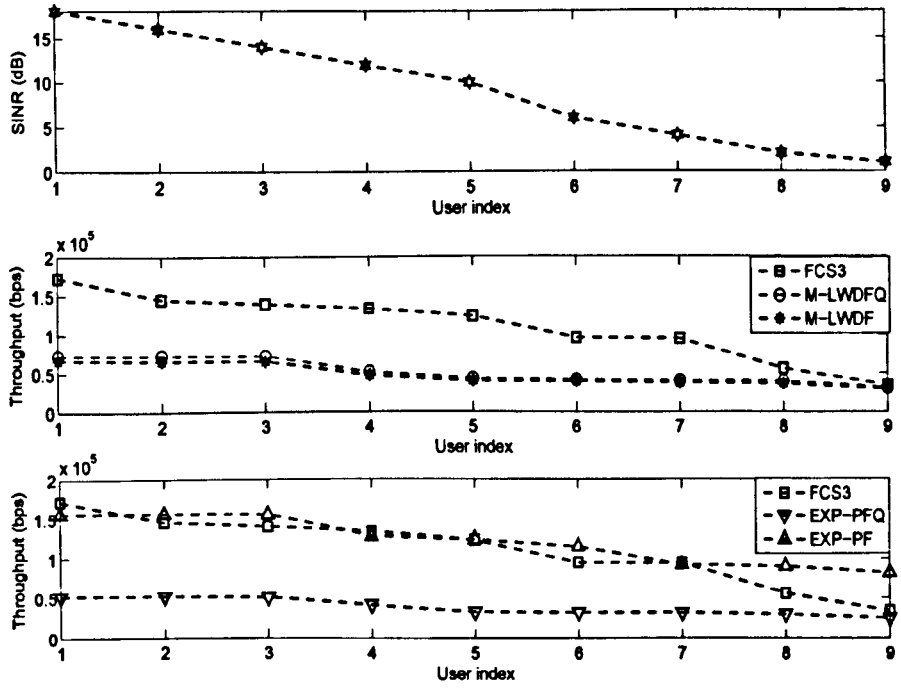


FIGURE 4.15: Performance of the best-effort flows for different scheduling rules.

for VoIP flows is the same as that of the M-LWDF rule. The M-LWDF and EXP/PF rule are delay based schedulers. These scheduling rules prioritize VoIP flows mainly because of the lower rate requirements. The token based version [70] of these scheduling rules penalizes the VoIP flows more because of the higher queue size of the video flows. EXP/PFQ performs worst for the VoIP flows mainly because the queue size of the VoIP flows always remains lower than the Video flows, which causes an exponential increase in the priority of the video flows. Therefore, all performance gain obtained by video flows penalizes to the VoIP and best-effort flows as shown in Figures 4.14 and 4.15. For best-efort flows (Figure 4.15), the performance of the EXP-PF rule is significantly better than other scheduling rules. The EXP/PFQ scheduling rule performs best for the video flows. However, it severly penalizes the VoIP flows.

All state-of-the-art scheduling rules prioritize best-effort flows by using the classical proportional fair equation (4.26). These rules prioritize delay sensitive flows by using the linear, logarithmic or exponential functions of the HoL delay as reported in Section 4.4.1. On the other hand, the FCS scheduling rule uses the same priority function for the best-effort and delay sensitive flows, as given in equation (4.18). The priority differentiation between the best-effort and delay sensitive traffic classes is controlled by adapting the maximum limit of the *output fuzzy set*. The same priority function for each traffic class allows the exploitation of multi-user channel diversities across all the flows. The priority of the best-effort traffic class is dynamic and changes according to the QoS

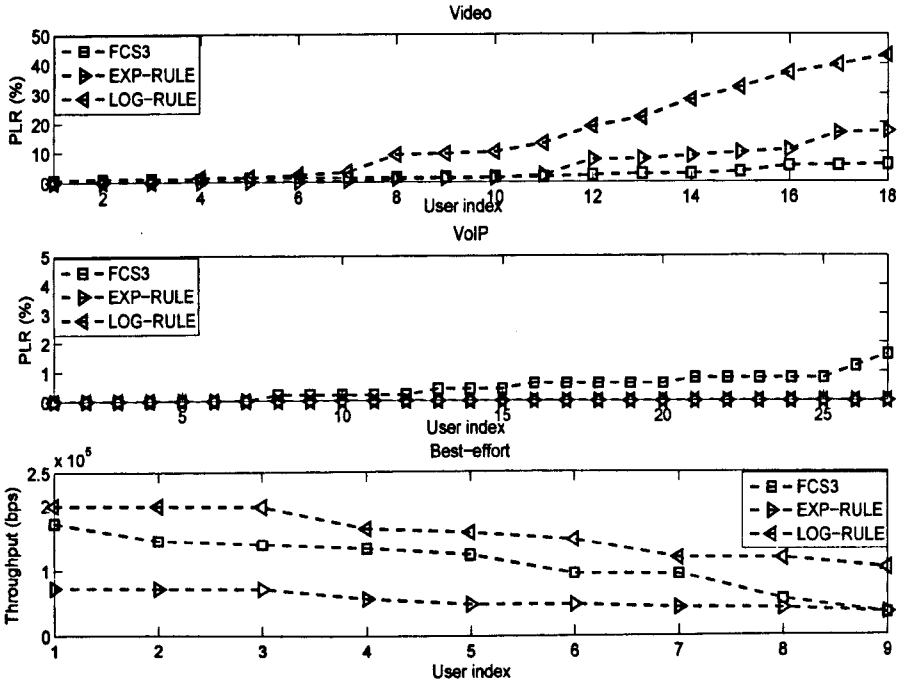


FIGURE 4.16: Performance comparison of the FCS strategy with the log and the exponential scheduling rules.

performance of the delay sensitive flows. It is important to note that video traffic class constituting the major part of the network traffic has a variable rate characteristics. The DRC fully exploits the low video traffic rate periods by reducing the priority difference between the delay sensitive and best effort flows thus achieving better fairness in terms of the total throughput achieved by each traffic class.

It is important to note that FCS strategy improves the performance of the lower channel quality video flows. Only EXP-PFQ performs better than the FCS strategy, but it achieves it by severely degrading the performance of the VoIP and best-effort flows. On the other hand, the log-rule achieves the best performance for the best-effort flows but it is highly inefficient for the video flows as shown in Figure 4.16. It is important to note that the log rule uses the normalized HoL delay as the time-domain priority. The logarithmic variation of the normalized HoL delay has a marginal increase in the priority of the lower channel quality flows. For instance, consider the log rule (equation (4.24)) with tunable parameter a_i set to 50. Let us analyze the priority difference between the flows having channel quality of 3 dB and 15 dB. The good channel quality flow with a normalized averaged delay of 0.3 results in a time domain priority of $\log[1.1 + (50.0.3)] = 2.78$. On the other hand, the poor channel quality flow having normalized HoL delay of 0.9 results in a time-domain priority of $\log[1.1 + (50.0.9)] = 3.83$. The logarithmic function marginalizes the priority of the poor channel quality flow as the

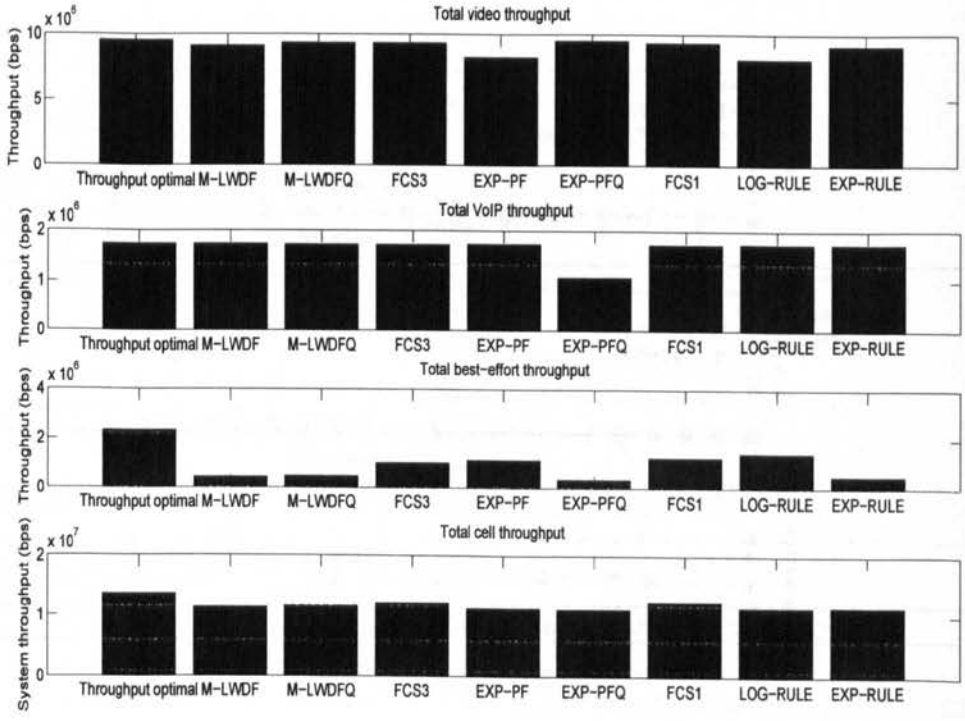


FIGURE 4.17: System throughput performance of all the considered scheduling rules.

delay urgency does not proportionately increases its priority. It is evident from Figure 4.16 that the log rule increases the PLR of the lower channel quality flows. The figure also shows the performance of the Exp-rule. According to the figure, the Exp-rule achieves better performance than the log-rule but it is highly unfair for the best-effort flows. The exponential function of the normalized HoL delay caters the delay urgency of the delay sensitive flows better than the log rule. It is important to note that the FCS strategy increases the PLR of the VoIP flows as compared to the linear, logarithmic and exponential delay based scheduling rules. However, the VoIP traffic class is packet loss tolerant and can tolerate the PLR threshold of 1%. The FCS strategy marginally violates the packet loss threshold of 2 VoIP flows, *i.e.*, the worst served VoIP flows have a PLR of 1.2 % and 1.6 %.

Figure 4.17 summarizes the performance of all the scheduling rules. The figure reports the throughput achieved by each of the traffic classes. The last sub-figure shows the total system throughput, which is simply the sum of the throughput achieved by each of the traffic classes. When compared to the optimum channel utilization strategy, the FCS scheduler compromises approximately 10.5 % of the total cell throughput while providing fairness and QoS provisions. It is clear from the figure that among all the aforementioned QoS aware scheduling rules, the FCS scheduling rule achieves the best inter-class fairness in terms of the throughput achieved by each traffic class.

4.5 Conclusion

A composite scheduling strategy is proposed for downlink scheduling at the MAC layer for delay sensitive traffic in wireless systems based on OFDMA. This strategy uses novel concept of providing fairness using fuzzy logic membership functions and its rule base, instead of relying on the rate based proportional fair strategies employed in the literature. Furthermore, the proposed framework provides service class differentiation among different traffic classes by utilizing the fuzzy logic priority scheme. The considered approach leads to a framework which provides intra-class as well as inter-class fairness. The design of the scheduling rule is robust and it serves well in diverse channel and rate requirements.

Chapter 5 presents scheduling strategies specifically for video streaming traffic. Scheduling strategy presented in chapters 3 and 4 are mainly packet scheduling rules. In Chapter 5, the scheduler exploits the scalable properties of the video streaming traffic and schedules layers based on the video frame importance.

Chapter 5

QoE-aware Fair Downlink Scheduling for Scalable Video Transmission over LTE Systems

5.1 Introduction

One of the requirements of video transmission systems over networks is the capability to adjust the system resources to the variable channel conditions. Scalable Video Coding (SVC) [75] provides considerable advantages, such as flexibility and convenience for achieving the desired visual quality and bit rate. SVC has been standardized to extend the capabilities of the H.264/advanced video coding (AVC) standard [76]. The fundamental principle of SVC is to generate a single compressed bitstream that can adapt to the varying bit rates of different transmission channels, display resolutions, and computational resource constraints of various receivers rapidly and easily. To obtain such flexibility, the SVC extension includes three types of scalability: spatial, temporal, and quality scalability, respectively resulting in an adjustment of the frame rate, of the spatial resolution, and of the image quality. An SVC stream is composed of a base layer and several enhancement layers. As long as the base layer is received, the receiver can decode the video stream. As more enhancement layers are received, the decoded video quality is improved. Temporal scalability refers to representing the same video sequence with different frame rates. The temporally scalable bit-stream is generated using hierarchical prediction structures. Spatial scalability refers to representing the video in different spatial resolutions or sizes. Normally, the picture of a spatial layer is based on the prediction from both lower temporal layers and spatial layers.

From a mobile operator point of view the goal of a scheduling strategy is to maximize the number of satisfied users for a given cost. From this point of view, achieving fairness does not mean to provide the same video quality to all video users, but in a loaded network users with good channel should enjoy a better video quality through the exploitation of channel quality. On the other hand, users with a relatively lower channel quality should have an acceptable quality by exploiting the layered structure of SVC, *i.e.*, scheduling the most important video layers. Hence, there is a trade-off between the exploitation of channel quality and the number of video layers scheduled. The main goal of this work is to address this trade-off.

There is a rich literature on content aware scheduling strategies. Most of the works make use of the multi-user content aware gradient-based or priority-based scheduling approach such as in [77, 78] [79] [80] [81] [82]. Some of the scheduling approaches exploit multi-user content diversity only while others exploit multi-user frequency and time diversity, but none of these works have exploited all the three types of diversity phenomenas, i.e, multi-user content, time and frequency diversities. Therefore one of the goals is to investigate the capacity gain that can be achieved when all three types of user diversity are exploited.

Game theory [83] and in general equilibrium theory was originally developed to study specific problems in economics and finance and was later applied in different areas, such as social science, physics, biology, engineering. In a scenario where multiple users have to share resources, the exploitation of the results of game theory leads to interesting results. In this work the concept of Nash equilibrium [84] for the design of a fair scheduling strategy is utilized in the downlink of an OFDMA based system. The application of the concepts of game theory in the area of rate control for communication networks was addressed in [84] where an analogy with the concept of proportional fairness was established. The concept was further addressed in [85], where it was shown that the Nash bargaining solution leads to proportional fairness for wireless systems. Where the main goal is to allocate resources by guaranteeing the Nash equilibrium, i.e., to maximize the product of payoff, using the logarithm of the throughput as a payoff. The payoff achieves the maximization of the product of throughput of all the users competing for resources. Hence, allocating resources in such a way achieves proportional fair scheduling. However, it is important to note that video quality is not a simple function of throughput. The nature of video traffic is highly variable in terms of bit rate and also different video frames have different priority, therefore solely relying on achieving target bit rate, optimum utilization of the network resources (in delivering video quality) is not achieved. Hence, the concept of Nash bargaining solution (NBS) is utilized by considering a function of the quality of the received video sequence as a payoff, instead of the achieved throughput (number of bits sent).

By exploiting the characteristics of video sources, and in particular of SVC, a utility metric is proposed by taking into account the dependency among frames in a video sequence. In SVC higher layer frames can only be decoded when all the frames from lower layers are correctly received. Therefore, a utility - *frame significance throughput* is proposed - which corresponds to some payoff only when one complete video frame is transmitted, since partially received video frames are usually not useful at the decoder. Next in order to achieve proportional fairness in terms of frame significance throughput, the maximization of the Nash product is considered. The utility design also exploits the highly variable nature of the SVC traffic in terms of frame sizes. The state of users transmitting lower size video frames in a shorter interval of time is exploited by giving other users, having lower channel quality or higher remaining backlog in the buffer or both, the incentive to transmit their frames and achieve an acceptable minimum payoff by efficiently utilizing the available resources.

The approach considered here leads to a scheduling framework which can be adapted to the business model of the operator. The operator can easily set the fairness level according to the system constraints. Note that the satisfaction of a well served users has a marginal improvement if its service level is increased further, for instance the user satisfaction is marginally increased if the average video quality, expressed in PSNR, is increased from 37 dB to 40 dB. On the other hand, if the service level is decreased below a threshold of acceptable level, the user satisfaction is decreased significantly. For instance the user satisfaction level is significantly increased if the average PSNR is increased from 27 dB to 30 dB.

This chapter is organized as follows. The system model is presented in Section 5.2. Section 5.3 presents the concept of quality-driven proportional fairness. Section 5.3 is divided in to several subsections starting with proposed novel metric frame significance throughput in temporal and quality scalable video streams in Sections 5.3.1 and 5.3.2, followed by the utility design in Section 5.3.3. The marginal utility design also called the Nash Product Scheduling (NPS) rule is presented in Section 5.3.4 followed by the concept of proportional fairness and design of gain in marginal utility in Sections 5.3.5 and 5.3.6. The proposed scheduling algorithm is presented in Section 5.3.7. Simulation scenario is presented in Section 5.4. Simulation results are analyzed in Section 5.5. Conclusions are drawn in Section 5.6.

Sections 5.7 and 5.8 proposed another QoE-aware scheduling strategy by exploiting the time-averaged bit throughput and the frame significance throughput known as the Opportunistic Proportional Fair (OPF) scheduling rule. The performance evaluation of both the NPS and OPF scheduling rules is discussed in Section 5.9.

TABLE 5.1: Mathematical symbols utilized in Chapter 5.

Symbol	Explanation
$R_{i,\varphi}^{(n)}$	Rate achieved by user i over PRB φ at the current scheduling instant n.
$R_{\text{ave}}^{(n)}$	Average throughput at current scheduling instant n taken over the transmission window of size t_w.
$\phi_{\varphi,i}^{(n)}$	Set of PRBs already allocated to a user at the current scheduling instant n.
$\Gamma_i(\chi_i)$; P_i	Payoff achieved when χ_i amount of resources is allocated to user i; payoff when no agreement is reached.
$\xi_{F,i}(F)$	Significance factor of frame F of user i.
$N_{\text{DEP}}(F)$	Total number of frames dependent on frame F of user i's flow.
N_{gop}	Total number of frames in a GOP.
$\mathbb{I}_{\{d_i=1\}}$	Indicator function, equal to 1 if its argument is true. Its argument is true if a frame is sent successfully before the deadline.
$\xi_{\text{GOP},i}(g)$	Frame significance throughput of current GOP g.
$\xi_{\text{GOP},i}(g-1)$	Frame significance throughput of the last GOP sent.
$H_i^{(n)}$	Delay elapsed of user i's GOP g before reaching the decoding deadline at TTI n.
H_{max}	Maximum decoding interval of GOP g or in other words delay budget of GOP g.
$N_{Q_i}^{(n)}$	Number of packets residing in the queue of flow i at scheduling instant n.
$\xi_{\text{GOP}_{\text{thr}}}$	Minimum frame significance throughput, which a user must receive.
$U^{(n)}(\xi_{\text{GOP},i})$	Utility, as a function of frame significance throughput, achieved by user i.
α	Trade off parameter between fairness and efficiency.
$U_i^{(n)}(\text{PSNR}_i)$	Utility, as a function of mean PSNR, achieved by user i.
$\text{PSNR}_i(g)$	Mean PSNR of GOP g.
$\text{PSNR}_i(g-1)$	Mean PSNR of GOP $g-1$.
PSNR_{thr}	Minimum mean PSNR, which a user must receive.
$\Lambda_{i,\varphi}^{(n)}$	Marginal utility gained by user i on PRB φ at scheduling instant n.
$G_{i,\varphi}^{(n)}$	Gain achieved by user i on PRB φ at scheduling instant n.
$\theta_{i,\varphi}^{(n)}$	Relative strength of user i's PRB φ at the scheduling instant n.
F_{tot_i}	Total number of frames streamed in to the buffer of user i.
meanPSNR_i	Mean PSNR of total number of frames streamed in to the buffer of user i.
avePSNR	Average of mean PSNR of all the active users in the network. It is used to quantify the efficiency achieved by different scheduling strategies.
stdPSNR	Standard deviation of mean PSNR of all the active users in the network. It is used to quantify the fairness achieved by different scheduling strategies.

5.2 System model

A single cell scenario is considered in which the serving eNodeB is at the the center of the cell. The serving eNodeB's Medium Access Control (MAC) scheduler controls all the available PRBs by allocating them to active flows competing for resources. The scenario comprises an orthogonal frequency division multiplexing (OFDM) single antenna Single Input Single Output (SISO) multiuser LTE system. In SISO system a PRB can only be assigned to one user at any scheduling instant, hence there is no overlapping in PRB allocation. Each user feeds back its CQI information. There are different feedback granularities in the standard, for instance the user can send just a single CQI for the whole bandwidth or a separate CQI for each PRB. A full Channel Quality Indicator (CQI) feedback is considered. If according to the scheduling decision more than one PRB, with different CQI values on each PRB, are assigned to a user, then it is necessary to calculate an average supported CQI value. The scheduler carries out the averaging, i.e., it maps the signal to interference noise ratio experienced by the allocated PRB into one (equivalent) Additive White Gaussian Noise (AWGN) SNR. One of the mostly used averaging methods is the Mutual Information Effective SNR Mapping (MIESM) [19] [25][26] [20]. After resource allocation, the scheduler uses the MIESM method to calculate a single CQI value to be used for all the allocated PRBs of a particular user.

A video server generating a pre-encoded video traffic workload is considered. Video is encoded according to the SVC standard and organized in independently decodable units called Group of pictures (GOP). The MAC layer buffers the video frames at the eNodeB. The buffer manager time stamps the whole GOP (in this work we consider a 16 frames GOP) of a user, and the scheduler should assign enough resources to schedule the whole GOP before its decoding deadline. The decoder buffers the video frames and once the deadline of the GOP is reached, video frames are displayed. When the scheduler starts the transmission of the first GOP of a user, the decoder at the receiver waits for the complete transmission of the GOP and starts the play-out as soon as the decoding deadline of the GOP is reached. When the decoder starts displaying the frames, it is important that the scheduler ensures that the next GOP reaches the decoder before its decoding deadline. Frames which are dropped by the scheduler due to deadline violation or which are not decoded at the receiver due to transmission impairments are concealed by the Frame Copy (FC) mechanism. In FC the last correctly received frame is displayed in place of the lost frame. Refer to Chapter 2 (Section 2.3.1) for a detailed description on the simulation platform.

5.3 Game Theory and Quality-driven Proportional Fairness

Derived from the principles of game theory, a novel concept of quality-driven proportional fairness and its application for scheduling of multiple users over a wireless LTE system is presented.

From game theory, the solution of the cooperative bargaining problem maximizes the Nash product N_p

$$N_p = \prod_{i=1}^{I} (\Gamma_i(\chi_i) - P_i) \tag{5.1}$$

where χ_i represents the fraction of resources allocated to user ("player") i, $\Gamma_i(\chi_i)$ is the payoff of player i when χ_i is allocated to it, P_i is the payoff of player i when no agreement is reached (zero in our case since in that case data are not transmitted). I is the total number of users in the cell.

In recent work [85], the throughput was considered as the payoff, in the assumptions that the goal of the users is to maximize their throughput. Instead when video traffic is considered the actual goal of the users is the maximization of their received video quality, which does not only depend on the received throughput.

If we consider a utility metric based on video quality denoted as $Q_i(\chi_i)$, as the payoff of players (users) when χ_i is allocated, then equation (5.1) becomes

$$N_p = \prod_{i=1}^{I} Q_i(\chi_i) \tag{5.2}$$

where

$$Q_i(\chi_i) = \Gamma_i(\chi_i) - P_i \tag{5.3}$$

and the optimization problem is

$$\max \prod_{i=1}^{I} Q_i(\chi_i). \tag{5.4}$$

Since the logarithm is a continuous monotonic function, solving the problem above is equivalent to the following

$$\ln\left(\max\prod_{i=1}^{I}Q_i(\chi_i)\right)=\max\left(\ln\prod_{i=1}^{I}Q_i(\chi_i)\right)=\max\sum_{i=1}^{I}\ln Q_i(\chi_i) \tag{5.5}$$

The next step is hence to design a suitable utility metric representing the video quality. In the following two options are considered: the case having a possibility to measure the video quality at the receiver side ("full reference") and the case where this is not possible. Note that the first case requires in a real system full knowledge of the original transmitted sequence, hence this is not actually realistic in most systems, but is considered here as a reference. For the second case above, a user payoff ("frame significance throughput") for both temporally and quality scalable video streams is proposed.

5.3.1 User payoff in temporal scalable video streams

A metric which utilizes the dependence among frames is proposed by defining a novel metric called frame significance throughput, $\xi_{\mathrm{GOP},i}(g)$, given as:

$$\xi_{\mathrm{GOP},i}(g)=\frac{\sum_{a=1}^{N_{\mathrm{gop}}}\left[\xi_{F,i}(F_a).\mathbb{I}_{\{d_i=1\}}\right]}{\sum_{a=1}^{N_{\mathrm{gop}}}\xi_{F,i}(F_a)} \tag{5.6}$$

with

$$\xi_{F,i}(F)=\frac{N_{\mathrm{DEP}}(F)+1}{N_{\mathrm{gop}}} \tag{5.7}$$

$\xi_{F,i}(F)$ is the significance factor of frame F of user i, $N_{\mathrm{DEP}}(F)$ is the total number of frames dependent on frame F of user i's flow and N_{gop} is the total number of frames in one GOP, $\mathbb{I}_{\{d_i=1\}}$ is the indicator function, equal to 1 if its argument is true. When the scheduler successfully schedules the frame, and all the frames from which it is predicted are transmitted, then $d_i=1$, indicating that the scheduled video frame is decodable at the receiver. On the other hand, $\mathbb{I}_{\{d_i=1\}}$ is equal to zero for frames not scheduled within the delay budget.

When a frame is dropped, then $N_{\mathrm{DEP}}(F)+1$ number of frames in the GOP are not decodable. Therefore the significance factor of a frame quantifies the importance of each frame within the GOP. The frame significance throughput quantifies the proportion of the most important frames sent in a GOP.

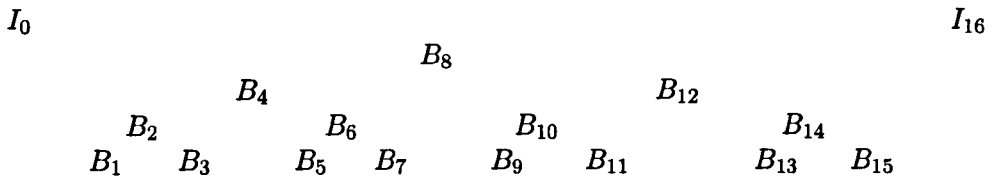


FIGURE 5.1: G16B15 (16 group of pictures with 15 B frames), 5 layers temporally scalable video stream.

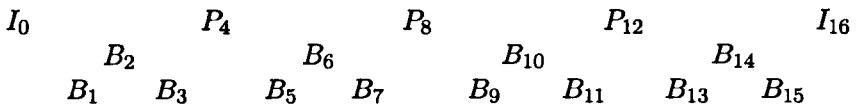


FIGURE 5.2: G16B12 (16 group of pictures with 12 B frames), 3 layers temporally scalable video stream.

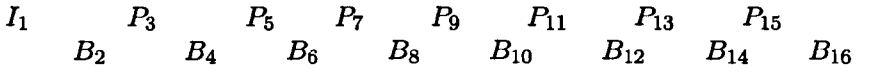


FIGURE 5.3: G16B8 (16 group of pictures with 8 B frames), 2 layers temporally scalable video stream.

Examples of the evaluation of the proposed metrics are reported in the following subsections, for different types of temporally scalable video streams.

5.3.1.1 Evaluation of the proposed metrics for different GOP structures

Consider a G16B15 (16 group of pictures with 15 B frames) temporally scalable video stream with a GOP size of 16 frames as shown in Figure 5.1. It is clear from the figure that there are in total five temporal layers for this stream. The figure also highlights the dependence among frames. For instance, layer one only contains the I_0 frame. When this frame is lost, none of the frames in the whole GOP is decodable. Frame B_8 is dependent on the key frame I_0 and on I_{16} (I_{16} is the key frame of the next GOP). Frame B_4 is dependent on frame I_0 and B_8. Similarly frame B_2 is dependent on frame B_4 and I_0. Frame B_3 is dependent on frame B_2 and B_4. B frames of layer 5 do not have frames depending on them.

Table 5.2 exemplify the calculation of the $\xi_{F,i}(F)$ metric. According to the table, first column reports the frame type F whereas the second and the third column presents the number of frames dependent on frame F ($N_{\text{DEP}}(F)$) and its significance factor ($\xi_{F,i}(F)$) respectively. Table 5.3 shows three scenarios for G16B15 GOP. According to the table when all the temporal layers are successfully scheduled, *i.e*, $\mathbb{I}_{\{d_i=1\}}$ is equal to 1 for all the frames within the GOP, then $\xi_{\text{GOP},i}(g) = 1$. In the second scenario all the frames of the last temporal layers are dropped by the scheduler, *i.e*, $\mathbb{I}_{\{d_i=1\}}$ is equal to zero

for frames B_1, B_3, B_5, B_7, B_9, B_{11}, B_{13} and B_{15}, the frame significance throughput is 0.877. Similarly if only the first three temporal layers are received and two temporal layers are dropped then the frame significance throughput drops to 0.6923.

Table 5.2 also reports the calculation of the metrics for a three layer temporally scalable video stream shown in Figure 5.2. The process to calculate the significance factor for each frame is same as above; layer one contains all the P frames, therefore the significance factor is different for each frame as shown in Table 5.2. According to Table 5.3, the frame significance throughput drops from 1 (when no layer is dropped) to 0.71 (when 2 layers are dropped) for a G16B12 GOP. Finally, the metrics for a two layer temporally scalable video stream, shown in Figure 5.3, is reported in Tables 5.2 and 5.3. According to Table 5.3, the base layer (all the P frames received correctly) frame significance throughput is 0.908. Therefore, the proposed methodology has approximately the same frame significance throughput for bit streams encoded with different layers and GOP structures.

TABLE 5.2: Calculation of frame significance factor, $\xi_{F,i}(F)$, for different structures of GOP.

Frame F, G16B15 GOP	$N_{\text{DEP}}(F)$, G16B15 GOP	$\xi_{F,i}(F)$, G16B15 GOP
I_0	15	1
B_8	14	0.9375
B_4	6	0.4375
B_2	2	0.1875
B_1	0	0.0625
Frame F, G16B12 GOP	$N_{\text{DEP}}(F)$, G16B12 GOP	$\xi_{F,i}(F)$, G16B12 GOP
I_0	15	1
P_4	14	0.9375
P_8	10	0.6875
P_{12}	6	0.4375
B_2	2	0.1875
B_1	0	0.0625
Frame F, G16B8 GOP	$N_{\text{DEP}}(F)$, G16B8 GOP	$\xi_{F,i}(F)$, G16B8 GOP
I_1	15	1
P_3	14	0.9375
P_5	12	0.8125
P_7	10	0.6875
P_9	8	0.5625
P_{11}	6 0	0.4375
P_{13}	4 0	0.3125
P_{15}	2 0	0.1875

TABLE 5.3: Frame significance throughput, $\xi_{GOP,i}(g)$, for different structures of GOP.

GOP	No. of layers dropped	$\xi_{GOP,i}(g)$
G16B15	0	1
G16B15	1	0.877
G16B15	2	0.6923
G16B12	0	1
G16B12	1	0.884
G16B12	2	0.71
G16B8	0	1
G16B8	1	0.908

5.3.2 User payoff in temporal and quality scalable video streams

The concept of frame significance throughput can be easily extended to accommodate quality scalability. The SVC stream, composed of quality and temporal scalability, comprises a base layer and one or quality more enhancement layers.

$$\xi_{Q_l} = \left[1 - \frac{l}{V_l}\right]^{N_{En}} \tag{5.8}$$

where V_l is the total number of layers in the stream, i.e, the base layer plus all the enhancement layers, N_{En} is the number of enhancement layers. The layer index is l, with the base layer represented by $l = 0$ and the last enhancement layer by $l = V_l - 1$. Therefore for video stream encoded with both quality and temporal scalability, the frame significance throughput $\xi'_{GOP,i}(g)$ is defined as follows

$$\xi'_{GOP,i}(g) = \sum_{l=0}^{V_l-1} \xi_{Q_l} \ \xi_{GOP,i}(g) \tag{5.9}$$

The payoff achieved is higher when the base layer is sent and it decreases exponentially with the number of enhancement layers present in the stream. Consider as an example a stream with two layers composed of a base layer and one enhancement layer and another stream with three enhancement layers besides the base layer. If both the layers, base and the enhancement, are successfully transmitted, the payoff achieved is 1.5 and in the four layer case the payoff achieved, when the base and one enhancement layer is transmitted, is 1.42. The lower payoff in the latter case indicates that there are further enhancement layers to be transmitted and that in general for a video stream with less enhancement layers, e.g., Coarse Grain Scalability (CGS), the increase in video quality with an additional layer is higher as compared to a stream with a large number of enhancement layers, e.g., Medium Grain Scalability (MGS). When the video stream

comprises of one base layer and no enhancement layers, then ξ_{Q_l} is one and equation (5.9) simplifies to (5.6).

5.3.3 Utility design

In this work bit throughput is referred as a payoff and logarithm of throughput as a utility. Similarly frame significance throughput is considered as a payoff and now frame significance throughput based utility is proposed in this section. According to the analysis shown by [85], the logarithm of throughput when used as a utility ensures proportional fairness. It is important to note that the logarithm of the bit throughput, when used as a utility, results in a strict utility gain at each Transmission Time Interval (TTI). For PRBs, when allocated to a user, the payoff is the number of bits transmitted. However, when the frame significance throughput is used as a payoff, one complete frame transmission corresponds to some payoff otherwise the payoff is null. However one frame transmission can take several TTIs due to variability in video traffic bitrate. Therefore, the utility gain after each TTI is a design issue. Another important point to consider is that video streaming has a stringent deadline. When the deadline has elapsed frames are dropped corresponding to a decrease in payoff.

Hence the utility design should have three important properties:

- The flow should receive a strict numeric utility gain, based on frame significance throughput, after each TTI.
- The deadline of the video flow must be taken into account.
- When throughput is used as a utility, proportional fairness in time is achieved by considering the throughput attained in previous TTIs such as the classical proportional fair rule in which an exponentially weighted low pass filter is utilized in order to update the average throughput. The utility design must have the time diversity exploitation in terms of frame significance throughput.

In order to achieve proportional fairness in time, the frame significance throughput achieved over the last GOP sent, is taken in to account. According to the considerations above, the payoff achieved or utility gained at scheduling instant n is defined as

$$U_i^{(n)}(\xi_{\mathrm{GOP},i}) = \left\{ \left[(1 - A_i^{(n)}).\frac{\xi_{\mathrm{GOP},i}(g-1)}{\xi_{\mathrm{GOP_{thr}}}} \right] + \left[A_i^{(n)}.\frac{\xi_{\mathrm{GOP},i}(g)}{\xi_{\mathrm{GOP_{thr}}}} \right] \right\}^{\alpha} \tag{5.10}$$

$$A_i^{(n)} = \frac{H_i^{(n)}}{H_{\max}} \tag{5.11}$$

where $H_i^{(n)}$ is the delay elapsed of GOP g before reaching the decoding deadline at TTI n. $H_{\max}$ is the maximum decoding interval of the GOP g. After $H_{\max}$, the remaining frames in the eNodeB buffer are dropped, $\xi_{\mathrm{GOP},i}(g-1)$ is the frame significance throughput of the last GOP, $(g-1)$, sent, $\xi_{\mathrm{GOP}_{\mathrm{thr}}}$ is the minimum frame significance throughput which a user must receive. If more than one user achieves the same frame significance throughput in the current GOP, this corresponds to the same payoff achieved by the users. The utility gain of users, attaining the same payoff, in a shorter interval of time will be higher. If users gain the same payoff in the same interval of time then the utility gain of user attaining a higher payoff in the last GOP sent will be higher.

The parameter α enables the operator of the system the flexibility to easily set the fairness level: the higher the value of α the higher will be the fairness. In addition, to be deadline aware and to exploit time diversity, the utility metric proposed here has the following features:

- The major proportion of the GOP sent in a shorter time yields larger utility and vice versa. For a user having a low channel quality, its queue will build up which results in some numeric payoff below the threshold (before the deadline of the GOP), thus more resources will be required by such a user. Such a user state will be quantified by a lower utility gain.
- Service differentiation can also be achieved by increasing the threshold of the frame significance throughput.
- Most of the times the wireless link is the bottleneck, but sometimes congestion can occur in the backbone network which results in a delayed delivery of frames to the eNodeB's buffer. Such an event will be quantified by the utility design in the form of lower proportion of a GOP sent while most of the delay budget of the GOP has elapsed.

One of the advantages of the metric described above is that it does not require full knowledge of the transmitted video sequences for its evaluation. As a benchmark, we define also a different metric, requiring full availability of the transmitted video sequence for its evaluation ("full reference").

5.3.3.1 Full reference based utility design

Assuming we have availability of a full reference video quality metric, such as PSNR, then (5.10) is transformed into:

$$U_i^{(n)}(\mathrm{PSNR}_i) = \left\{ \left[(1 - A_i^{(n)}) . \frac{\mathrm{PSNR}_i(g-1)}{\mathrm{PSNR}_{\mathrm{thr}}} \right] + \left[A_i^{(n)} . \frac{\mathrm{PSNR}_i(g)}{\mathrm{PSNR}_{\mathrm{thr}}} \right] \right\}^{\alpha} \tag{5.12}$$

The higher the number of frames received in the GOP, the higher will be the average PSNR in the GOP. For example, when the first frame, key frame I_1 in Figure 5.3, of the GOP is successfully transmitted, the information in terms of PSNR available is shown in Table 5.4. Table 5.4 shows that the PSNR decreases if only the key frame in the GOP is successfully transmitted and the last successfully received frame, I_1, is copied in place of the missing frames B_2, P_3 till B_{16}. $\mathrm{PSNR}_i(g)$ is the average PSNR of the frames in the current GOP, whereas $\mathrm{PSNR}_i(g-1)$ is the average PSNR of the last GOP.

TABLE 5.4: Sample from the offset distortion video trace file of the GOP structure of Figure 5.3. I_1 frame is copied in place of the lost frames of the GOP in Figure 5.3.

Frames	Quality of Frames in PSNR [dB]
I_1	42.97138
I_1	43.46792
I_1	36.74172
I_1	34.4448
I_1	33.0239
I_1	30.93178
I_1	27.85437
I_1	25.1831
I_1	24.13
I_1	23.08108
I_1	22.96167
I_1	20.93817
I_1	18.0939
I_1	17.57556
I_1	15.40225
I_1	14.02072

Equations (5.10) and (5.12) can be used alternatively in (5.5).

5.3.4 Marginal utility (NPS rule)

Instead of maximizing the sum of the user utilities, the main goal is to maximize the marginal utility, that is the gain achieved in utility when a PRB is allocated to a user. This is equivalent to approximating the steepest ascent maximization method. Mathematically, the marginal utility is expressed as

$$\wedge_{i,\varphi} = \ln[(U_i^{(n)}(\xi_{\mathrm{GOP},i})) + G_{i,\varphi}] - \ln(U_i^{(n)}(\xi_{\mathrm{GOP},i})) \tag{5.13}$$

where $G_{i,\varphi}$ is the gain of user i upon receiving PRB φ.

Maximizing the marginal utility instead of the full utility serves three goals.

- The marginal utility is calculated on each PRB, thus exploiting statistically independent multi-user frequency selective fading.
- Finding a user that maximizes the gain in utility leads to greater sum utility than a user which maximizes full utility.
- The utility design achieves proportional fairness in time by exploiting multi-user, time and content diversity. This leads us to design a gain which is a function of the rate achieved on each PRB and has the characteristic of proportional fairness in frequency. Thus the marginal utility design achieves proportional fairness in time and frequency.

We address in the following two implementation aspects:

5.3.4.1 Null Utility

The use of equation (5.13) in the proposed utility, according to equation (5.10), poses some limitations in the event where the achieved utility is zero. Although the occurrence of such an event has a low probability, it can occur when the frame significance throughput of the last GOP sent is zero. The frame significance throughput can be zero if the I frame of the GOP is dropped. When the transmission of the next GOP starts, the achieved utility will be zero, and its logarithm ∞ according to the equation (5.13). For the aforementioned reason, we transform equation (5.13) into:

$$\wedge_{i,\varphi} = \ln[1 + (U_i^{(n)}(\xi_{\mathrm{GOP},i})) + G_{i,\varphi}] - \ln[1 + (U_i^{(n)}(\xi_{\mathrm{GOP},i}))] \tag{5.14}$$

Equation (5.14) avoids dealing with ∞ value, which is not the case in equation (5.13).

5.3.4.2 Initialization

Another issue about the implementation detail is when a new user enters the system, there is no information about the payoff achieved in the last GOP $g-1$. In such case we initialize $\frac{\xi_{\mathrm{GOP},i}(g-1)}{\xi_{\mathrm{GOP}_{\mathrm{thr}}}} = 1$, i.e, the frame significance throughput achieved in the last GOP is equal to the threshold frame significance throughput. According to this assumption, the payoff achieved in the current GOP g will depend upon the channel condition and load. If the channel quality of the user is good or the system is under light load or both, then the payoff achieved by the user will be higher than the threshold payoff.

5.3.5 Proportional Fairness

Before defining the gain term, $G_{i,\varphi}$, the concept of proportional fairness in time and frequency is revised in this section.

In CDMA systems, all the available resources are allocated to a single user for a given TTI. Proportional fairness exploits only time diversity. Hence in a single carrier system user i^* is selected according to equation [62]:

$$i^* = \text{argmax}_i \frac{R_i^{(n)}}{R_{\text{ave},i}^{(n)}} \tag{5.15}$$

$$R_{\text{ave},i}^{(n)} = R_{\text{ave},i}^{(n-1)} \left(1 - \frac{1}{t_w}\right) + \frac{1}{t_w} R_i^{(n-1)} \tag{5.16}$$

where $R_i^{(n)}$ and $R_{\text{ave},i}^{(n)}$ are the instantaneous and average throughput at current scheduling instant n for user i. The average throughput is calculated in equation (5.16) over the transmission window of size t_w, e.g., 1000 TTIs. $R_i^{(n-1)}$ and $R_{\text{ave},i}^{(n-1)}$ are the instantaneous and average rate achieved in the previous scheduling instant $n-1$ by user i.

Note that equation (5.15) results in an inefficient system for OFDMA, since all the PRBs would be allocated to a single user, neglecting statistically independent multiuser frequency diversity. For multicarrier systems, equation (5.15) becomes [64]:

$$i^* = \text{argmax}_i \frac{R_{i,\varphi}^{(n)}}{R_{\text{ave},i}^{(n)}} \tag{5.17}$$

where $R_{i,\varphi}^{(n)}$ is the rate achieved by user i over PRB φ at the current scheduling instant n. This approach attains proportional fairness in time and frequency. Another approach which attains proportional fairness in time and frequency is to select the user according to [64] :

$$i^* = \text{argmax}_i \left(\frac{R_{i,\varphi}^{(n)}}{R_{\text{ave},i}^{(n)} + \sum_{j \in \phi_{\varphi,i}^{(n)}, j \neq i} R_{i,j}^{(n)} T} \right) \tag{5.18}$$

where $\phi_{\varphi,i}^{(n)}$ is the set of PRBs already allocated to a user at the current scheduling instant and T is the time duration of the TTI. This approach takes in to account the PRBs allocated to a user in the current TTI. Both the previous approaches, equation (5.17) and (5.18), attain proportional fairness in time and frequency, by taking into account

the rate achieved in the previous TTIs. According to [63] the following equation attains proportional fairness in frequency.

$$i^* = \text{argmax}_i \left(\frac{R_{i,\varphi}^{(n)}}{\sum_{j \in \phi_{\varphi,i}^{(n)}, j \neq i} R_{i,j}^{(n)} T} \right) \tag{5.19}$$

It is important to note that proportional fairness in frequency can be unfair for the cell edge user. For instance when the number of PRBs is lower than the number of users, users with good channel conditions will be allocated one PRB each, whereas users experiencing lower channel quality will not get any resource, i.e., users with good channel will starve users with lower channel quality. When the number of PRBs is equal to the number of users, each user will get one PRB each. Thus equation (5.19) achieves proportional fairness when number of PRBs is equal or greater than the number of users.

It has been proved in [84] that the logarithm of the rate, used as a utility, achieves proportional fairness, i.e.,

$$\max \sum_{i=1}^{I} \ln R_i \tag{5.20}$$

where we can consider alternatively:

$$R_i = \frac{R_{i,\varphi}^{(n)}}{\sum_{j \in \phi_{\varphi,i}^{(n)}, j \neq i} R_{i,j}^{(n)} T} \tag{5.21}$$

or

$$R_i = \frac{R_{i,\varphi}^{(n)}}{R_{\text{ave},i}^{(n)} + \sum_{j \in \phi_{\varphi,i}^{(n)}, j \neq i} R_{i,j}^{(n)} T} \tag{5.22}$$

5.3.6 Gain

The goal of this section is to design a gain $G_{i,\varphi}$ term exploiting multi-user frequency diversity. We want to achieve proportional fairness in frequency. Equation (5.19) provides proportional fairness in frequency but it suffers from fairness when number of PRBs are less than the number of active users. We aim to design a gain term which provides proportional fairness in frequency and can be utilized irrespective whether the number

of users are below or above the number of available PRBs. Therefore, the gain term is formulated as:

$$G_{i,\varphi}^{(n)} = R_{i,\varphi}^{(n)}.\theta_{i,\varphi}^{(n)} \tag{5.23}$$

where $\theta_{i,\varphi}$ is the relative strength of PRB φ of user i. Mathematically it is defined as:

$$\theta_{i,\varphi}^{(n)} = \frac{R_{i,\varphi}^{(n)}}{\frac{\sum_{m=1}^{M_{PRB}} R_{i,m}^{(n)}}{M_{PRB}}} \tag{5.24}$$

The rationale behind the introduction of $\theta_{i,\varphi}$ is to ensure that each user is scheduled on its best PRB, thus fully utilizing multiuser frequency diversity. In the beginning of each scheduling instant, $\theta_{i,\varphi}$ is calculated for each user i on each PRB φ. M_{PRB} is the total number of PRBs in the system. It is important to note that equation (5.23) is different from all the equations in Section 5.3.5 as it does not consider the PRBs allocated to the user at the current scheduling instant. Alternatively it considers the relative strength of each PRB of all the users before the allocation process. Below is a numerical example explaining the significance of this factor.

Considering $RB_{\text{efficiency}}(i,\varphi)$ as the rate achieved by each user on each PRB φ; an example is reported in the the following matrix

$$RB_{\text{efficiency}} = \begin{pmatrix} 3.3223 & 5.5547 & .2344 & 0.6016 & 1.4766 & 1.4766 \\ 2.7305 & 5.5547 & 1.1758 & 0.8770 & 1.4766 & 1.4766 \\ 3.9023 & 5.5547 & 2.7305 & 2.4063 & 1.4766 & 0.8770 \\ 3.3223 & 5.5547 & 3.3223 & 2.7305 & 0.6016 & .1523 \\ 5.1152 & 5.5547 & 3.9023 & 2.4063 & 0.2344 & 1.4766 \\ 5.1152 & 5.5547 & 3.9023 & 1.4766 & 0.1523 & 1.1748 \end{pmatrix}$$

Each row in the above matrix reports the spectral efficiency in bits/second/hertz (taken from Table 2.2 in Chapter 2) for one resource block and for the different users, whereas each column reports the spectral efficiency values for user i over the different resource blocks. In other words, rows of the matrix represent the user index i and columns represents the PRB index φ. Therefore the above matrix is a representation of the channel quality (in bits/second/hertz) for the different users over the the different PRBs. There are 6 rows in total, i.e., the bandwidth is 1.4 MHz (6 PRBs available for scheduling) for the case of this example. Equation (5.24) is applied to calculate the relative strength of each PRB which results in a matrix as shown below:

$$\theta_{i,\varphi} = \begin{pmatrix} 0.8480 & 1.0000 & .0921 & 0.3438 & 1.6352 & 1.3353 \\ 0.6969 & 1.0000 & 0.4612 & 0.5012 & 1.6352 & 1.3353 \\ 0.9960 & 1.0000 & 1.0731 & 1.3753 & 1.6352 & 0.7931 \\ 0.8480 & 1.0000 & 1.3056 & 1.5605 & 0.6662 & .1377 \\ 1.3056 & 1.0000 & 1.5336 & 1.3753 & 0.2596 & 1.3353 \\ 1.3056 & 1.0000 & 1.5336 & 0.8439 & 0.1687 & 1.0633 \end{pmatrix}$$

The numerical example shows that the usage of the relative strength of each PRB not only improves the system efficiency but also introduces fairness; for instance if 2 or 3 PRBs of a user are under a deep fade and the remaining PRBs are experiencing lower fading, this factor will assign higher weights to PRBs experiencing lower fading (although their attainable rates are lower). In the numerical example above, as users 4, 5 and 6 are having lower average channel condition compared to users 1, 2 and 3, the relative strength factor will assign higher weights to good PRBs of user 4, 5, and 6. For example, PRB 4 of user 4 is having a lower attainable rate compared to the PRB 4 of user 1, 2 and 3, but the relative strength of user 4 on PRB 4 is the highest. When we take the product of rate and relative strength of each PRB according to the equation (5.23), each user is scheduled on its best PRB depending up on the utility gained. A user with lower utility achieved in the previous TTIs will require more resources; this factor will ensure that such user will be assigned its best PRBs with a higher probability. The utility design attains proportional fairness in time by considering frame significance throughput, in case of no-reference based utility, and PSNR, in case of full-reference based utility. The gain design ensures proportional fairness in frequency but not considering the PRBs already allocated to a user in the current TTI as in equation (5.19), instead a product of the instantaneous rate and relative strength of each PRB is considered. The proposed algorithm, described in the following, performs a linear search on each PRB. Another advantage of using the relative strength of each PRB gives information about the amount of fading on other PRBs which are not yet allocated.

5.3.7 NPS scheduling algorithm

In multiuser downlink scheduling scenario, eNodeB is the central authority responsible for enforcing the agreements among all the users. In the following, two assumptions are considered.

- The virtual negotiations have already taken place among all the active users in order to enforce the optimal NBS.

- The most important frames, i.e., the frames with the highest significance factor within the GOP, are transmitted first by the video streaming server, thus each user receives the most important frames first.

The scheduling algorithm consists of three steps:

- Find the user that has the highest gain in utility, $\Lambda_{i,\varphi}$ according to equation (5.14), among all users when PRB φ is allocated to it, i.e., for each PRB φ, find the user i^* such that: $i^* = \text{argmax}_i \Lambda_{i,\varphi}$.
- Allocate PRB φ to user i^*. If more than one user have the highest marginal utility then allocate PRB to the user with the highest gain, $G_{i,\varphi}$.
- Delete the PRB φ from the available set of PRBs and repeat steps 1-3 until all PRBs are allocated.

5.3.7.1 Features of the scheduling algorithm

The features of the algorithm are summarized below:

- Low complexity.
  The proposed algorithm allocates a PRB to the user after performing a linear search among all the active users thus the computation complexity is the product of the number of users and the number of PRBs in the system. Hence linear complexity enables real-time implementation of the algorithm.
- Variable fairness.
  The parameter α can be varied to change the fairness level. This gives the operator of the system the flexibility to change the fairness level
- Metric independent.
  The proposed utility design is independent of the metric. Both the full reference and no reference based metric can be employed in the utility design. This allows both the content providers and operators to adapt to different reference and non-reference based metrics.
- Good trade off between efficiency and fairness.
  The marginal utility principle achieves a good trade of between fairness and efficiency by efficiently exploiting statistically independent multi-user time, frequency and content diversities.

5.4 Simulation Environment

The list of assumptions employed in the simulations are as follow: The channel quality of each user remains constant during the subframe period of 1 ms, although it changes from subframe to subframe. CQI feedback from UE to the eNodeB is error free. The error free assumption of the feedback channel is satisfied by using efficient and heavily coded feedback stream. It is assumed that equal downlink transmit power is allocated on each PRB. It is assumed that, at any time instant, pathloss is fixed on each PRB. Multipath induced fading is modeled by a tap-delay based model known as Typical Urban. The model is developed according to the guidelines in [49, 50]. The simulation parameters are reported in Table 5.5.

TABLE 5.5: Simulation parameters for video application on LTE schedulers.

PARAMETERS	VALUE
Application	Video Traffic Trace of Die Hard
Video Streaming rate	166, 224.55 and 556 kbps
Video Trace Sequence	Die Hard
Temporal Layers	5
Quantization Parameter	28
GoP pattern	G16B15
Encoder	JSVM(9.15)
Frame Rate	30 Frames Per Second
Concealment algorithm	Frame Copy
Bandwidth	10 MHz
Mode	Tx = 1 and Rx=1 or SISO
Pathloss model	Hata-Cost-231 model (urban pathloss model)
HARQ	Synchronous retransmissions (up to 3 retransmissions)
Channel	3GPP-TU (Typical Urban)
Channel Fading	Block Fading
UE speed	15 to 100 km/h (UEs moving at variable speed)
UE distribution	uniform
Cell radius	1 Km
Maximum time duration of GOP in the buffer	$H_{\max} = 500$ ms

Three different bitrate video streams with different PSNR are considered as shown in Table 5.6.

TABLE 5.6: Parameters of video streaming model

Temporal Level	Rate with each additional layer [Kbps]	Quality with each additional Layer [dB]
Layer 5	166	40.0972
Layer 4	126	36.7332
Layer 3	92.14	32.9785
Layer 2	66.55	28.9812
Layer 1	50.37	26.2946
Temporal Level	Rate with each additional layer [Kbps]	Quality with each additional Layer [dB]
Layer 5	224.55	40.5614
Layer 4	162.46	34.7578
Layer 3	108.15	29.033
Layer 2	69.5	24.9001
Layer 1	43.3	22.33
Temporal Level	Rate of each layer [Kbps]	Quality with each additional Layer [dB]
Layer 5	556	37.64
Layer 4	404.37	30.83
Layer 3	258.36	24.96
Layer 2	154.33	21.25
Layer 1	88.86	19.27

5.5 Results

To investigate the fairness and efficiency analysis, a five layered temporally scalable video stream is considered with high, medium and low rate of 556, 224.55 and 166 Kbps respectively as shown in Table 5.6. The information available in the table is relevant to different portions of video sequence, each 5 seconds long, of the *Die Hard* video trace file. Traffic statistics, quality statistics, and traces for this video sequence are publicly available in [86]. The increase in quality when extra layers are added, in terms of mean PSNR, is highly variable as shown in Table 5.6. The mean PSNR of each user is calculated as:

$$\text{meanPSNR}_i = \frac{1}{F_{\text{tot}_i}} \sum_{f=1}^{F_{\text{tot}_i}} PSNR_{i,f} \tag{5.25}$$

where $\text{PSNR}_{i,f}$ is the PSNR value at the f-th frame of user i, meanPSNR_i and F_{tot_i} are the mean PSNR and the total number of frames requested by user i.

According to Table 5.6, when 4 out of 5 temporal layers are correctly received, the mean PSNR for high rate traffic is 30.83 as compared to 34.75 and 36.73 for medium and low rate frame sequences respectively. Therefore for high rate traffic more than 3 temporal layers are necessary for an acceptable perceived video quality according to the mean PSNR to MOS mapping shown in Table 5.7. For medium and low rate traffic, three temporal layers are enough for fair perceived video quality. The scenario comprises a uniformly distributed users within the cell, the distribution of different rate users are given in Table 5.8. It is ensured in the simulations that high, low and medium rate users are uniformly distributed at the cell edge as well as at the centre of the cell. In the following, the video quality received for users with the following scheduling strategies are analyzed: bit throughput, frame significance throughput and PSNR based proportional fair schedulers. For bit throughput based proportional fair strategy, equation (5.17) is utilized. This equation provides proportional fairness in time and frequency by exploiting multi user time and frequency diversities.

TABLE 5.7: Mean PSNR to MOS mapping [7]

PSNR [dB]	MOS value	Quality
< 20	1	Bad
20-25	2	Poor
25-31	3	Fair
31-37	4	Good
> 37	5	Excellant

TABLE 5.8: Number of different rate users in the the cell for 84 user case

High rate	Medium rate	Low rate
22	28	34

Figure 5.4 reports the cumulative distribution function (cdf) of the mean PSNR for the different users in the system. A summary of the relevant results in terms of mean PSNR to MOS mapping (Table 5.7) is given in Table 5.10. According to the cdf of bit throughput based proportional fair scheduler, for approximately 15 percent of the users the mean PSNR is between 26 dB and 31 dB, that is, 13 users out of 84 are receiving a fair perceived video quality, whereas 9 users are enjoying good video quality and for 62 users the perceived quality is excellent. Note that the cdf curve undergoes a steep rise from 0.26 to 1, which shows that for 22 high rate users the mean PSNR is low compared to the medium and low rate users. Therefore the video quality difference among the users is very high. This is more evident from the fact that the bit throughput based utility favors low to medium rate users rather than high rate users. Another important point to consider is that the lowest PSNR value with the bit throughput based proportional fair rule is approximately 26 dB, which corresponds to the fact that this scheduling strategy

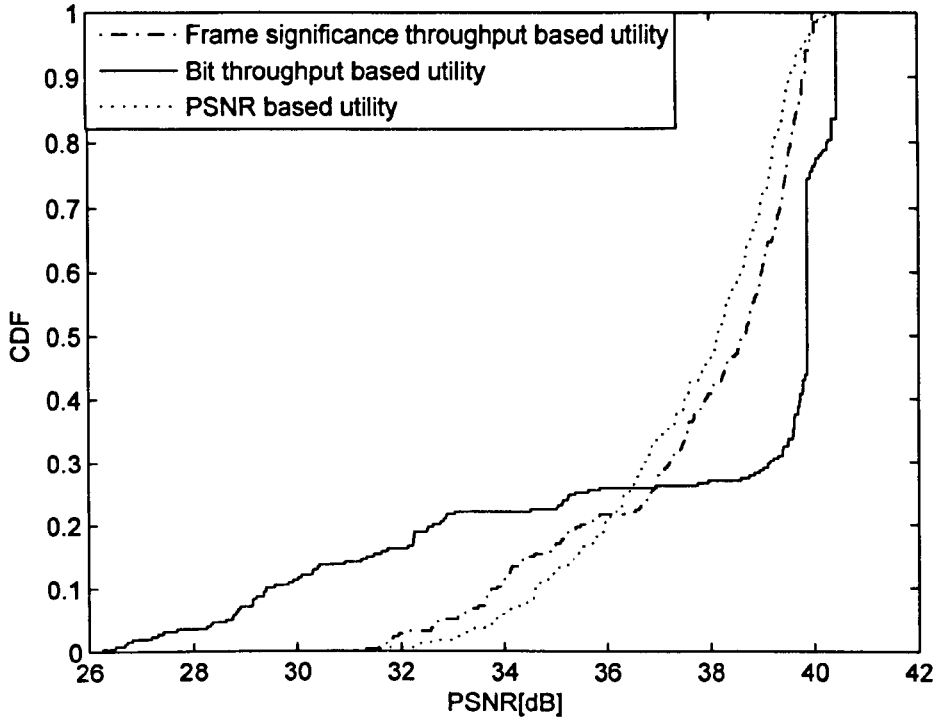


FIGURE 5.4: Mean PSNR CDFs for 84 users, $\alpha = 2.5$ for PSNR based utility and $\alpha = 3$ for frame significance throughput based utility.

cannot serve more than 84 users: if more users are added to the network then this will have an adverse affect on the perceived video quality of high rate and low channel quality users. On the other hand the frame significance throughput based scheduling strategy considerably improves the performance of cell edge and high rate users. In this case none of users experience a mean PSNR below 31 dB, as shown by the frame significance throughput based scheduling strategy cdf in Figure 5.4, therefore all the users have a perceived quality which is good to excellent according to mean PSNR to MOS mapping. The frame significance throughput based scheduling strategy using the marginal utility principle improves the fairness without compromising much on the efficiency. This is confirmed by Table 5.9 which reports the average of the mean PSNR of all the users according to equation (5.26) and the standard deviation according to equation (5.27).

$$\text{avePSNR} = \frac{\sum_{i=1}^{I} \text{meanPSNR}_i}{I} \tag{5.26}$$

The higher avePSNR, the higher the system efficiency in terms of overall video quality. To measure the fairness in terms of PSNR, the standard deviation of PSNR is considered.

$$\text{stdPSNR} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(\text{meanPSNR}_i - \text{avePSNR})^2} \tag{5.27}$$

TABLE 5.9: Average and standard deviation of PSNR, 84 user case, for different scheduling strategies

Scheduler	avePSNR [dB]	stdPSNR [dB]
Bit throughput based Proportional fair	37.75	3.92
Frame significance throughput based proportional fair	37.65	2.2
PSNR based Proportional fair	37.1	1.83

TABLE 5.10: Perceived quality for 84 user according to mean PSNR to MOS mapping

Scheduler	Fair	Good	Excellent
Bit throughput based Proportional fair	13	9	62
Frame significance throughput based proportional fair	0	25	59
PSNR based Proportional fair	0	28	56

The lower the standard deviation, the fairer the quality among the users. The scheduler based on frame significance throughput achieves the same efficiency as the throughput based scheduler, but with an improvement in standard deviation of 1.72 dB, as shown in the Table 5.9. The performance of the PSNR based scheduler is approximately the same as the scheduler based on the frame significance throughput, but with an improved fairness as shown by the cdf curve of the PSNR based scheduler in Figure 5.4. A summary of the results in Figure 5.4 in terms of perceived quality according to mean PSNR to MOS mapping is given in Table 5.10. Figure 5.4 shows that the metric based on the frame significance throughput is a good trade-off metric in multiuser video scheduling. This metric prioritizes the frames according to their dependence on other frames, which is not the case in simple bit throughput based multiuser video scheduling. On other hand, its complexity is very low when compared to full reference based metrics such as PSNR, which require complete information about the original video sequences.

In order to investigate the capacity gain, the distribution of the 84 users is kept fixed and 16 medium rate users are added to the network, as shown in Table 5.11. The cdf of

TABLE 5.11: Number of different rate users in the the cell for 100 user case

High rate	Medium rate	Low rate
22	44	34

TABLE 5.12: Perceived quality for 100 user case according to mean PSNR to MOS mapping

Scheduler	α	Fair	Good	Excellent
Frame significance throughput based proportional fair	3	11	45	44
Frame significance throughput based proportional fair	6	10	69	21
PSNR based Proportional fair	2.5	8	56	36
PSNR based Proportional fair	5	1	90	9

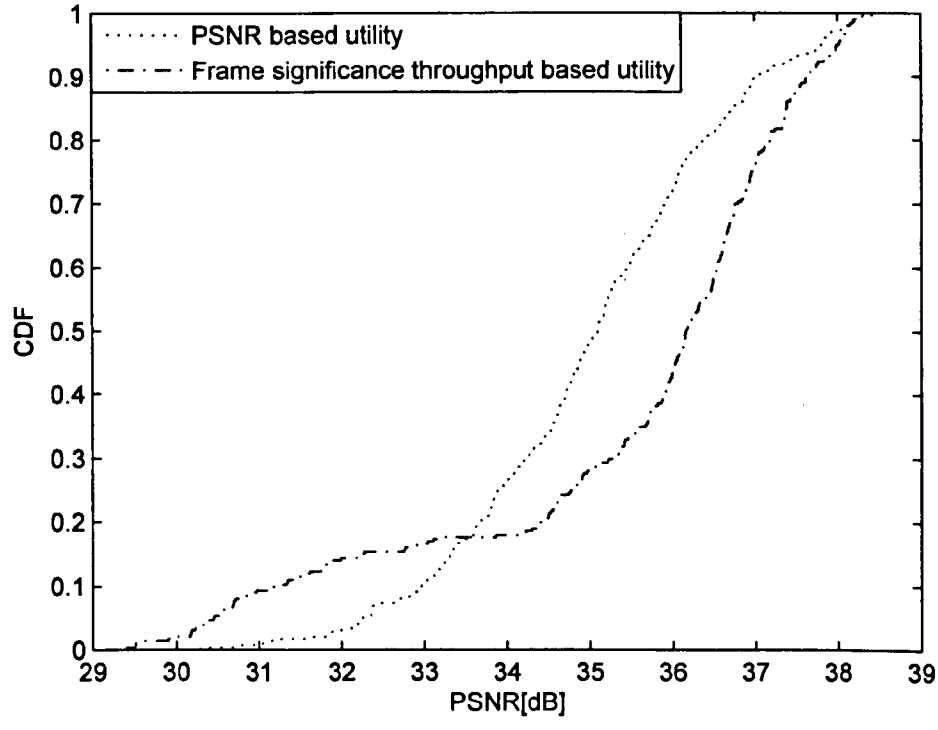


FIGURE 5.5: Mean PSNR CDFs for 100 users. $\alpha = 5$ for PSNR based utility and $\alpha = 6$ for frame significance throughput based utility.

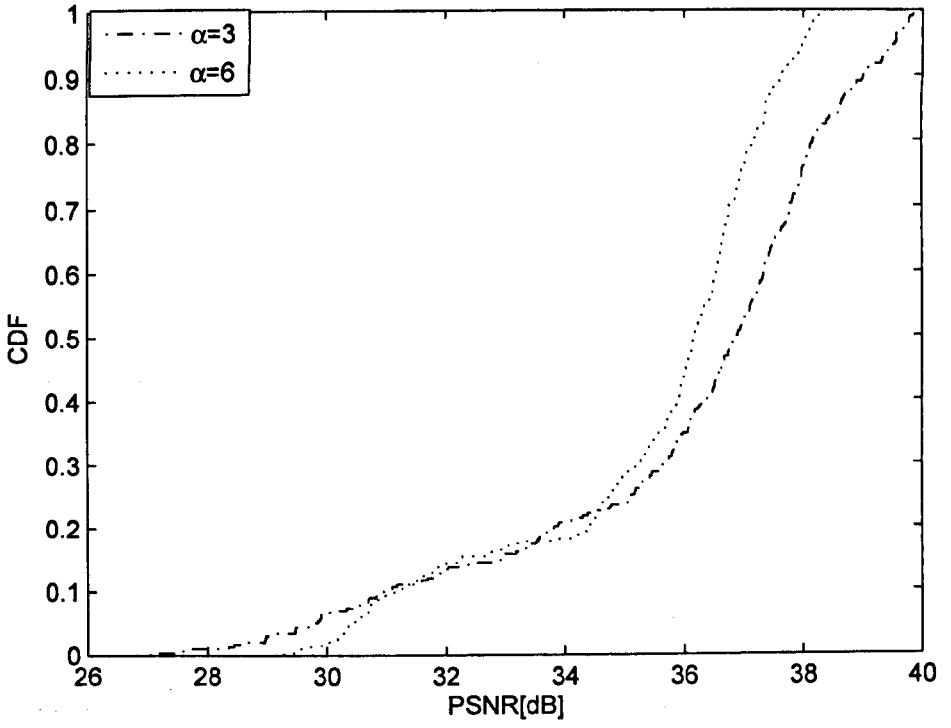


FIGURE 5.6: Impact of α on the frame significance throughput based scheduler. 100 users in the cell.

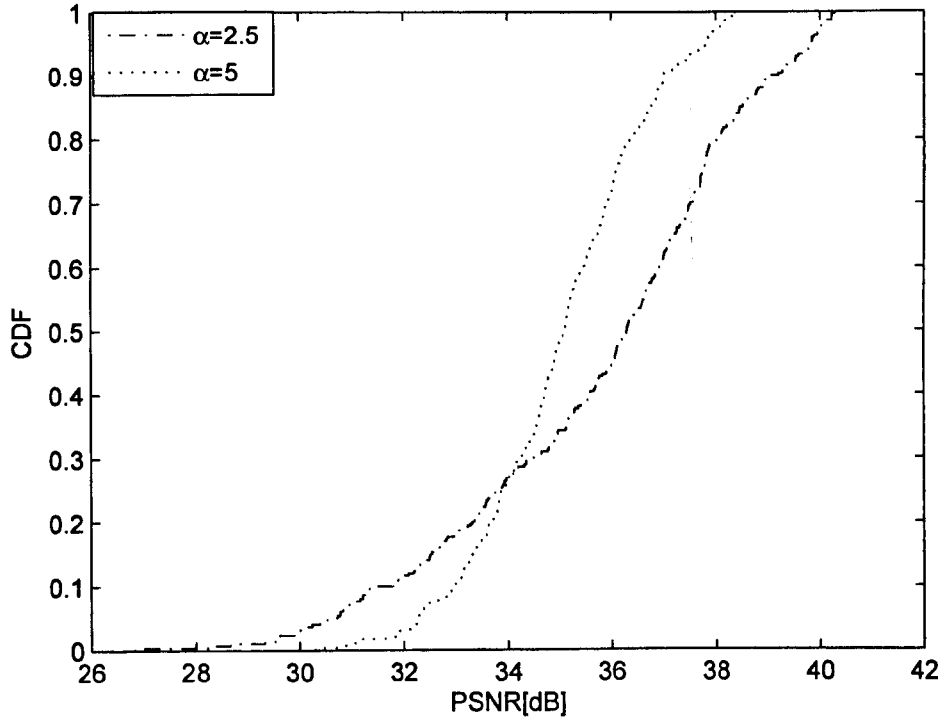


FIGURE 5.7: Impact of α on the PSNR based scheduler. 100 users in the cell.

the Mean PSNR for the case with 100 users is shown in Figure 5.5. When the trade-off parameter α is increased from 3, in case of frame significance throughput utility, to 6, the minimum mean PSNR is increased from 27 dB to 29.4 dB, as shown in Figure 5.6. This increase is significant, since, as already stated, the satisfaction of well served users has a marginal improvement if their service level is increased further, compared to users having a lower level of service satisfaction. Note that the impact of the trade-off parameter is more evident when the PSNR based utility is used, as shown by Figure 5.7. When α is increased from 2.5 to 5, the performance of 22 percent high rate users is increased significantly, as shown by the meeting point of the two curves at 0.25 in Figure 5.7. Therefore, the perceived quality is approximately good to excellent for all the users according to the mean PSNR to MOS mapping, in case of PSNR based utility, as shown in Table 5.12. Hence, by increasing the trade-off parameter, the capacity gain can be further increased, that is, more than 100 users can be accommodated in the considered scenario. From the simulation results, a trade-off parameter value of 5, in case of PSNR based utility, is considered a good trade-off between efficiency and fairness, whereas for the utility metric based on the frame significance throughput, an α value of 6 achieves a good trade off between efficiency and fairness. Minimum capacity gain of 20 percent can be achieved with the proposed utility metric based on the frame significance throughput when α is set to 3, and more than 20 percent gain potential is available by increasing the trade off parameter above 3. The capacity gain can be further enhanced when the PSNR based utility with a higher value of α is used.

5.6 Conclusion

A multi-user downlink scheduling strategy is analyzed for delay sensitive video applications. The solution of the Nash product for the cooperative bargaining problem is exploited, where the eNodeB's scheduler is the central authority responsible to enforce the agreement in the centralized scheduling scenario. The concept of maximizing the product of the payoff of the different users is considered. The payoff is maximized by using the marginal utility principle. The considered approach makes use jointly of multiuser time, content and frequency diversity, by efficiently designing the utility and gain. A novel approach in the design of the utility metric, based on what defined as "frame significance throughput" is proposed, where a user receives payoff after transmitting a complete frame; the higher the priority of frame, the higher the payoff. The scheduling strategy based on the proposed utility metric outperforms a simple bit throughput based video scheduling as it takes into account different frame priorities. On the other hand it is less complex than the alternatively proposed utility relying on the full reference PSNR, which requires full information about the original frame sequences.

Simulation results show that at least 20 percent gain in capacity is possible as compared to the bit throughput based proportional fair scheduler which exploits time and frequency diversities. Furthermore, the scheduling framework gives the flexibility of varying the fairness and more capacity gain can be achieved by increasing the trade off parameter.

5.7 Extension of multi-user time diversity by considering the time-averaged bit throughput

A novel approach in the design of the utility metric based on frame importance maximizes the product of the payoff of different users. The NPS strategy based on the proposed utility metric outperforms a simple bit throughput based video scheduling as it takes into account different frame priorities. On the other hand it is less complex than the alternatively proposed utility relying on the full reference PSNR, which requires full information about the original frame sequences. It is important to note that multiuser diversity among the users has been exploited in the literature by considering the time-averaged bit throughput such as the classical PF rule in [64]. The NPS strategy utilizes frame significance throughput as the multi-user time-averaged diversity. Simulation results show that at least 20 percent gain in capacity is possible as compared to the bit throughput based proportional fair scheduler. In this section, the main goal is to add another dimension to the time-averaged multiuser diversity. Therefore OPF strategy is proposed, which is a function of of both frame significance throughput and time-averaged bit throughput.

The frame significance throughput considers the importance of a frame and provides fairness by prioritizing the most important video frames of all the flows. The main feature of the OPF strategy is to improve the exploitation of the multi-user time-averaged diversity. The time-averaged diversity is termed as the statistically independent variations among the video traffic rate of different users. In NPS strategy, the main goal of the scheduler was to maximize the product of the achieved video quality of each of the competing flows. The time-averaged bit throughput was not considered in the scheduling decisions. In this work, the main goal is to exploit the achieved time-averaged bit throughput w.r.t the achieved video quality thus adding another dimension to the multi-user time-averaged diversity. The details of the OPF strategy is given in the subsequent sections.

5.8 OPF scheduling framework

For delay sensitive traffic, [40] developed the M-LWDF scheduling strategy which assigns PRB φ at scheduling instant n to user i^* according to the following rule:

$$i^* = \text{argmax}\left(\gamma_i \frac{R_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} A_i^{(n)}\right) \tag{5.28}$$

where

$$R_{i,\text{ave}}^{(n)} = R_{i,\text{ave}}^{(n-1)}\,(1 - \frac{1}{n_w}) + \frac{1}{n_w}\,R_i^{(n-1)}. \tag{5.29}$$

where $A_i^{(n)}$ is the HOL packet delay and γ_i is a constant whose value is adjusted to account for different delay requirements for different users. $R_{i,\text{ave}}^{(n)}$ and $R_{i,\text{ave}}^{(n-1)}$ are the time-averaged rate, also called bit throughput, achieved over a transmission window of size n_w at scheduling epochs n and $n-1$ respectively, whereas $R_i^{(n-1)}$ is the number of bits sent at the previous scheduling epoch $n-1$ and $R_{i,\varphi}^{(n)}$ is the achievable instantaneous rate of user i on PRB φ.

According to the literature [42], among the existing state-of-the-art strategies this rule provides the best trade-off between fairness and efficiency for delay sensitive video traffic. This strategy is very simple to implement which is very important in LTE because the scheduling interval is 1 ms. However like the proportional fair strategy, this rule does not consider the video quality in the scheduling decision. As discussed above, content aware scheduling strategies such as in [79][87] [88] increase the computation complexity of the system as they require a full reference video quality metric.

The exploitation of the trade-off between fairness and efficiency is achieved by utilizing the time-averaged bit throughput and the proposed frame significance throughput based payoff. The frame significance throughput based payoff, $P_i^{(n)}(\xi_{\text{GOP},i})$, is derived from equation (5.10) (utility function used the NPS strategy) and is reported below:

$$\begin{aligned} P_i^{(n)}(\xi_{\text{GOP},i}) = \Big\{ &(1 - A_i^{(n)})\;\xi_{\text{GOP},i}(g-1) \\ &+ A_i^{(n)}\;\xi_{\text{GOP},i}(g) \Big\} \end{aligned} \tag{5.30}$$

At scheduling epoch n, PRB φ is assigned to user i^* satisfying the following rule:

$$i^* = \text{argmax}\left(\frac{R_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}}.(1 - P_i^{(n)}(\xi_{\text{GOP},i}))\right). \tag{5.31}$$

5.8.1 Multiuser time-averaged diversity

The scheduling rule consists of two PRB independent or time-averaged parameters. The payoff, $P_i^{(n)}(\xi_{\text{GOP},i})$, and the time-averaged rate, $R_{i,\text{ave}}^{(n)}$, parameters exploit statistically

independent multi-user time diversity. Since video traffic exhibits a variable rate characteristics, the use of these two parameters utilize the statistically independent variations among the video traffic rate of different users. In the event of users achieving equal payoff when scheduler sends the same proportion of significant frames, the priority will be given to the user achieving frame significance throughput with lower average rate. On the other hand, users attaining lower frame significance throughput either due to lower channel quality or higher frame size or both are prioritized by the term $1-P_i^{(n)}(\xi_{\text{GOP},i})$. The higher the significance and the lower the size of the scheduled video frame of a user, the higher the payoff and lower the average throughput for that user. Therefore, the ratio $\frac{1-P_i^{(n)}(\xi_{\text{GOP},i})}{R_{i,\text{ave}}^{(n)}}$ achieves fairness by considering the achieved bit throughput, the importance of the video frames sent and the HOL delay of the GOP g.

5.8.2 Multiuser frequency diversity

The instantaneous rate $R_{i,\varphi}^{(n)}$ on PRB φ based on the CQI feedback improves the system efficiency by exploiting the multi-user channel diversity in the frequency domain. In a wireless system, multi-path propagation effects cause a variable amount of fading on the PRBs of each user. Due to such variability in the fading, some of the PRBs of a user experience a higher amount of fading. Therefore, a parameter called relative strength of a PRB, $\theta_{i,\varphi}^{(n)}$, introduced in Section 5.3.6 quantifies the amount of fading on each PRB.

$$\theta_{i,\varphi}^{(n)} = \frac{R_{i,\varphi}^{(n)}}{\overline{R}_i^{(n)}} \tag{5.32}$$

where

$$\overline{R}_i^{(n)} = \frac{\sum_{m=1}^{M_{PRB}} R_{i,m}^{(n)}}{M_{PRB}}. \tag{5.33}$$

is the average PRB capacity of user i at scheduling instant n and M_{PRB} is the total number of PRBs in the system. This parameter assigns higher weights to the PRBs experiencing lower fading. Hence, least faded PRBs are assigned higher weights which not only improves system fairness but system efficiency as well. In order to exploit this parameter in the scheduling decision, the product of instantaneous rate $R_{i,\varphi}^{(n)}$ and relative strength $\theta_{i,\varphi}^{(n)}$ is considered as given below:

$$i^* = \text{argmax}\left[\frac{R_{i,\varphi}^{(n)}\,\theta_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}}.\left(1-P_i^{(n)}(\xi_{\text{GOP},i})\right)\right]. \tag{5.34}$$

Replacing in (5.34) the expression of $\theta_{i,\varphi}^{(n)}$ given in (5.32)

$$i^* = \text{argmax} \left[\frac{[R_{i,\varphi}^{(n)}]^2}{\overline{R}_i^{(n)}} . \frac{\left(1 - P_i^{(n)}(\xi_{\text{GOP},i})\right)}{R_{i,\text{ave}}^{(n)}} \right]. \tag{5.35}$$

At scheduling epoch n, the priority of user i depends upon two instantaneous parameters, $R_{i,\varphi}^{(n)}$ and $\overline{R}_i^{(n)}$, and two time-averaged priority parameters, $R_{i,\text{ave}}^{(n)}$ and $P_i^{(n)}(\xi_{\text{GOP},i})$.

5.8.3 Controlling fairness and efficiency

Controlling the trade-off between fairness and efficiency provides a flexible scheduling framework. In the priority function design, the payoff design quantifies the frame significance throughput and is the main source of controlling fairness among the users. On the other hand, the instantaneous rate $R_{i,\varphi}^{(n)}$ on each PRB controls the system efficiency. In order to enable the operator of the system the flexibility to easily set the fairness level, a fairness control parameter α_f and an efficiency control parameter α_c are utilized as given below:

$$i^* = \text{argmax} \left[\frac{(R_{i,\varphi}^{(n)})^{\alpha_c}}{\overline{R}_i^{(n)}} . \frac{\left(1 - P_i^{(n)}(\xi_{\text{GOP},i})\right)^{\alpha_f}}{R_{i,\text{ave}}^{(n)}} \right]. \tag{5.36}$$

The priority of user i on PRB φ is exponentially proportional to the instantaneous rate, $R_{i,\varphi}^{(n)}$, and the term $[1 - P_i^{(n)}]$ according to the operator defined parameters α_c and α_f respectively. A higher value of α_c results in higher efficiency as the system is highly channel aware. On the other hand, a higher value of α_f results in higher fairness among the users in terms of frame significance throughput. By varying these two parameters, a good trade off between efficiency and fairness can be achieved.

5.9 Performance evaluation

The performance of the proposed OPF scheduling rule is compared with the state-of-the-art strategies. The virtual token based and the delay based M-LWDF scheduling rule are considered as the benchmark. The virtual token based M-LWDF [70] scheduling rule is given as:

$$i^* = \text{argmax} \left(\gamma_i \frac{R_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} N_{Q_i}^{(n)} \right). \tag{5.37}$$

where $N_{Q_i}^{(n)}$ is the queue size of user i at the eNodeB. The delay version of the M-LWDF stated in equation (5.28) is also used as the benchmark. Both the scheduling rules exploit time diversity by using time-averaged throughput. In order to provide fairness, the M-LWDF rule uses the HOL delay whereas the M-LWDFQ uses the queue size of each user. The NPS strategy based on the frame significance throughput is also considered as the benchmark. The efficiency of the proposed scheduler with respect to the benchmark ones is analyzed in terms of video quality, in particular as average of the mean PSNR of all the users.

To measure the fairness, the standard deviation (equation (5.27)) of the meanPSNR is considered. The lower the standard deviation, the fairer the received video quality among the users. On the other hand, the higher the average of the mean PSNR the higher will be the system efficiency. The fairness and efficiency performance of all the considered priority functions on each PRB is discussed in the following subsections.

5.9.1 Simulation scenario

In order to evaluate the performance of the proposed scheduling strategy, scalable video transmission over an LTE system is simulated. The simulation parameters are reported in Table 5.13. Three video streams with different rates, resulting in different mean PSNR values are considered as reported in the Section 5.4. Similar to the previous simulation scenario, low, medium and high rate video sequences with different temporal complexities are considered. The main goal of the simulations is to analyze the fairness and efficiency performance of the NPS and OPF strategies under varying load and diverse channel characteristics with video streams having different average rate and mean PSNR characteristics. Initially 30 users are simulated which are divided equally into low, medium and high rate. The performance of the considered scheduling strategies is analyzed by adding more users in the network. The load is increased by adding 3 users, one user from each of the average rate traffic, in the cell until the total load is 42 users.

5.9.2 Results and discussion

In order to select the optimum values of α_c and α_f different combinations were simulated, not shown here for brevity. The OPF(α_c,α_f) rule with OPF(3,6) and OPF(4,6) provide a good trade off between efficiency and fairness. The fairness and efficiency performance at different load, *i.e.* number of users, is shown in Figure 5.8 where the number of users is reported in the square labels.

TABLE 5.13: Simulation parameters for video application on LTE schedulers.

PARAMETERS	VALUE
Bandwidth	5 MHz
Pathloss model	Hata-Cost-231 model (urban pathloss model)
HARQ	Synchronous retransmissions (Up to 3 retransmissions)
Channel	3GPP-TU (Typical Urban)
UE speed	15 to 100 km/h (UEs moving at variable speed)
Cell radius	1 Km
CQI averaging method	MIESM [20]
Maximum time duration of GOP in the buffer, Video frame rate	$D_{\max} = 500$ ms, 30 fps

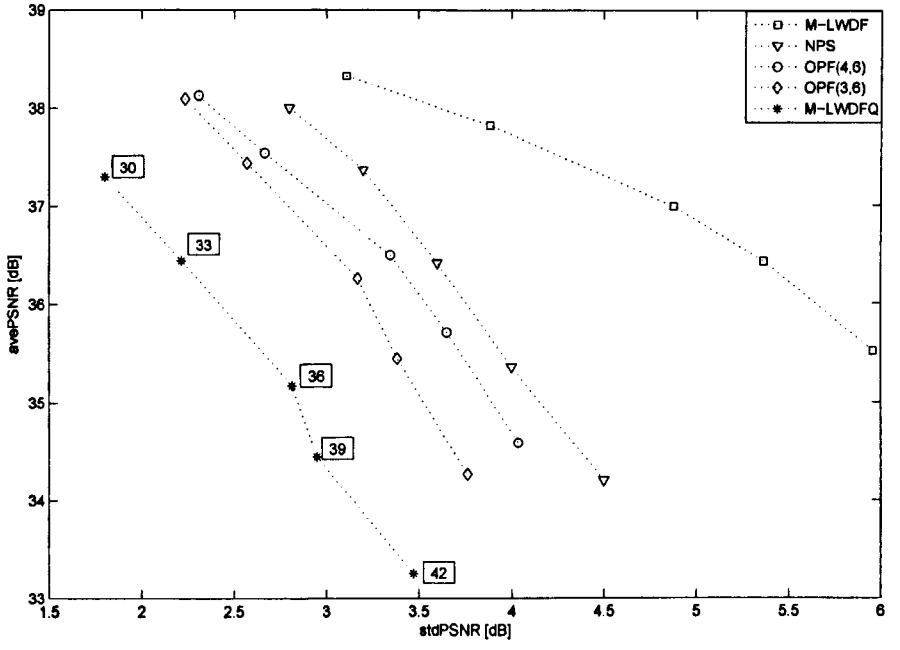


FIGURE 5.8: avePSNR vs. stdPSNR for different number of users.

According to Figure 5.8, the OPF scheduling rule outperforms the other strategies in terms of fairness and efficiency. The M-LWDF rule, widely considered the best one for our scenario, performs worse. Only the virtual token based M-LWDFQ rule provides better fairness, but at the expense of a lower average PSNR. For instance, when the number of users in the cell is 30, the average and the standard deviation of the mean PSNR with the OPF scheduling rule is approximately 38.1 dB and 2.5 dB respectively. On the other hand with the NPS strategy, the efficiency performance reduces to 37.9 dB and the standard deviation of the mean PSNR increases to 3 dB. The OPF scheduling rule considers both the frame significance throughput as well as the bit throughput which enhances the time-averaged diversity exploitation, hence users achieving frame significance throughput with lower average rate are prioritized by the OPF rule. On the other hand, the NPS strategy exploits time-averaged diversity by considering only the frame significance throughput. Furthermore, the use of the efficiency and fairness control parameters, α_c and α_f, effectively manages the trade-off between fairness and efficiency which is not the case in the NPS strategy. The advantage of the NPS strategy is that it requires very few parameters to calculate on each PRB. However with increased set of parameters, the computation complexity of the OPF rule is still manageable as it varies with the product of number of users and PRBs.

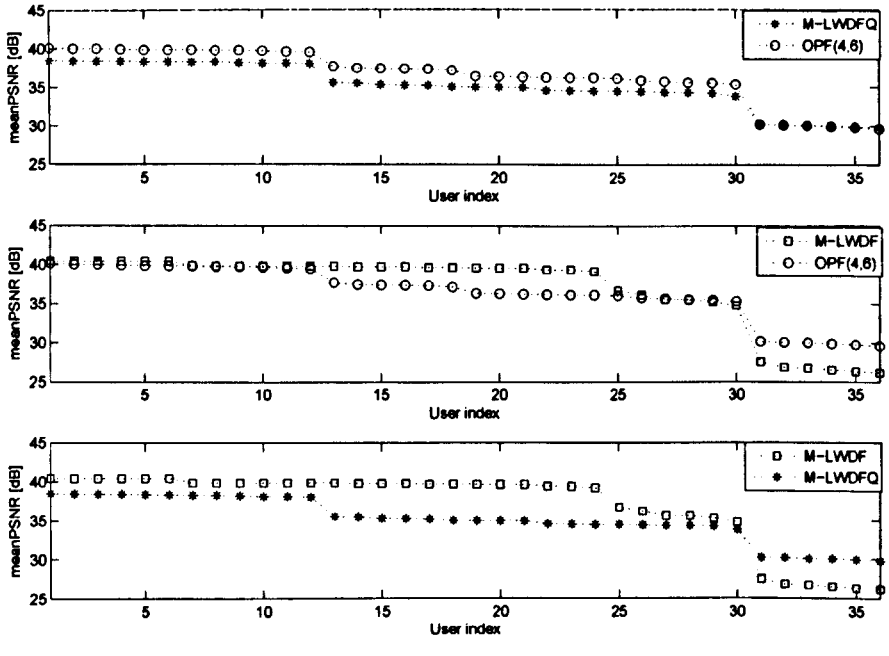


FIGURE 5.9: User index vs. meanPSNR for 36 users in the cell.

Figures 5.9 and 5.10 report the mean PSNR of all the users for the 36 and 42 users load respectively. The OPF and M-LWDFQ scheduling rules improve the performance of the

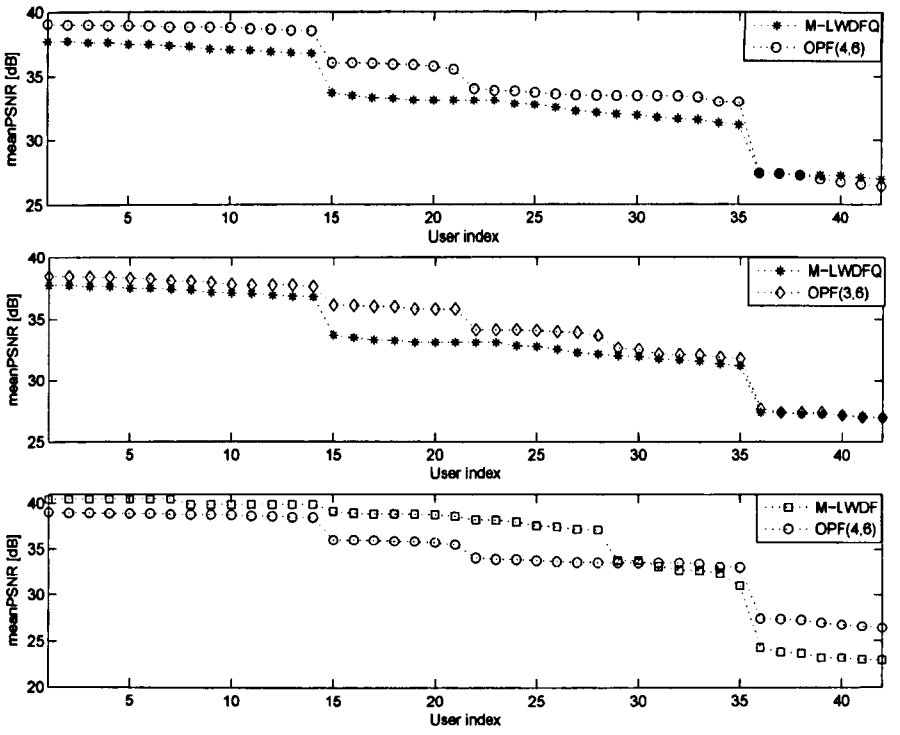


FIGURE 5.10: User index vs. meanPSNR for 42 users in the cell. Mean PSNR to perceived video quality is shown in Table 5.7.

worst served users as compare to the M-LWDF scheduling rule. However M-LWDFQ achieves it by degrading the performance of the best served users when compared to the M-LWDF rule. On the other hand the mean PSNR reduction of the best served users in OPF is minimal when compared to the M-LWDF scheduling rule. The OPF scheduling rule provides hence the best trade-off between fairness and efficiency by exploiting the statistically independent variability of the video traffic by using the payoff design which is a function of frame significance throughput and delay.

Under high load, the OPF priority rule schedules the most significant frames of the high rate and low channel quality users, thus dropping the remaining low priority frames and hence attaining an acceptable minimum video quality, as shown in Figure 5.10. When the performance of the M-LWDFQ and OPF priority rule is compared, the difference in the mean PSNR of some of well served users is as high as 4 dB, as shown in Figure 5.10. Thus M-LWDFQ provides fairness by significantly degrading the performance of the well served users. On the other hand, the M-LWDF scheduling rule achieves a mean PSNR of less than 25 dB ("poor" perceived video quality according to the mean PSNR to MOS mapping for 7 users in the cell.

Chapter 6 proposes a packet scheduling strategy based on QoE-based packet prioritization. Scheduling strategies proposed in this chapter consider only video streaming

traffic, whereas the scheduling framework proposed in Chapter 6 is flexible to accommodate all traffic types, delay sensitive and best-effort traffic, while providing a QoE-aware resource allocation.

Chapter 6

Prioritized packet scheduling for multi-class traffic over an LTE system

6.1 Introduction

For resource allocation, the optimization goal is either throughput maximization [89] [90] or the maximization of throughput with a fairness constraint, such as the consideration of queue length information in the scheduling decision [91] [92][93]. With content aware scheduling approaches, the optimization goal is the maximization of the video quality [94]. For instance in [95] [96], a concept of incrementally additive distortion among video packets is used to determine the importance of video packets for each user. Essentially, the increase in distortion due to the loss of a video packet is a function of all other video packets that are dependent on it and cannot be decoded if it is not sent. This information is used to drop video packets in the event of congestion over the wireless interface, beginning with the least important video packet. This buffer management strategy is combined with various scheduling approaches that set the priorities across users in order to either maximize throughput, ensure proportional fairness or minimize late packets. Scheduling across users, however, does not explicitly exploit the relative importance of video packets. In [80], the subflow concept is introduced in which a video flow (bitstream) is divided into several subflows based on their delay constraints as well as based on the relative priority in terms of the overall distortion of the decoded video. This is combined with a prioritized scheduling function of the 802.11e WLAN MAC. One drawback of this approach is its limitation to the 802.11e MAC which allows a limited number of priority classes. This results in limited subflow differentiation of the video stream and thus

limited gains from the multi-user content diversity. Furthermore, multi-user channel diversity is not exploited. In [81], a gradient-based scheduling and resource allocation algorithm is proposed, which prioritizes the transmissions of different users as a function of content, deadline requirements and past allocation history. Simulation results show that the proposed algorithm outperforms content-blind and deadline-blind algorithms with a gain of as much as 6 dB in terms of average PSNR in a congested network. This scheme does not explicitly consider channel conditions in its allocation process. In a wireless environment, this could lead to poor overall performance, with a few users with very poor channel conditions using almost all the available channel resources to satisfy their video quality requirements. In order to avoid this, it is proposed in [81] to deny access to users with channel quality less than a predetermined threshold, which limits wireless coverage. In [77], a heuristic approach is used to determine the importance of frames across users based on the frame types (I, P, or B), or their positions in a group of pictures. A priority based scheduling algorithm is proposed, with the priority function taking into account the importance of different frame types, channel conditions, buffer state and the relative start time of the video streams of the users. At the beginning of a time slot the scheduler computes the priorities of all users and schedules the one with the highest priority to transmit. This scheme is compared to non-content aware opportunistic scheduling and is shown to significantly improve the overall frame loss rate, and also to ensure that the higher priority frames have a lower frame loss rate.

The content-aware scheduling strategies presented in [95] [96] [80] [81] [77] [97] [82] require complex application layer information such as distortion, rate of each video layer, decoding deadline associated with each of the video packets. These algorithms require extensive cross-layer signaling [98][99] so that rate adaptation can be performed at the radio link layer. It is important to note that different requirements of video content information increase the scheduling complexity at the MAC layer. Therefore such scheduling algorithm may pose problems from an implementation point of view as the scheduling interval in LTE is only 1 ms. Another important point to consider is that these algorithms are designed specifically for video streaming applications. There is no information on the traffic handling of other traffic types like VoIP, video-conferencing or best-effort traffic. Unlike other previous works, [100] proposes a novel concept of utility based cross-layer optimization framework. The main idea is to exchange key parameters across the application, transport, link and physical layers. The cross-layer optimization framework allows the network operators to perform radio resource allocation based on users' satisfaction. The main steps involved in cross-layer optimization framework are:

- Collecting key information from each of the layers through cross-layer signaling;

- The cross-layer information is gathered in a decision center. The decision center performs the overall optimization by considering the variables from each of the layers like channel quality, time-averaged throughput, video content information and buffer status;

- After the joint optimization, the decision center sends the resource allocation information to the MAC layer. The scheduler then performs the packet scheduling based on the radio resource allocation decisions. If any rate adaptation is required then the optimizer sends the rate adaptation instructions to the server.

For solving wireless multimedia radio resource allocation problems there is a good amount of research work being carried out using cross-layer optimization techniques. For example, [101][102] [103][104] perform cross-layer optimization with the goal of performing efficient link layer packet scheduling for delay sensitive video traffic. The scheduler specifically considers the video content with the goal of maximizing the number of satisfied users. Similarly, [105][106] [107] [108] perform cross-layer optimization with the goal of efficient joint channel and source coding. [100] performs joint optimization of the link and physical layer with the goal of assigning the radio resources by considering the video content and channel quality.

The aforementioned cross-layer optimization techniques only consider the video application and design a scheduling rule by considering the video content and channel quality. [109] proposed a general resource allocation framework which can accommodate all the traffic classes (VoIP, video, web browsing, ftp). Authors proposed MOS as the optimization metric for each of the traffic types and design utility which quantifies users' satisfaction in terms of the MOS. [110] further extended the work by introducing different objective functions such as the modified max-min MOS where the objective function guarantees a minimum MOS to each of the flows. The cross-layer optimization technique by considering the MOS of each application shows a remarkable improvement in terms of the number of satisfied users as compared to the throughput based resource allocation schemes. The cross layer optimization technique considered in [109] [110] assumes that the video server is located close to the base station which allows quick rate adaptation of the video application. The rate adaptation is assumed to be performed by changing the quantization parameters of the video streams. However, video servers are located outside the wireless RAN thus limiting flexibility of delay free rate adaptation. Cross layer optimization techniques involve extensive cross layer signaling which increases the overheads and involves additional delay. In dynamic wireless environment, timely rate adaptation is very important. When the video server is located outside the RAN, the additional delay imposed in the end-to-end link probing from video server to the base station decreases the probability of timely rate adaptation. Thus, the performance of the

cross-layer optimization strategy decreases under congestion. Furthermore, the cross-layer optimization strategy requires the video server to support the required cross-layer signaling protocols such as [111] [112] [113] to support the video rate adaptation thus adding compatibility issues.

The European projects PHOENIX [107] [108] and OPTIMIX [94] proposed a framework where the cross-layer adaptation task is split in two main control entities, one at the application layer ("Application controller"), with the task of performing rate-control and adaptation based on information from the lower layers, and one at the PHY/MAC layer ("Base station controller"), with the task of adapting PHY/MAC parameters based on the characteristics of the video flows. The two controllers are supposed to work on different timescales and require information obtained through cross-layer signaling. The works in [114] [115] [116] [117] propose a realistic scenario in which the video streaming content is stored at the video server outside the RAN. In order to perform "in network rate adaptation" they propose two modules which are located inside the RAN. The two modules are the Traffic Engineering and Traffic Management module. The main task of the Traffic Management module is to act as the downlink optimizer for resource allocation, whereas the main task of the Traffic Engineering module is to act as a controller for performing rate adaptation in the RAN. With the two modules located in the RAN, the proposed resource allocation optimization cycle is 1 sec. The Traffic Engineering module performs rate adaptation either based on packet dropping or transcoding. [116] proposed three objective functions at the optimizer. One of the objective functions is the maximization of the MOS based utility, according to which the rate adaptation is done to maximize the mean MOS (mean user perceived quality). According to the objective function, resources are first reserved to the users with good channel quality and low rate demanding application. The authors also proposed a max-min fairness based objective function, where the main goal of the objective function is to allocate resources such that all the users get the same perceived quality. However, the authors do not propose any scheduling algorithm to be used in conjunction with the proposed cross-layer resource allocation frame work. The scheduling algorithm is important in determining the overall performance of an LTE system. In order to reduce the resource allocation optimization cycle to 1 sec, the framework proposed by [114] [115] [116] [117] requires extensive cross-layer signaling.

To reduce the required cross-layer signaling and constant end-to-end link probing, a novel cross-layer scheduling framework is proposed in this chapter which reduces the required cross-layer signaling for QoE aware resource allocation. The proposed work is based on the following key features.

- Application layer packet marking: The emerging SVC standard (scalable extension of the H.264/AVC standard) splits the encoded video stream in different layers of different importance. It is possible to mark packets belonging to different layers with layer information. The SVC standard facilitates the truncation of bit streams thus allowing graceful degradation of video quality in the event of wireless channel variations or network congestion. An SVC stream has a base layer and several enhancement layers. As long as the base layer is received, the receiver can decode the video stream. As more enhancement layers are received, the decoded video quality is improved. SVC consists of temporal, spatial and quality scalability. Temporal scalability refers to representing the same video in different frame rates. Spatial scalability refers to representing the video in different spatial resolutions or sizes. Normally, the picture of a spatial layer is based on the prediction from both lower temporal layers and spatial layers. Quality (or SNR) scalability refers to representing the same video stream in different SNR or quality levels. Video packet marking such as in [118] [119] [120] allows flexible rate adaptation at the link layer. The packet marking at the link layer is proposed which decreases the constant link probing from the base station to the video servers.

- QoE-curve based marking of SVC layers into priority classes (mapping scheme elaborated by Bo Fu, DOCOMO): The marking is done at the P-GW thus enabling the mobile operator to perform optimal prioritization achieving the maximum overall QoE under the constraint of network resources. Operators can define a utility based on bandwidth and MOS. Video layers achieving maximum MOS for a given bandwidth are prioritized thus allowing utility based prioritization. A simple example can be a QoE-bitrate utility as the priority metric, which prioritizes videos achieving higher MOS at a lower bit rate. Packets from these videos are assigned a higher priority class. Operators can define a fixed number of priority classes and can also mark VoIP, FTP and web browsing traffic to these priority classes. For instance, the video base layer and VoIP traffic can be assigned the same highest priority class and the least important priority classes can be assigned to delay tolerant traffic such as the FTP and web browsing.

- Opportunistic scheduler: The scheduler exploits the QoE-based packet marking. The main goal of the scheduler is to exploit the cross-layer information and minimize the delay bound violations of the most important priority class thus minimizing the delayed packet loss impact on the perceived video quality. The design goals of the prioritized packet scheduler are discussed in Section 6.3.

- Furthermore, rate adaptation at the link layer is proposed by applying class-based admission control on the video layers as discussed in Section 6.6.

TABLE 6.1: Mathematical symbols utilized in Chapter 6.

Symbol	Explanation
i, j, n	User/flow index, Priority class index, Current scheduling epoch.
$J_{\max}$	Maximum number of priority classes in the system.
$H_i^{(n)}$	HoL packet delay of user/flow i at scheduling epoch n.
$P_{i,j}^{(n)}$	Packet priority index of user i's HOL packet.
$\mathrm{PRF}_{i,\varphi}^{(n)}$	Priority function of user/flow i on PRB φ.
$\chi_{i,\varphi}^{(n)}$	Instantaneous Channel quality on PRB φ also known as the normalized subband spectral efficiency.
$\chi_i^{(n)}$	Average PRB quality, in terms of spectral efficiency, of flow i at scheduling instant n.
$\overline{\chi_i}^{(n)}$	The normalized average wideband channel quality of user i over the moving average window of size n_c epochs.
$\chi_{\max}$	A constant, *i.e.*, the spectral efficiency (5.5547 bits/sec/Hz) corresponding to the maximum CQI feedback.
M_{PRB}	The number of PRBs available for allocation at each scheduling epoch.
$W_i^{(n)}$	Weight of the HOL packet.
$R_{i,\mathrm{ave}}^{(n)}$ and $R_{i,\mathrm{ave}}^{(n-1)}$	Exponential time-averaged throughput (over the window of size n_w) at scheduling instant n and $n-1$.
$R_i^{(n-1)}$	Number of bits transmitted at scheduling epoch $n-1$.
$N_{Q_i}^{(n)}$	Number of packets residing in the queue of flow i at scheduling instant n.
$H_{\max}$	Maximum delay budget of user i's packet.
$A_i^{(n)}$	Normalized HoL delay. It is the ratio of $H_i^{(n)}$ and $H_{\max}$.
$\overline{A^{(n)}}$	Average of the normalized HoL delay taken over I flows present in the system.
β	Weight of the number of bits transmitted. This factor is multiplied by $R_i^{(n-1)}$. It is a function of $\delta_i^{(n)}$.
$\delta_i^{(n)}$	Weight of the bits sent by flow i at scheduling instant n and is used to control the system's fairness through parameter η.
η	A parameter controlling the impact of the averaged wideband channel quality on the number of bits sent.
Δ	A constant, its value is set to 1 in the simulations.
α_c	System efficiency control parameter.
α_f	Packet priority weight control parameter.
$\mathrm{plr}_{j^*}^{(t_w)}$	Packet loss rate of class j^* over t_w scheduling epochs.
$P_{\mathrm{transmit}_j^*}^{(m)}$	Number of transmitted packets of class j^* over the moving average transmission window t_w.
$P_{\mathrm{drop}}^{(m)}$	Number of dropped packets over the moving average transmission window t_w.
$H^{(t_w)}$	The congestion status of the network by calculating the average of the $\overline{A^{(n)}}$ over t_w scheduling epochs.
$I_{\mathrm{block}_{j^*}}$	Number of flows to block after t_w scheduling epochs.
$I_{\mathrm{re-admit}_{j^*}}$	Number of flows to readmit after t_w scheduling epochs.
S_{schmitt}	The output of the hysteresis based admission controller.
$S_{\mathrm{thr_l}}$ and $S_{\mathrm{thr_h}}$	Lower and higher threshold limits of the Schmitt trigger.
$P_{H^{(t_w)}}$	Probability of delay bound violation based on the congestion status of the network.
ω	Parameter used to Control the speed of the readmission.
C	A constant used to control the resource allocation between the delay sensitive and best-effort traffic classes.

The remainder of this chapter is organized as follows. Section 6.2 presents the considered system model, followed by the descriptions of the main goals of the scheduling strategy in Section 6.3. Section 6.5 presents the proposed Packet Priority Scheduler (PPS) and its performance under Variable Bit Rate (VBR) and Constant Bit Rate (CBR) traffic. In particular, after a description of the scheduling metric and an analysis of the factors composing it, the simulation scenario considered for its performance evaluation is described in 6.5.2 and results are reported in 6.5.3 and 6.5.4 for VBR and CBR traffic, respectively. After evaluating the efficiency and priority awareness characteristics of the PPS scheduling function, Section 6.6 presents a novel class based admission control policy. The class based admission control policy exploits the basic characteristics of the PPS scheduling function and performs rate adaptation. The joint PPS scheduling function and admission control policy is evaluated in Section 6.7 where a QoE based mapping of SVC layers to priority classes is considered. QoE based packet marking is performed at the core network, whereas at the eNodeB packet marking is exploited with the goal of achieving user satisfaction when the network is congested. A sum-MOS based packet marking (demanding video flows in terms of bitrate are less favored) is considered in Section 6.7. Finally, a composite scheduling rule for best-effort and delay-sensitive traffic is considered in Section 6.8. In this section, different traffic types marked with different packet priorities are considered. A max-min based packet marking (video users receiving same MOS quality) in Section 6.8 is considered. The packet marking of different traffic types helps in prioritizing the flows based on their channel quality, queue status and packet priority, thus fully exploiting multi-user channel diversity. Detailed simulation results of the composite scheduling rule are reported in Section 6.8.2, where the performance of the benchmark scheduling rules is also discussed.

6.2 System model

An OFDM SISO multiuser LTE system is considered with a single cell scenario in which the serving eNodeB is at the center of the cell. The serving eNodeB's MAC scheduler controls all the available PRBs by allocating them to the active flows competing for resources. A video server generating a pre-encoded video traffic workload is considered. Video is assumed to be encoded in different layers according to the SVC standard, and temporally organized in units which can be decoded independently from each other, each referred as GOP. Packets of the video stream are divided into priority classes, as shown in Figure 6.1, with each priority class representing the packet's importance towards the video quality. Each flow is assigned a buffer at the eNodeB. Furthermore, the buffer of each flow is divided into sub-queues depending upon the incoming packet's priority class as shown in Figure 6.1. The packets of each priority class entering into the buffer

are time stamped by the scheduler. The scheduler should assign enough resources to schedule the packet before its delay budget expires. Packets violating the delay budget are dropped from the queue. All the packets of the flows have the same delay budget irrespective of their priority classes. The priority of user i on class j at scheduling instant n is calculated according to the following rule:

$$P_{i,j}^{(n)} = J_{\max} - j_i^{(n)} \tag{6.1}$$

where $j_i^{(n)}$ is the priority class index of user i's packet and $J_{\max}$ is the maximum number of priority classes in the system and is fixed, *i.e*, it can be 4, 8 or 16. Table 6.2 reports the priority class index with the maximum number of priority classes in the system set to 8.

TABLE 6.2: Priority classes.

Priority Class Index (PCI), $j_i^{(n)}$	$P_{i,j}^{(n)}$
0 (most important class)	8
1	7
2	6
3	5
4	4
5	3
6	2
7 (least important class)	1

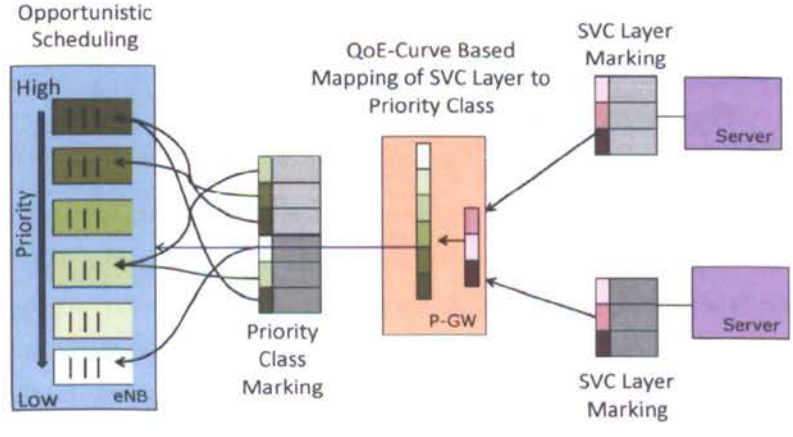


FIGURE 6.1: Considered cross-layer architecture [scheme elaborated by D. Staehle, DOCOMO.].

6.3 Goals

Delay based scheduling rules, such as the M-LWDF rule [42] (considered as the best delay-sensitive scheduling rule), are highly unfair for video flows characterized by high rate and low channel quality. For instance, as shown in Figure 6.2, high average throughput and low channel quality (consider the equation shown in Figure 6.2) of flow 1 causes the average amount of waiting time for the packets of flow 1 to be lower than that of flow 2. As shown in Example 1 of the figure, these scheduling rules achieve higher average waiting time for the packets of flows having higher average throughput and lower channel quality. On the other hand, flows having good channel quality and lower average throughput have a very low average waiting time. When the system load is increased, the probability of delay bound violation of the lower channel quality and higher average throughput flows is very high, which results in an unfair system.

Consideration of packet priority in the scheduling decision of such rules can reduce the system efficiency. For instance, consider Example 2 in Figure 6.2 where each flow has two packet priority classes. In order to avoid the delay bound violation of the high rate and low channel quality flows, the priority difference between the two classes should be very high. If the M-LWDF (one of the best delay based scheduling rule) priority function is weighted with the packet priority, then higher priority packets entering the buffer will have a minimal amount of waiting time, *i.e.*, higher priority packets entering the buffer are instantly served irrespective of the channel quality of other flows. This reduces the exploitation of the most important phenomenon in wireless LTE systems, *i.e.*, multi-user channel diversity. In LTE, multi-user channel diversity has more significance since it is a multi-carrier system which has another dimension of exploitation known as the multi-user frequency diversity. However, weighing delay aware schedulers with packet priority results in a less complex system as such scheduling rules are simple to implement. Low complexity scheduling rules are very important in LTE mainly because of the short scheduling interval of 1 ms.

In this chapter, the main goal is to address the aforementioned weaknesses in an innovative way such that the low complexity feature of such scheduling rules is preserved. The delay urgency and packet priority are utilized in determining the weight of the queue. The weight of the queue is dynamic and changes according to the system load. For instance if a flow has a good channel quality and is well served, *i.e.*, the probability of delay violation is low, then such a flow does not require any packet prioritization. On the other hand, a flow having lower channel quality and higher HoL delay requires packet prioritization, so that under high load the most important packets are scheduled within the delay bound. The the packet priority index is utilized by designing a function which changes according to the system load and flow's HoL packet delay. The scheduling

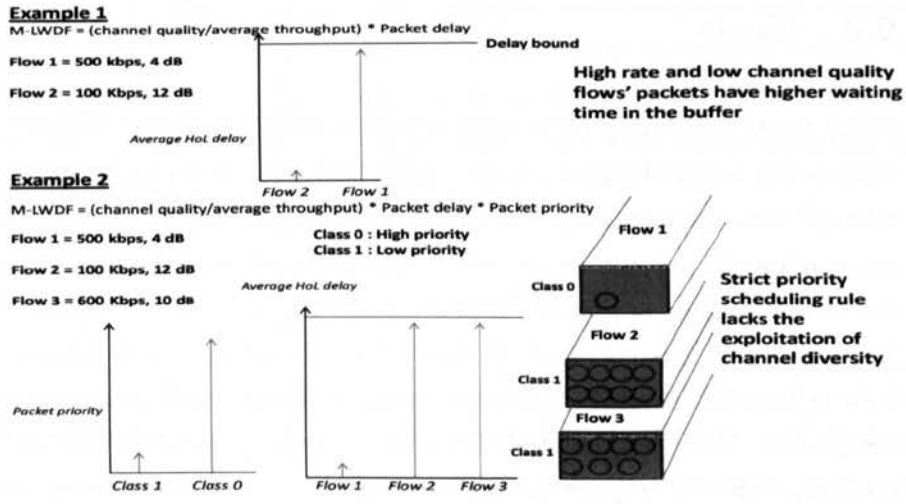


FIGURE 6.2: Issues of the delay based and strict priority scheduling rules.

design makes use of the packet priority in a novel way by exploiting packet's HoL delay and the overall system delay. The main objective of the scheduler is to minimize the probability of delay bound violation of the most important packets without compromising much on the system efficiency. Another design goal is to robustly tune the balance between packet priority and system efficiency. To summarize, the main goal to achieve a low complexity scheduling strategy which provides a trade-off among the packet priority, HoL delay, channel quality and time-averaged throughput.

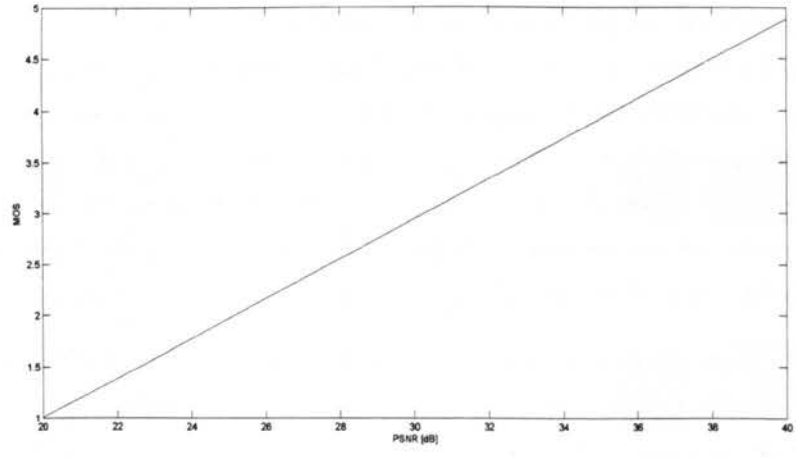


FIGURE 6.3: Linear relationship between PSNR and MOS.

6.4 MOS Model

In order to achieve MOS based packet marking, a linear relationship between PSNR and MOS mapping has been utilized as shown in Figure 6.3. QoE actually encompasses many different aspects and objective video quality is just one of them. There are many QoE metric available in the literature such as in [121] [122] [123] [124] each with its merits and demerits. Although PSNR is not an accurate QoE metric, as it has its share of weaknesses as outlined in [125], PSNR is a widely utilized metric in the literature because of its simplicity and fast computation characteristics.

The main feature of the considered cross-layer architecture is its flexibility to be used with any video quality metric. Only a mapping function between the objective video quality metric and MOS is required. The linear mapping between PSNR and MOS has been widely used in the literature such as in [114] [115] [116] [117].

6.5 Proposed scheduling strategy and metric

In the proposed MAC layer scheduling strategy, each PRB is assigned to the user maximizing a defined scheduling metric.

The following scheduling metric is proposed:

$$\mathrm{PRF}_{i,\varphi}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{R_{i,\mathrm{ave}}^{(n)}} * W_i^{(n)}[N_{Q_i}^{(n)}]. \tag{6.2}$$

This priority function design depends upon four parameters as shown in Figure 6.4:

- $N_{Q_i}^{(n)}$ is the number of packets currently residing in the buffer of flow i at scheduling instant n.
- $W_i^{(n)}$ is the weight of the HoL packet. Mathematically, it is defined as:

$$W_i^{(n)} = \exp\left\{\left[P_i^{(n)}\right]^{\overline{A^{(n)}}} * A_i^{(n)}\right\} \tag{6.3}$$

where

$$\overline{A^{(n)}} = \frac{1}{I}\sum_{i=1}^{I} A_i^{(n)} \tag{6.4}$$

and

$$A_i^{(n)} = \frac{H_i^{(n)}}{H_{\max}}. \tag{6.5}$$

$\overline{A^{(n)}}$ is the average of the normalized HoL delay taken over the I flows present in the system. $\overline{A^{(n)}}$ is also known as the normalized system delay at scheduling epoch n. $H_i^{(n)}$ is the HoL packet delay and $H_{\max}$ is the maximum delay budget of user i's packet at scheduling instant n. $P_{i,j}^{(n)}$ is the priority value of the class according to Table 6.2. It is important to note that the proposed weight design depends on the system load. The higher the system load, the higher will be the normalized system delay $\overline{A^{(n)}}$, which results in a higher weight for the packets of the most important priority classes. If the system delay is low, packets from different priority classes have approximately the same weights. Considering that the packet priority in the scheduling decision can reduce the system throughput, the main goal in the this work is to design a scheduling strategy where the packet priority changes according to the system load.

- $\chi_{i,\varphi}^{(n)}$ is the channel quality on PRB φ.
- $R_{i,\text{ave}}^{(n)}$ is the time-averaged throughput. Mathematically, it is defined as:

$$R_{i,\text{ave}}^{(n)} = R_{i,\text{ave}}^{(n-1)} * (1 - \frac{1}{n_w}) + \frac{1}{n_w} * \beta R_i^{(n-1)}. \tag{6.6}$$

where $R_{i,\text{ave}}^{(n-1)}$ is the average throughput at scheduling instant $n-1$. $R_i^{(n-1)}$ is the number of bits transmitted at scheduling instant $n-1$. n_w is the size of the time-average window. β is the weight of the number of bits transmitted, discussed in Section 6.5.1.2.

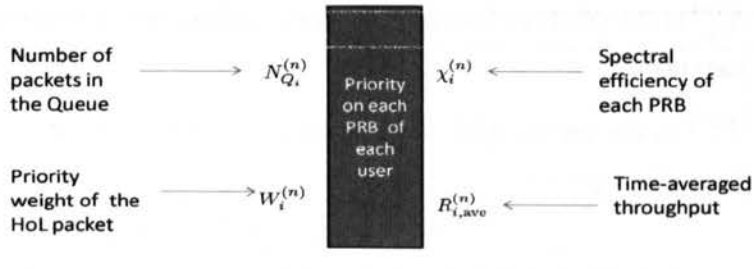


FIGURE 6.4: Prioritized packet scheduling rule.

6.5.1 Analysis of the factors composing the priority function

The significance of the weighted queue rule (product of priority weight and queue size) and the proportional fair rule (ratio of channel quality and time-averaged throughput) in the scheduling design are discussed below.

6.5.1.1 Weighted queue component

The delay based priority weight makes the scheduling rule dynamic: when the system load increases, the probability of delay bound violations also increases. The impact of the exponential weight design in equation (6.2) can be controlled by introducing a tunable parameter ζ as shown in equation 6.7.

$$W_i^{(n)} = \exp\left\{\frac{\left(P_i^{(n)} * \zeta\right)^{\overline{A^{(n)}}}}{\zeta} * A_i^{(n)}\right\}. \tag{6.7}$$

The weight $W_i^{(n)}$ in the scheduling decision prioritizes the most important packets. If the system load is low, packets from different priority classes have approximately the same weight. The impact of the exponential weight at different system loads and HoL delays is shown in Figures 6.5 and 6.6. According to Figure 6.6, packets with PCI of 0 and 4 have approximately the same weights until a normalized average delay of 0.7 when ζ is set to 10. On the other hand, when ζ is set to 1 the priority difference between the packets of the two priority classes increases exponentially after normalized average delay of 0.5.

6.5.1.2 Time-averaged throughput

The exponential averaging method has a higher convergence time (convergence to the time-averaged throughput). This method is highly suitable for non-real-time applications as it achieves fairness in terms of throughput over a longer time scale. For delay sensitive applications, however, the convergence time is greater than the packet's delay budget. Therefore, the number of transmitted bits is weighted with the normalized averaged wideband channel quality. The normalized average wideband channel quality (expressed in terms of spectral efficiency) $\overline{\chi}_i^{(n)}$ of user i over the moving average window of size n_c is given as:

$$\overline{\chi}_i^{(n)} = \frac{1}{\chi_{\max}}\left[\frac{1}{n_c}\sum_{k=n-n_c}^{n}\chi_i^{(k)}\right] \tag{6.8}$$

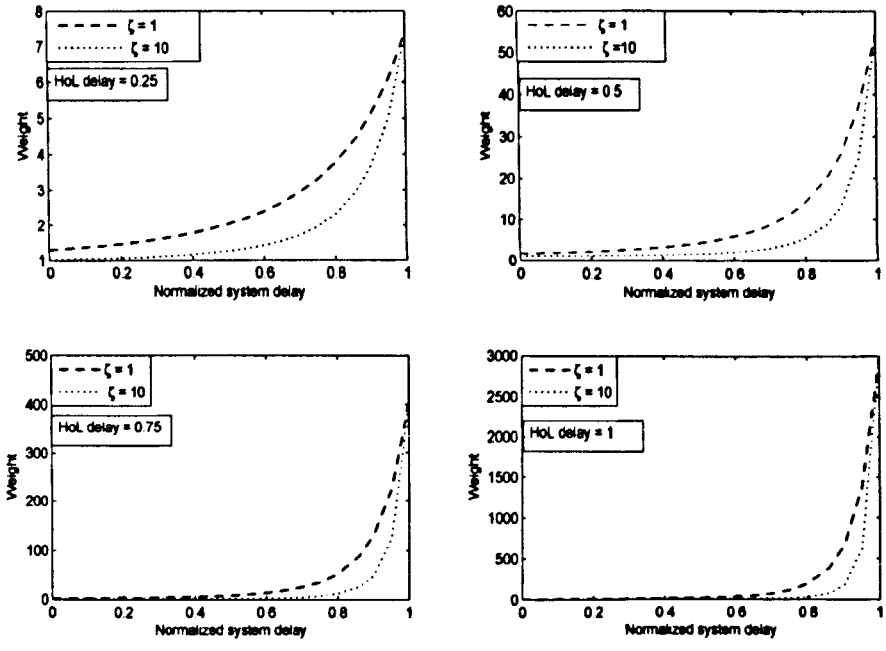


FIGURE 6.5: Impact of the exponential weight at different normalized system delay $\overline{A^{(n)}}$ and HoL delay $A^{(n)}$.

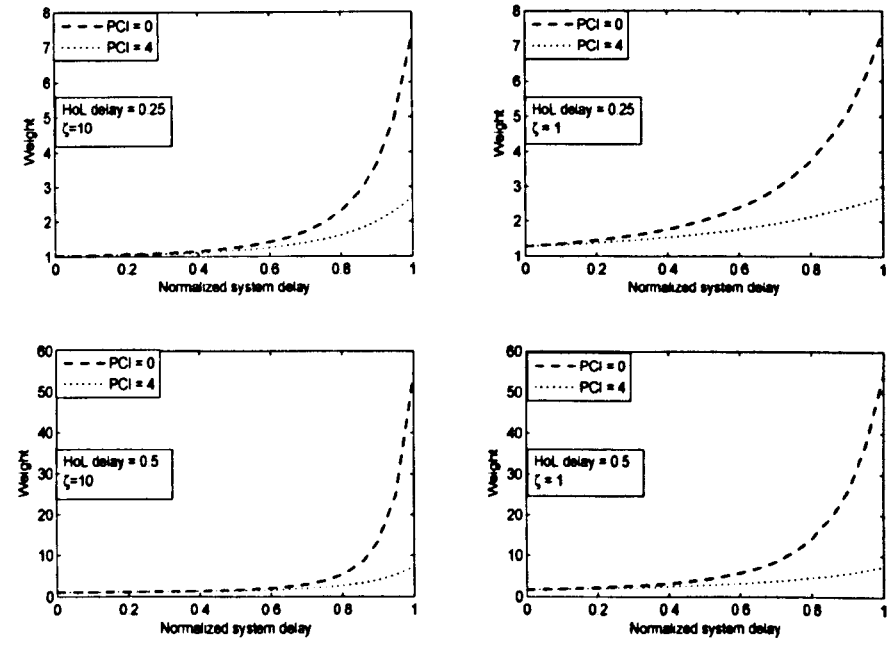


FIGURE 6.6: Exponential weight difference for different priority classes.

with

$$\chi_i^{(n)} = \frac{1}{M_{\text{PRB}}} \sum_{\varphi=1}^{M_{\text{PRB}}} \chi_{i,\varphi}^{(n)} \tag{6.9}$$

where $\chi_i^{(n)}$ is the wideband spectral efficiency of user i at scheduling instant n and $\chi_{i,\varphi}^{(n)}$ is the subband spectral efficiency of user i at PRB φ. $\chi_{\max}$ is the maximum spectral efficiency in terms of the CQI feedback. In order to utilize the averaged wideband channel quality, the parameter β is defined below:

$$\beta = \begin{cases} \delta_i^{(n)}, & \text{if } A_i^{(n)} < 1 \\ \frac{-n_w}{R_i^{(n-1)}}, & \text{if } A_i^{(n)} == 1 \end{cases} \tag{6.10}$$

where

$$\delta_i^{(n)} = \eta * \overline{\chi}_i^{(n)} + (1 - \eta) * \Delta. \tag{6.11}$$

η is used to control the impact of the averaged wideband channel quality on the number of bits sent. Δ is a constant whose value is set to 1 in the simulations; when η is set to zero all the flows have the same $\delta_i^{(n)}$, i.e., equal to Δ. Typical values of η lie in between 0.2 and 0.5. The higher the value of η, the lower will be the system efficiency. $\delta_i^{(n)}$ is the weight of the bits sent by flow i at scheduling instant n and is used to control the system's fairness through parameter η.

With this definition, the time-averaged throughput is defined as:

$$R_{i,\text{ave}}^{(n)} = R_{i,\text{ave}}^{(n-1)} * (1 - \frac{1}{n_w}) + \frac{1}{n_w} * \beta R_i^{(n-1)}. \tag{6.12}$$

It is important to note that the main motivation of considering the time-averaged channel quality is to decrease the difference in the average waiting time between the flows having diverse channel quality and average rate requirements as shown in Figure 6.7. This leads to an increase in the average system delay $\overline{A^{(n)}}$, which in turn leads to the prioritization of the higher priority packets.

The penalty associated with packet's delay budget violation is $\frac{-n_w}{R_i^{(n-1)}}$. $R_{i,\text{ave}}^{(n)}$ will decrease by unity whenever there is a delay violation, *i.e.*, $A_i^{(n)} == 1$. Under high load, the probability of packet's delay bound violation increases. It is important to note that

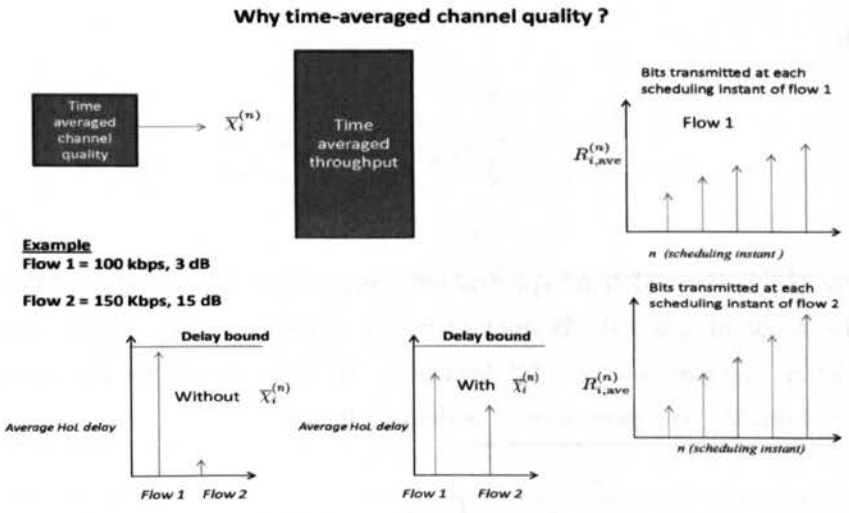


FIGURE 6.7: Significance of time-averaged channel quality.

flows' packets violating the packet delay bound are dropped from the queue. When a packet is dropped from the queue, the priority of the flow is reduced as the number of packets N_Q residing in the buffer is decreased. In order to prioritize such flows, $R_{i,\text{ave}}^{(n)}$ is decreased by unity upon every delay bound violation.

In order to manage the trade-off between system efficiency and packet priority awareness, two control parameters are proposed: an efficiency parameter, α_c, and a packet priority weight control parameter, α_f. Hence, the priority function of the PPS(α_c,α_f) rule proposed in (6.2) becomes:

$$\text{PRF}_{i,\varphi}^{(n)} = \frac{[\chi_{i,\varphi}^{(n)}]^{\alpha_c}}{R_{i,\text{ave}}^{(n)}} * [W_i^{(n)}]^{\alpha_f} [N_{Q_i}^{(n)}]. \tag{6.13}$$

6.5.2 Simulation scenario

In order to assess the performance of the proposed scheduling function, users with different video rate characteristics are simulated. Specifically, VBR and CBR video flows are considered. The main rationale for selecting different video rate flows is to analyze the priority awareness and system efficiency performance of the PPS scheduling rule under different traffic types. Specifically, for VBR traffic high, medium and low rate users with average rates of 500 kbps, 200 kbps and 140 kbps respectively are considered. Furthermore, all the traffic exhibits a peak rate which is equal to twice the average rate. Users are distributed uniformly in the cell. The performance of each scheduling rule is analyzed at different load scenarios. In all the scenarios, the users are divided equally

into the low, medium and high rate video streams. Initially, 42 users are simulated, with 14 users from each of the low, medium and high rate video sources. This corresponds to an average input traffic rate of 11.76 Mbps with an average spectral efficiency requirement of 2.352 bits/sec/Hz (considering a 5 Mhz bandwidth). The channel quality (in terms of SINR) of each of the video users is set such that the system capacity is 11.5 Mbps (2.3 bits/sec/Hz). Therefore, there is a higher packet delay bound violation probability as the average input traffic is higher than the system capacity. The input average traffic is further increased by increasing the number of video users to 48 (adding 2 video users from each of the considered video traffic sources); the system capacity in terms of bits/sec/Hz is kept the same. For the next two scenarios, the load is further increased to 54 (18 video users from each of the considered video sequences) and 60 (20 video users from each of the considered video sequences) users. The average spectral efficiency requirement of the input average traffic in the 60 user case is 3.36 bits/sec/Hz as compared to the system capacity of 2.352 bits/sec/Hz.

The average rate of the important priority classes (class 1 and 2) is greater than the lower priority classes. These priority classes correspond to the base layer video quality. Thus, if a user receives these classes with no packet losses, then a base layer video quality is guaranteed. On the other hand, if delay bound violations occur for a base layer video quality class, then the higher layer priority classes cannot be decoded. These traffic characteristics are selected after analyzing several videos with different motion and spatial characteristics. Specifically, videos with one base and one enhancement layer is chosen. The enhancement layer is further divided into 6 sublayers according to the medium granular scalability (MGS) SVC codec. In other words, class 1 and class 2 correspond to the base layer video quality and the remaining 6 priority classes correspond to the MGS sublayers. Table 6.3 reports the simulation parameters adopted for the LTE system and the wireless channel.

The following scheduling rules are considered in the simulations:

- Scheduler 1: Proposed PPS.
- Scheduler 2: M-LWDF scheduler [42].
- Scheduler 3 : Queue-aware M-LWDF rule (M-LWDFQ) scheduler [70].

The considered benchmark scheduling rules, M-LWDF and M-LWDFQ, do not consider the packet priority in the scheduling function. The main motivation is to analyze the system efficiency penalty (through system packet loss ratio) when packet priority is considered in the scheduling decision. Since M-LWDF gives better performance for videos, therefore the delay and queue based schedulers as a benchmark.

TABLE 6.3: Simulation parameters - Downlink LTE scheduling for multi-class traffic.

PARAMETERS	VALUE
Bandwidth, Carrier frequency	5 MHz, 2.1 GHz
UE distribution, Cell radius	Uniform, 1 km
Channel	3GPP-TU (Typical Urban)
Pathloss model	Hata-Cost-231 model
HARQ	Up to 3 synchronous retransmissions
Channel Fading	Block Fading (1 ms)
UE speed	15 to 100 km/h (users moving independently at variable speed)
CQI averaging method	MIESM [19] [20]
packet's delay budget $H_{\max}$	200 ms

6.5.3 Performance of the scheduling function under VBR traffic

For each scenario, results are reported in the form of a table (presenting the system PLR of each priority class) and a figure which presents the PLR of each user. The simulation results of 42 users are given in Table 6.4 and Figure 6.8. Similarly the simulation results of 48, 54 and 60 users are given in Tables 6.5, 6.6, 6.7 and Figures 6.9, 6.10, 6.11 respectively. The tables present the packet losses in each of the priority classes whereas the fairness performance of each of the considered scheduling rules is shown in the figures.

In the table, N_t denotes the total number of packets streamed into the buffer of each user from each priority class. The table also reports the packet loss ratio of each scheduling rule. Table 6.8 presents the system packet loss ratio of each of the considered scheduling rules. The system packet loss ratio represents the system efficiency by considering the total number of scheduled and dropped packets.

Some observations on the performance of the considered scheduling rules are reported below:

- The M-LWDF scheduling rule penalizes the high rate and low channel quality flows and shows a poor fairness performance at all load conditions as reported in Figures 6.8, 6.9, 6.10 and 6.11. It is important to note that PLR above 10 % in the most important layers results in a major degradation of the video quality. For high motion videos, the acceptable PLR of the important layers is as low as 2 to 5 %. Considering 10 % as the overall acceptable PLR (if dropping occurs in the important priority classes) then at 42 user load, the M-LWDF scheduling rule accommodates only 36 users. When load is increased to 48, the total number of

satisfied users (having PLR greater than 10 %) is 36. Similarly the total accommodated users is 34 for 54 and 60 users respectively. The M-LWDF scheduler suffers from poor fairness characteristics and severely penalizes low channel quality and high rate users.

- The M-LWDFQ scheduler considers the queue size. According to Table 6.8, this scheduler achieves the best system efficiency. However, if 10 % is considered as the overall acceptable PLR (if dropping occurs in the important priority classes) then, at 42 user load, the M-LWDF scheduling rule accommodates only 36 users. When the load is increased to 48, the total number of accommodated users (having PLR less than 10 %) is 24. Similarly when the load is increased to 54, the total number of accommodated users is only 20 and at the highest load approximately all users have PLR of more than 10 %. This scheduler gives good fairness in terms of PLR and achieves the best efficiency performance among all the scheduling rules. However, this scheduler requires a strict admission control since allowing more users to enter the system would increase the system throughput but decrease the overall capacity in terms of the number of satisfied users.

- At lower load (when system PLR is less than 10 %) the proposed PPS scheduler gives more priority to the good channel flows as the normalized average system delay is low. At higher load, the scheduler delivers the most important layers to each of the flows thus improving the system fairness and decreasing the system efficiency. The PPS scheduling strategy does not require a strict flow admission control. The exponential priority weights increases the resource allocation probability of important priority classes thus reducing the packet losses in these classes. At higher load (60 users case) PPS strategy with $\zeta = 1$ performs better than $\zeta = 10$. It is important to note that ideally we would like to achieve results where we receive a 100 % PLR for the least important classes and the PLR for important priority classes is as low as 0 %. In the current scenario, any further increase in load would cause PLR in higher priority classes. One of the main goals is to achieve a resource allocation strategy where no flow based admission control (denying admission to new flows) policy is required. Therefore, the PPS scheduling function along with the novel priority class based admission control discussed in Section 6.6 where rate adaptation is achieved through blocking of flows' less important priority classes instead of completely blocking the flows from the network. In the next section, the performance of the scheduling function under CBR traffic is analyzed.

TABLE 6.4: Packet loss ratio of each traffic class for different priority functions. Total number of users in the cell is 42.

Priority Class Index	N_t	Packet loss ratio, PPS($\zeta = 1$)	Packet loss ratio, PPS($\zeta = 10$)	Packet loss ratio, M-LWDF	Packet loss ratio, M-LWDFQ
0	2983	0	0	0.042	0.0459
1	2633	0	0.0005	0.0621	0.0740
2	1673	0	0.0017	0.0381	0.0462
3	1754	0	0.0040	0.0448	0.0335
4	1980	0.0327	0.0152	0.0400	0.0151
5	2360	0.0586	0.0330	0.0604	0.0233
6	2320	0.2007	0.0942	0.0867	0.0283
7	2276	0.3533	0.1422	0.0807	0.0261

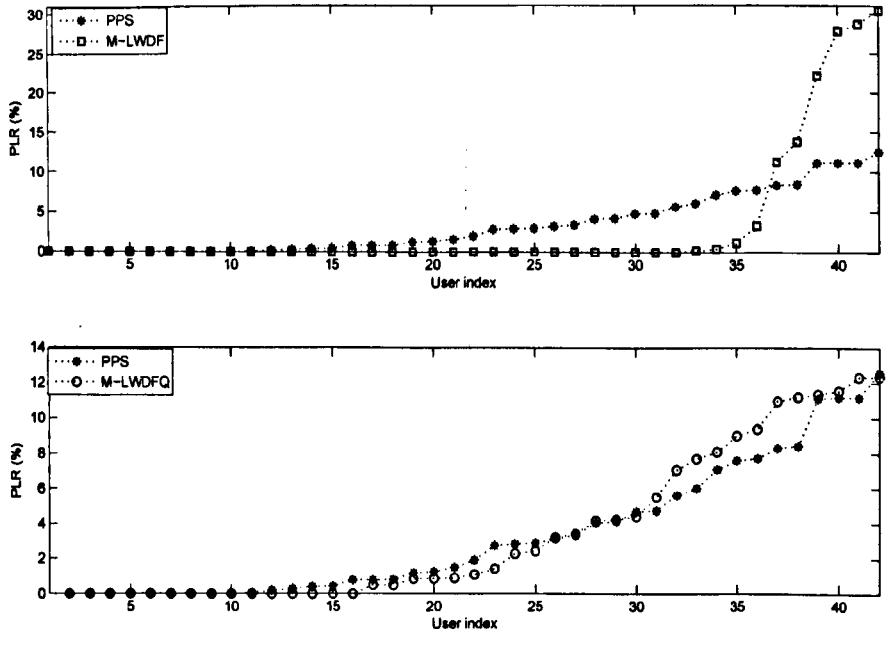


FIGURE 6.8: PLR(%) vs. user index of different priority functions. The total number of users in the cell is 42.

TABLE 6.5: Packet loss ratio of each traffic class for different priority functions. The total number of users in the cell is 48.

Priority Class Index	N_t	Packet loss ratio, PPS($\zeta = 1$)	Packet loss ratio, PPS($\zeta = 10$)	Packet loss ratio, M-LWDF	Packet loss ratio, M-LWDFQ
0	2983	0	0	0.1741	0.1921
1	2633	0	0.00095	0.2575	0.3646
2	1673	0.0015	0.0082	0.2213	0.3474
3	1754	0.0028	0.0070	0.1710	0.1428
4	1980	0.1517	0.1212	0.1661	0.0884
5	2360	0.1907	0.1547	0.2210	0.1367
6	2320	0.4585	0.3653	0.2672	0.1887
7	2276	0.6261	0.5030	0.2493	0.1459

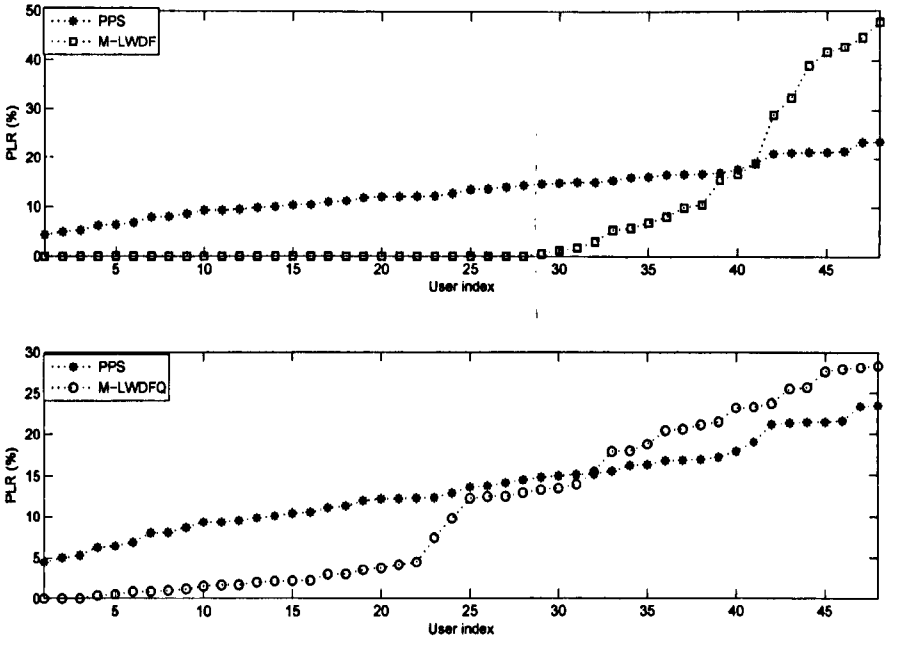


FIGURE 6.9: PLR(%) vs. user index of different priority functions. The total number of users in the cell is 48.

TABLE 6.6: Packet loss ratio of each traffic class for different priority functions. The total number of users in the cell is 54.

Priority Class Index	N_t	Packet loss ratio, PPS($\zeta = 1$)	Packet loss ratio, PPS($\zeta = 10$)	Packet loss ratio, M-LWDF	Packet loss ratio, M-LWDFQ
0	2983	0	0	0.1741	0.1921
1	2633	0.0017	.0072	0.2575	0.3646
2	1673	0.0329	.0336	0.2213	0.3474
3	1754	0.0592	.0468	0.1710	0.1428
4	1980	0.2582	.2392	0.1661	0.0884
5	2360	0.3526	.3380	0.2210	.1367
6	2320	0.6312	.59	0.2672	0.1887
7	2276	0.7273	.6623	0.2493	0.1459

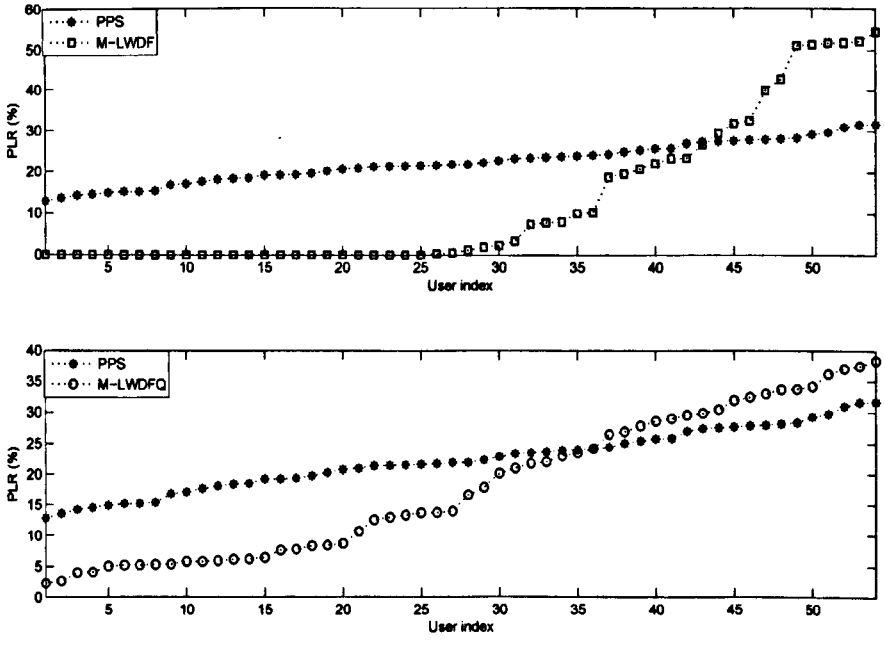


FIGURE 6.10: PLR(%) vs. user index of different priority functions. The total number of users in the cell is 54.

TABLE 6.7: Packet loss ratio of each traffic class for different priority functions. The total number of users in the cell is 60.

Priority Class Index	N_t	Packet loss ratio, PPS($\zeta = 1$)	Packet loss ratio, PPS($\zeta = 10$)	Packet loss ratio, M-LWDF	Packet loss ratio, M-LWDFQ
0	2983	0	0	0.2511	0.2736
1	2633	0.0084	0.0228	0.3178	0.4730
2	1673	0.0952	0.1333	0.2579	0.4291
3	1754	0.1277	0.1961	0.2198	0.1935
4	1980	0.3434	0.4015	0.2173	0.1259
5	2360	0.4479	0.5216	0.2742	0.2020
6	2320	0.7289	0.6806	0.3509	0.2435
7	2276	0.83	0.7425	0.3263	0.2049

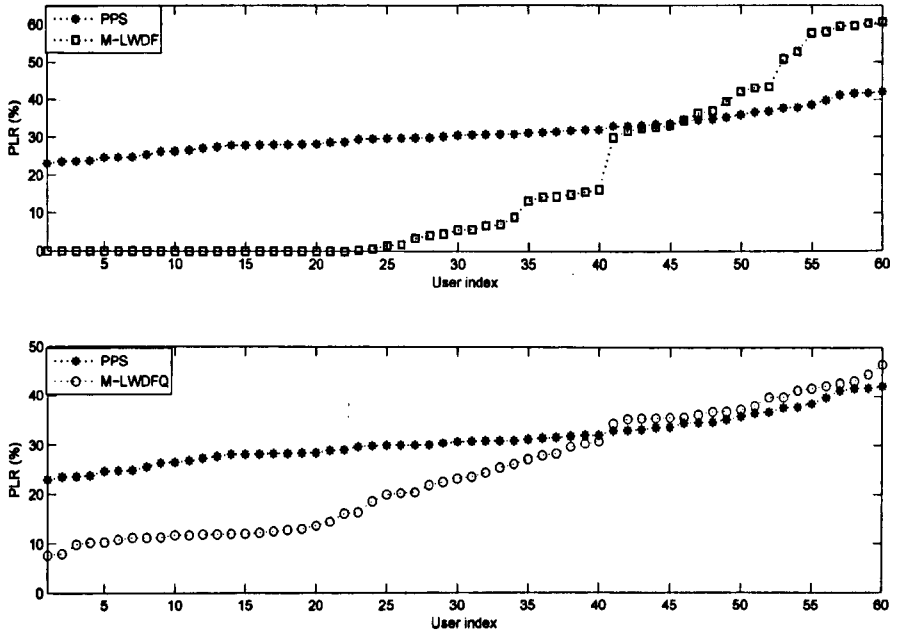


FIGURE 6.11: PLR(%) vs. user index of different priority functions. The total number of users in the cell is 60.

TABLE 6.8: System packet loss ratio.

Load	PPS with $\zeta = 1$	PPS $\zeta = 10$	M-LWDF	M-LWDFQ
42 users	.0788	0.0348	0.0591	.0341
48 users	0.1746	0.1391	0.1419	0.1086
54 users	0.2483	0.23	0.2210	0.1836
60 users	0.3070	0.3173	0.2829	0.2468

6.5.4 Performance of the scheduling function under CBR traffic

This section analyzes the performance of the channel and weight control parameters α_c and α_f. The simulation results in the above section reveal that setting $\zeta = 1$ results in better performance at high load. Therefore, a simple exponential weight reported in equation (6.3) is utilized. A high load scenario of 600 Kbps CBR flows is selected. For CBR traffic, equal amount of packets are streamed into the buffer of each traffic class. The input average traffic rate for the CBR traffic is 21.6 Mbps (36 CBR users with 600 Kbps rate, requiring an average spectral efficiency of 4.32 bits/sec/Hz as compare to the same system capacity as in the VBR scenarios (2.3 bits/sec/Hz)).

The results are reported in the form of a table (presenting the system packet loss ratio of each priority class) and a figure which presents the overall system packet loss ratio and throughput. The simulation results of 36 users are given in Table 6.9 and Figure 6.12. According to Figure 6.12, the PPS scheduling function with $\alpha_c = 2$ and $\alpha_f = 1.5$ achieves the best trade-off between packet priority and system efficiency. A higher value of α_c increases the system throughput, but at the expense of higher packet losses in the most important priority classes, as shown in Table 6.9. PPS(8,1) achieves the best system efficiency (system throughput of approximately 13 Mbps) but with a packet loss ratio of .0435 in the most important priority class. It is important to note that, if a flow's base layer is not decodable at the receiver due to higher packet losses, then quality enhancement layers received with zero packet losses are generally of no use as they are predicted from the base layer. In such a case, the resources utilized in scheduling the quality enhancement layer are wasted. Therefore, packets of the most important priority class must be received without any packet losses. In order to analyze the performance of the proposed scheduling rule in terms of resource utilization, a novel performance metric Resource Utilization Index (RUI) is reported below:

$$\mathrm{RUI} = \frac{1}{J_{\max} * I} \sum_{j=0}^{J_{\max}} \sum_{i=1}^{I} \mathbb{I}_{\{\mathrm{PLR}_{i,j} \leq \mathrm{PLR}_{\mathrm{thr}}\}} \tag{6.14}$$

where i and j are the user and class indexes respectively and $\mathbb{I}_{\{\mathrm{PLR}_{i,j} \leq \mathrm{PLR}_{\mathrm{thr}}\}}$ is an indicator function, equal to 1 if its argument is true. When user i's PLR for class j is less than or equal to the threshold $\mathrm{PLR}_{\mathrm{thr}}$ then the indicator function is 1. If the PLR exceeds the threshold, then the indicator function is 0. I and $J_{\max}$ are the total number of flows and classes in the system. The flows achieving a packet loss rate of less than the threshold for all the classes attain an RUI of 1. Furthermore if a flow's higher priority class has a PLR above the prescribed threshold then the indicator function

$\mathbb{1}_{\{PLR_{i,j} \leq PLR_{thr}\}}$ is zero for all the subsequent lower priority classes of such a flow. The RUI for all the considered values of α_c and α_f are:

Case: PPS(8,1)

$$RUI = 0.14 \tag{6.15}$$

Case: PPS(6,1)

$$RUI = 0.18 \tag{6.16}$$

Case: PPS(4,1)

$$RUI = 0.2 \tag{6.17}$$

Case: PPS(2,1.5), PPS(2,3), PPS(2,5), PPS(2,10)

$$RUI = 0.25 \tag{6.18}$$

Consider both system efficiency and RUI as the performance measure, then PPS(2,1.5) achieves the best trade-of between packet priority and system efficiency.

TABLE 6.9: Packet loss ratio of each traffic class for the PPS(α_c,α_f) scheduling rule. The total number of users in the cell is 36.

Priority Class Index	PPS(8,1)	PPS(6,1)	PPS(4,1)	PPS(2,1)	PPS(2,1.5)	PPS(2,3)	PPS(2,5)	PPS(2,10)
0	.0435	.0200	.0063	.0011	0	0	0	0
1	.1769	.1228	.0650	.0176	.0075	.0011	.0004	.0059
2	.2456	.2230	.2028	.2187	.2324	.2878	.3370	0.42
3	.2963	.2833	.3165	.4443	.4774	.5239	.5576	0.6030
4	.5231	.6043	.6796	.7731	.7558	.7114	.6874	0.6782
5	.5726	.6461	.7414	.8070	.8270	.8493	.8633	.8988
6	.6142	.6869	.7557	.8111	.8603	.9580	.9903	1
7	.8145	.8503	.8622	.8795	.9234	.9708	.9903	1

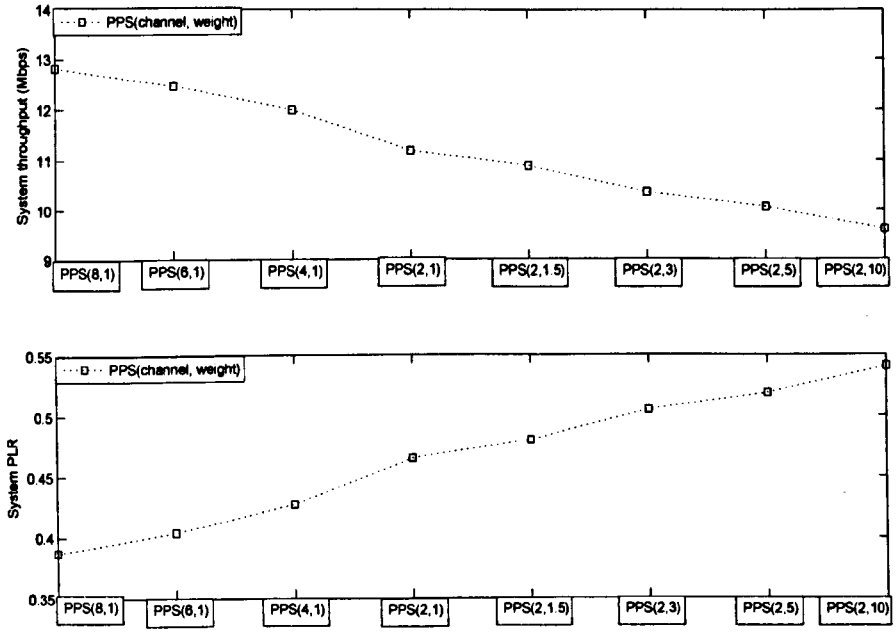


FIGURE 6.12: Efficiency performance of the PPS(α_c,α_f) scheduling rule at different values of α_c and α_f.

6.5.5 Conclusion

The performance of the scheduling function under VBR and CBR traffic is analyzed under CBR and VBR traffic. For instance if the PLR performance at the maximum input load, in Section 6.5.3, is analyzed (average input traffic's spectral efficiency requirement of 3.36 bits/sec/Hz as compare to the system capacity of 2.3 bits/sec/Hz), the resulting system packet loss ratio is approximately 0.3 with 0 and .0084 packet loss ratios in the most important priority classes. Any further increase in the average input traffic rate would incur packet losses in priority classes with index 0 and 1. By considering the M-LWDF scheduling rule as the benchmark, the increase in system capacity is 1.76 times (34 users with zero PLR in the base layer for the M-LWDF rule as compared to the 60 flows with zero PLR in the base layer for the PPS rule).

When the input traffic is CBR, there is a substantial improvement in the system robustness in terms of packet losses in the most important priority classes. For instance when the average input traffic load is 4.32 bits/sec/Hz (system capacity of 2.3 bits/sec/Hz) in Section 6.5.4, the packet loss ratios in priority classes 0 and 1 are 0 and .0075, respectively. The variability in the input traffic increases the probability of delay bound violations in the most important priority classes. Video traffic exhibits variable traffic

characteristics, therefore restricting the admission of flows is mandatory when the system's average traffic rate is above a specific threshold. Under higher system load, the PPS scheduling function incurs no packet losses in the most important priority class. From the simulation results, the maximum tolerable load in terms of the bits/sec/Hz is 3.36 (VBR traffic) and 4.32 (CBR traffic) against the system capacity of 2.3 bits/sec/Hz.

In order to increase the system capacity in terms of the number of satisfied users, the subsequent section utilizes the PPS scheduling function along with the novel concept of class based admission control policy.

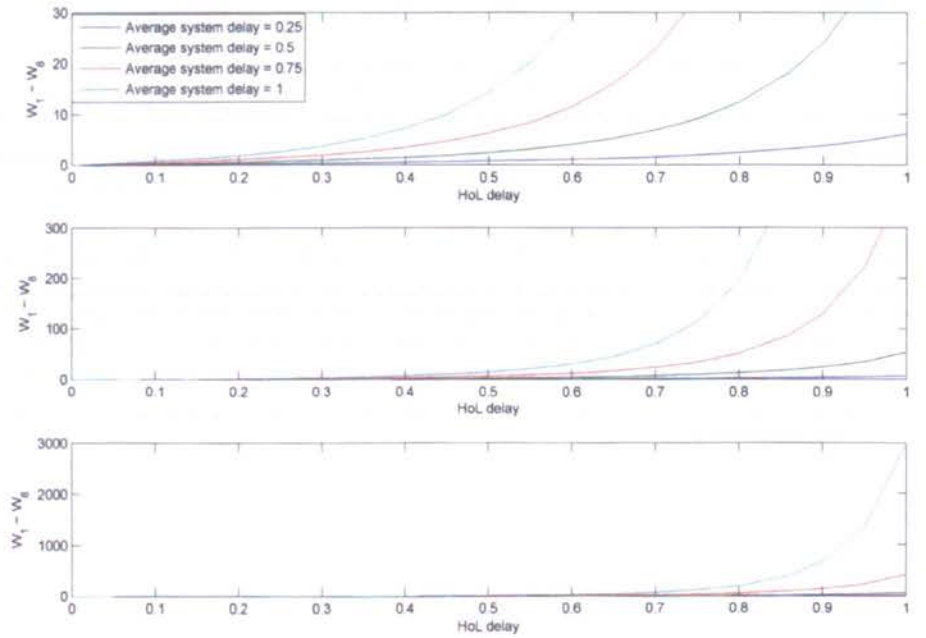


FIGURE 6.13: Priority weight difference between the least and most important priority classes.

6.6 Admission control

The main goal of the PPS scheduler is to prioritize the most important video layers. The scheduling function exploits QoE based packet marking and schedules the most important priority classes. It is important to note that at higher system delay, the lower priority packets residing in the buffer are dropped when the delay limit is reached. Lower priority packets, residing in the buffer till the delay bound, increase the average system delay, which in turn increases the resource allocation probability of the higher priority classes. Figure 6.13 shows the priority weight difference between the least and the most important priority classes. For instance consider a system with 8 priority

classes as shown in Figure 6.13, the delay based exponential priority weight decreases the resource allocation probability of the least important priority classes. According to Figure 6.13, the system becomes strictly priority aware when the normalized average system delay is one. Higher system delay makes the scheduler channel unaware as the highest priority packets are assigned resources irrespective of the channel quality which leads to a significant reduction in the system efficiency. Therefore, it is proposed to apply an admission control policy on the lower priority classes because of the following reasons.

- Restricting the admission of the lower priority packets under higher system delay will decrease the average system delay and increase the scheduler's channel awareness thus improving the system efficiency. It is important to note that a video layer is generally not decodable when its packet loss rate is above a specific threshold. Resources assigned to the half sent video layers are wasted as the video layer is not useful at the receiver. Therefore instead of scheduling a portion of a video layer, it is useful to drop the complete video layer. Thus applying admission control on the lower priority classes decreases the delay bound violation of higher priority class packets. Table 6.10 shows the decrease in the packet loss ratio of higher priority classes when lower priority classes are blocked by applying an admission control policy. According to the table, the RUI is approximately 0.25 when lower priority classes are not blocked and the scheduling function is continuously overloaded. When three of the least important priority classes are blocked the RUI increase to 0.38. Similarly when 4 priority classes are blocked, the delay bound violation of the higher priority classes further decreases which increases the RUI to 0.45. Detailed design and analysis of a class based admission control policy is discussed in Section 6.6.1.

TABLE 6.10: PLR of each traffic class for the PPS(α_c,α_f) scheduling rule with and without priority class blocking. Total number of users in the cell is 36.

Class	PPS(2,1), RUI = 0.25	PPS(2,1.5), RUI = 0.25	PPS(2,3), RUI = 0.25	PPS(2,1.5), RUI = 0.45, 4 classes blocked	PPS(2,1.5), RUI = 0.38, 3 classes blocked
1	.0011	0	0	0	0
2	.0176	.0075	.0011	.0013	.0018
3	.2187	.2324	.2878	.0069	.0143
4	.4443	.4774	.5239	.1013	0.2794
5	.7731	.7558	.7114	1	0.5652
6	.8070	.8270	.8493	1	1
7	.8111	.8603	.9580	1	1
8	.8795	.9234	.9708	1	1

- Since video traffic exhibits a highly variable traffic, the instantaneous quality variations of video flows will be reduced by imposing an admission control policy. Class based admission control will increase the resource allocation probability of the delay tolerant TCP traffic class. The scheduling function along with the admission control policy in a multi-class traffic scenario is discussed in Section 6.8.

6.6.1 Hysteresis principle for admission control

In electronics, a Schmitt trigger is a circuit with positive feedback and a loop gain greater than 1. The circuit is named a "trigger" because the output retains its value until the input changes sufficiently to trigger a change. When the input is higher than a certain chosen threshold, the output is high. When the input is below a different (lower) chosen threshold, the output is low, and when the input is between the two levels, the output retains its value. This dual threshold action is called hysteresis and implies that the Schmitt trigger possesses memory and can act as a bistable circuit (latch or flip-flop). Figure 6.14 shows the basic operation of a Schmitt trigger. The principle of Schmitt trigger in order to design a class based admission control policy is utilized.

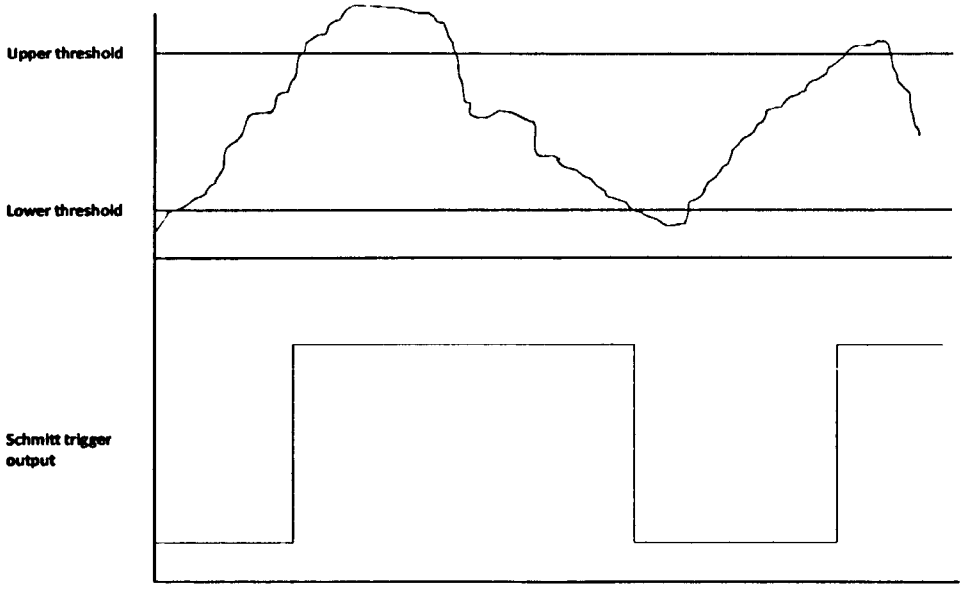


FIGURE 6.14: Basic operation of Schmitt trigger.

An admission control decision consists of blocking or admitting the packets of least important priority classes. The admission control decision of the least important priority class is taken after t_w scheduling epochs. A window based admission control policy based

on dual threshold action is introduced where a decision whether a flow's lowest priority class packets are allowed to enter the buffer is taken after t_w scheduling instants. According to the PPS scheduling metric, the priority weight decreases the resource allocation probability of the least important priority class when the normalized instantaneous system delay is high. In order to facilitate the admission control decisions, the ratio of the dropped and transmitted packets of the current lowest priority class represented by index j^* is monitored. The packet loss ratio of the lowest priority class j^* is:

$$\mathrm{plr}_{j^*}^{(t_w)} = \frac{\sum_{m=n-t_w}^{n} P_{\mathrm{drop}}^{(m)}}{\sum_{m=n-t_w}^{n} \left(P_{\mathrm{transmit}_j^*}^{(m)} + P_{\mathrm{drop}}^{(m)} \right)} \tag{6.19}$$

where

$\mathrm{plr}_{j^*}^{(t_w)}$: Packet loss ratio of class j^* over t_w scheduling epochs.

$P_{\mathrm{transmit}_j^*}^{(m)}$: Number of transmitted packets of class j^* over the moving average transmission window t_w.

$P_{\mathrm{drop}}^{(m)}$: Number of dropped packets over the moving average transmission window t_w.

The congestion status of the network is measured by utilizing the average system delay as shown below:

$$H^{(t_w)} = \frac{1}{t_w} \sum_{m=n-t_w}^{n} \overline{A^{(m)}} \tag{6.20}$$

where

$$\overline{A^{(n)}} = \frac{1}{I} \sum_{i=1}^{N} A_i^{(n)} \tag{6.21}$$

$\overline{A^{(n)}}$ is the average of the normalized HoL delay taken over I number of flows present in the system and $H^{(t_w)}$ indicates congestion in the network by calculating the average of the $\overline{A^{(n)}}$ over t_w scheduling epochs. The decision on the flows to block depends upon the packet loss ratio discussed in Algorithm 3. Flows are blocked or re-admitted according to the following rules:

$$I_{\mathrm{block}_{j^*}} = \lfloor I_{j^*} * H^{(t_w)} * S_{\mathrm{schmitt}} * \mathrm{plr}_{j^*}^{(t_w)} \rfloor \tag{6.22}$$

$$I_{\mathrm{re-admit}_{j^*}} = \lfloor I_{j^*} * (1 - H^{(t_w)}) * (1 - S_{\mathrm{schmitt}}) \rfloor \tag{6.23}$$

where $I_{\text{block}_{j^*}}$ is the number of flows blocked for class j^* and $I_{\text{re-admit}_{j^*}}$ is the number of flows re-admitted. S_{schmitt} is the output of the Schmitt trigger. The output of the Schmitt trigger, S_{schmitt}, upon reaching the higher and lower threshold limits is given as:

$$S_{\text{Schmitt}} = \begin{cases} P_{H^{(t_w)}}, & \text{if } H^{(t_w)} \geq S_{\text{thr}_h} \\ \rho, & \text{if } H^{(t_w)} \leq S_{\text{thr}_l} \end{cases} \tag{6.24}$$

where

$$\rho = S_{\text{schmitt}} * (1 - \frac{1}{\omega}) + \frac{1}{\omega} * P_{H^{(t_w)}} \tag{6.25}$$

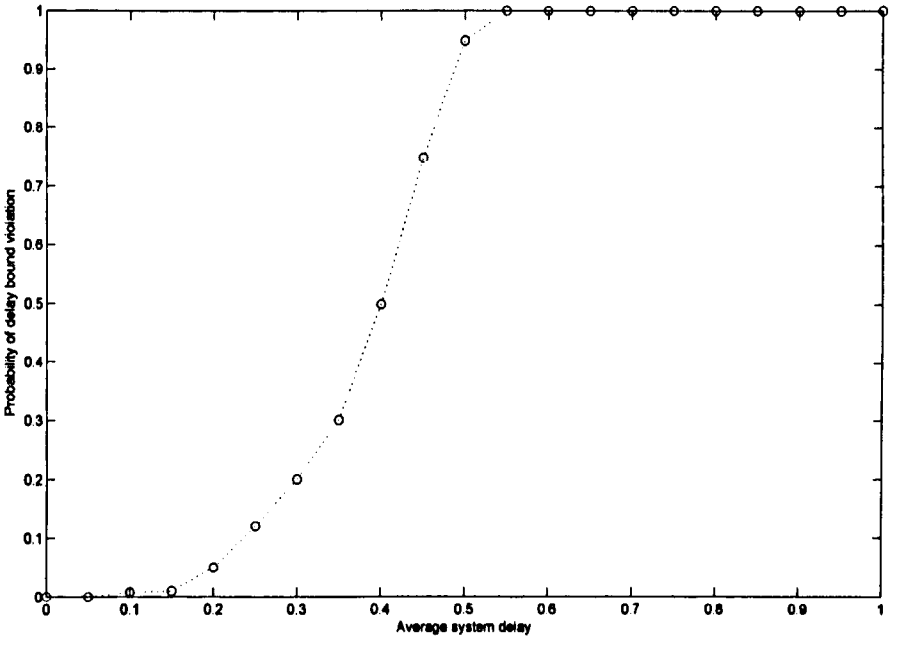


FIGURE 6.15: Probability of delay bound violation at different system delay.

S_{thr_l} and S_{thr_h} are the lower and higher threshold limits of the Schmitt trigger. $P_{H^{(t_w)}}$ is the probability of delay bound violation based on the congestion status of the network. ω is the exponential moving average weight. The probability of delay bound violation, $P_{H^{(t_w)}}$, can be derived mathematically from the scheduling function. Figure 6.15 shows the probability of delay bound violations based on multiple simulations. The basic operation of the Schmitt trigger based admission control is shown in Figure 6.16. According to the figure, the output of the Schmitt trigger latches to $P_{H^{(t_w)}}$ when the congestion parameter $H^{(t_w)}$ crosses the higher threshold limit (the higher limit of the threshold is

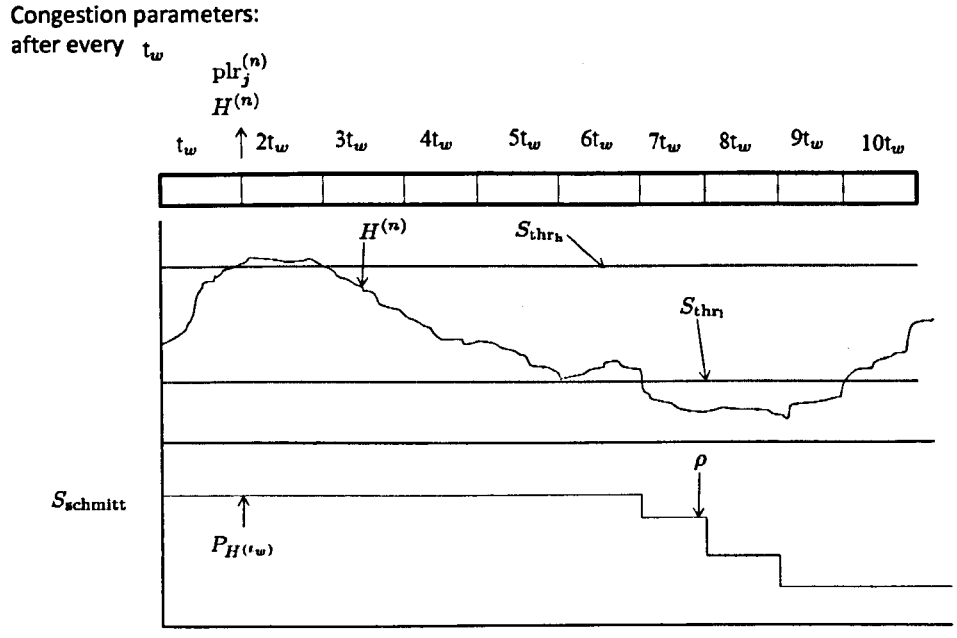


FIGURE 6.16: Basic operation of Schmitt trigger based admission control.

set to the point where the delay bound violation probability is close to 1). When the congestion parameter $H^{(t_w)}$ reaches the lower threshold limit, the output is decreased according to the exponential averaging equation given in (6.25). The output decreases until the congestion parameter $H^{(t_w)}$ is below the lower threshold limit S_{thr_l}. The output is latched if the congestion parameter $H^{(t_w)}$ crossover the lower threshold limit. Further details on the admission control policy is given in the following subsection.

6.6.1.1 Admission control pseudo-code

The admission control pseudo-code is shown in Algorithm 3. According to the algorithm, the number of transmitted packets of the current lowest priority class j^* and the number of packets dropped due to delay bound violations are calculated at each scheduling epoch. After t_w scheduling epochs, the packet loss rate (considering the number of transmitted packets of the least important class j^* and total number of dropped packets due to the violation of the delay bound limit) and congestion parameter $H^{(t_w)}$ are computed according to (6.19) and (6.20) respectively. For instance, if the scheduling function schedules 30 ($P^{(m)}_{\text{transmit}^*_j} = 30$) packets of different flows' lowest priority class and the total number of packets dropped due to delay bound violations is 20 ($P^{(m)}_{\text{drop}} = 20$) over the scheduling window of t_w epochs then the admission controller, after t_w scheduling epochs, calculates the packet loss ratio $\text{plr}^{(t_w)}_{j^*}$ using (6.19) as given below:

$$\text{plr}_{j^*}^{(t_w)} = \frac{20}{30+20} = 0.4 \tag{6.26}$$

Now let us assume that the average system delay, $H^{(t_w)}$, over the scheduling window of t_w epochs is 0.6 and the higher threshold limit S_{thr_h} is set to 0.5. When $H^{(t_w)}$ crosses the threshold limit of 0.5 then S_{schmitt} is latched to $P_{H^{(t_w)}}$ as shown in Figure 6.16. Average system delay of 0.5 characterizes a congested system where the probability of delay bound violation is approximately 1 as shown in Figure 6.15. Therefore, $P_{H^{(t_w)}}$ at $H^{(t_w)} = 0.5$ is set to 1. Considering 20 flows in the system having packets of the least important class then according to equation (6.22), ($\lfloor 20 * 0.6 * 1 * 0.4 \rfloor$) 4 flows' priority class j^* will be blocked.

According to the algorithm, if $I_{\text{block}_j^*}$ is greater than zero then $\delta(j^*, i)$ (the admission control row vector containing the ids of the blocked flows for priority class j^*) is updated. The decision on which of the flows are blocked is based on the ratio of channel quality and time-averaged throughput. $\delta(j^*, i)$ is updated with the ids of the blocked flows. The total number of blocked flows for class j^* in the considered example is 4, therefore $\delta(j^*, i)$ is updated with the ids of the 4 flows having the least ratio of channel quality and time-averaged throughput ($|\delta(j^*, i)| = 4$ is the cardinality of $\delta(j^*, i)$). In the considered example, assuming $j^* = 5$ informs the buffer manager that the admission controller has blocked all the flows having packets of classes 6 and 7 (system with 8 priority classes having PCI from 0 to 7) and the current lowest priority class in the system has a class index of 5. The row vector, $\delta(j^*, i)$, at the entrance of the buffer with $j^* = 5$ implies that all flows' packets marked with indexes 6 and 7 are blocked from entering the queues at eNodeB. Furthermore, flows having packets marked with $j^* = 5$ and having ids in the row vector are blocked from entering the queues.

After blocking 4 flows, the system delay is monitored over the next window of scheduling epochs. If the average system delay is reduced and there are no packet losses then number of flows will be re-admitted. For instance, if the average system delay (after blocking packets of 4 flows' current least important priority class) reduces to 0.4 then according to the algorithm, equation (6.23) will determine the number of flows to re-admit as given below:

$$I_{\text{re-admit}_j^*} = \lfloor 20 * (1 - 0.4) * (1 - 1) \rfloor = 0 \tag{6.27}$$

According to the pseudo-code, none of the flows will be re-admitted as $I_{\text{re-admit}_j^*} = 0$. Hence $\delta(j^*, i)$ will not be updated (none of flows' id will be removed from the row vector) as shown in the pseudo-code. S_{schmitt} is latched to 1 and will only change its

value when $H^{(t_w)}$ will cross the lower threshold. This is an indication that the scheduler is in the optimal performance range (input traffic is in the achievable rate region as there are no delay bound violations). The lower threshold limit is set to a point where the packet delay bound violation probability is zero. If we consider that the lower threshold limit is set to 0.2 and the average system delay $H^{(t_w)}$ is below 0.2 (over the moving average window) then upon reaching the lower threshold, S_{schmitt} is equal to ρ as given in equation (6.25). If $H^{(t_w)}$ is equal to or below the lower threshold limit for a higher number of window cycles, then S_{schmitt} will decrease exponentially which will increase the number of re-admitted flows as shown in Figure 6.16. The selection on which of the flows to re-admit is also based on the proportional fair rule. Flows with the highest ratio of channel quality and time-averaged throughput are re-admitted by removing flows ids from the row vector $\delta(j^*, i)$. If all the flows of the least important class are re-admitted then $|\delta(j^*, i)| = 0$. In the next admission control cycle, priority class with index $(j^* = j^* + 1)$ will be re-admitted as shown in the pseudo-code of the algorithm.

Algorithm 3 Schmitt trigger based admission control

```
Set Simulation_time
repeat
  for n to n = n + t_w do
    Calculate P^(n)_transmit*_j and P^(m)_drop at each scheduling epoch
  end for
  Calculate plr^(t_w)_j*, for class j*, according to (6.19)
  Calculate H^(t_w) according to (6.20)
  if plr^(t_w)_j* > 0 then
    For class j* calculate the number of flows to block, I_block*_j, according to (6.22)
    if I_block*_j > 0 then
      Update δ(j*, i)
      if |δ(j*, i)| ≥ N then
        j* = j* − 1
      end if
    end if
  else
    For class j* calculate the number of flows to re-admit, I_re−admit*_j, according to
    (6.23)
    if I_re−admit*_j > 0 then
      Update δ(j*, i)
      if |δ(j*, i)| ≤ 0 then
        j* = j* + 1
      end if
    end if
  end if
until END OF Simulation_time
```

In the subsequent sections, the performance of the joint admission control and scheduling function under QoE (sum MOS maximization) based packet marking scheme is analyzed.

6.7 Performance evaluation of the proposed joint admission control and scheduling algorithm with packet marking

In this section, the performance of the admission control strategy by utilizing the exponential based PPS scheduling function is analyzed. Furthermore, QoE based marking of *SVC layers into priority classes* is considered in the performance evaluation. The main goal of packet marking is to achieve the maximum overall QoE under the constraint of the available network resources. Packet marking facilitates rate adaptation under the event of congestion, by assigning less important priority classes to the layers achieving lower video quality at higher bitrates. Thus, the packets of video layers contributing less to MOS at the expense of higher bitrates are marked with higher priority class index. The higher priority class index indicates the lower priority of the packets which is exploited by the joint admission controller and scheduling function by dropping such packets under the event of network congestion. The simulation scenario and the results achi

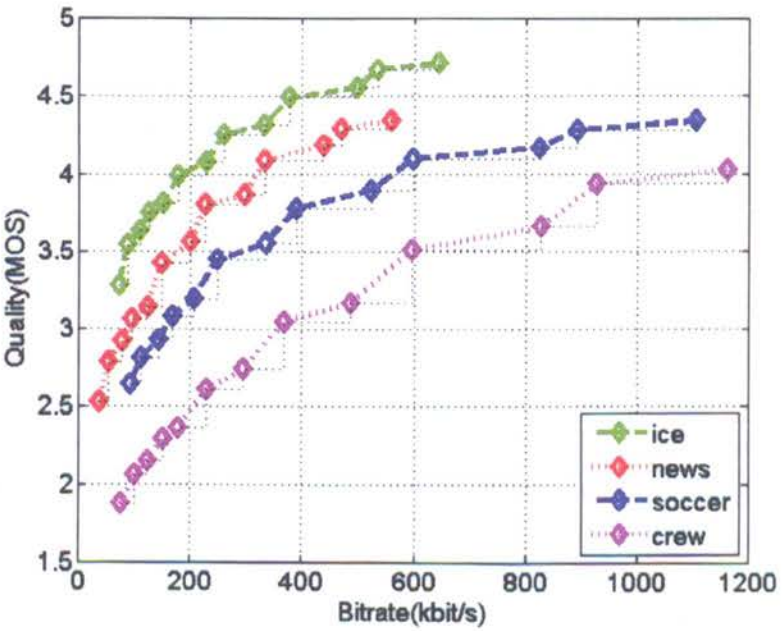


FIGURE 6.17: Quality (MOS) vs. Bitrate (Kbps) characteristics for each of the considered videos sequences (by Bo Fu, DOCOMO.).

TABLE 6.11: Priority marking description (elaborated by Bo Fu, DOCOMO.).

Priority classes	mean MOS achievements	Priority Class Index (PCI), $j_i^{(n)}$	$P_{i,j}^{(n)}$
Class 1	[1.0, 2.2]	0	12
Class 2	[2.2, 2.4]	1	11
Class 3	[2.4, 2.6]	2	10
Class 4	[2.6, 2.8]	3	9
Class 5	[2.8, 3.0]	4	8
Class 6	[3.0, 3.2]	5	7
Class 7	[3.2, 3.4]	6	6
Class 8	[3.4, 3.6]	7	5
Class 9	[3.6, 3.8]	8	4
Class 10	[3.8, 4.0]	9	3
Class 11	[4.0, 4.2]	10	2
Class 12	[4.2, 5]	11	1

6.7.1 Simulation scenario

In order to assess the performance of the proposed admission control and scheduling strategy, scalable video transmission over an LTE/LTE-A system is simulated. Packet marking is performed at the core network. The video packets are categorized in 12 priority classes, each associated to a mean MOS target range, as shown in Table 6.11. The higher the mean MOS, the higher the assigned priority class index. The table shows the priority class index assigned to the packets of each priority class. Furthermore, the priority $P_{i,j}^{(n)}$ of each packet associated with the marked priority class index at the MAC layer is also shown in the table. Different video streams with different rate and quality characteristics are selected as shown in Figure 6.17. The video sequences are encoded with the SVC codec and comprise a base layer and 12 quality layers. The increase in the MOS score along with the addition of each quality layer is shown in Figure 6.17. The wireless simulation parameters are the same as in Table 6.3, except the selection of the bandwidth. A 3 MHz (15 PRBs) bandwidth system is selected. The selection of the lower bandwidth system stems from the fact that it allows less video flows, enabling us to analyze the PLR and MOS performance of each of the video flows in the system. According to the simulation results reported in Section 6.5.4, the parameters $\alpha_c = 2$ and $\alpha_f = 1.5$ achieve the best trade-off between packet priority and system efficiency. Therefore, same parameters are selected for the scheduling function. For the class based admission controller, the two limits of the hysteresis are $S_{\text{thr}_l} = 0.2$ and $S_{\text{thr}_h} = 0.5$. These limits are set according to Figure 6.15 which shows the probability of delay bound violation at different system delay values. According to the figure, the probability of delay bound violation is approximately 1 at the normalized averaged system delay of

0.5. When the system delay is below 0.2, the probability of delay bound violation is very low. Therefore, the hysteresis output S_{Schmitt} latches to 1 ($P_{H^{(t_w)}} = 1$), when $H^{(t_w)} = 1$. The hysteresis output decreases exponentially when $H^{(t_w)}$ crosses the lower threshold of 0.2.

The maximum load scenario of the PPS scheduling function in Section 6.5.3 (under VBR traffic) incurred a system PLR of approximately 30 %. It is important to note that with the joint scheduling function and admission control policy, the system is more robust in protecting the important video layers from any delay bound violations. Therefore, 5 different high load scenarios are considered where the incurred system PLR starts from 34.75 % (Scenario 1). The input average traffic (by adding more video flows) is increased to the point where the delay bound violations starts to occur in the base layer of the considered video sequences. In the following, five different high load scenarios are reported:

- 8 video users: In this scenario 2 Ice, 2 News, 2 Soccer and 2 Crew video flows are simulated. This scenario corresponds to an average input traffic rate of 7 Mbps (2.33 bits/sec/Hz). The channel quality (in terms of SINR) of each of the video flows is set such that the system capacity is 5.2 Mbps (1.73 bits/sec/Hz).

- 12 video users: In this scenario the load is further increased by adding more video flows. Specifically, 2 Ice, 6 News, 2 Soccer and 2 Crew video flows are simulated. This scenario corresponds to an average input traffic rate of 9.2 Mbps (3.066 bits/sec/Hz) whereas, the system capacity is 1.73 bits/sec/Hz.

- 10 video users: In this scenario the average input traffic rate is approximately similar to the previous scenario. However, the video mix is different from the previous scenario. Specifically, 2 Ice, 2 News, 2 Soccer and 4 Crew video flows are simulated. This scenario corresponds to an average input traffic rate of 9.2 Mbps (3.066 bits/sec/Hz), whereas the system capacity is similar to the previous scenario.

- 16 video users: The input traffic rate is further increased to 14 Mbps (4.66 bits/sec/Hz) by simulating 4 video flows from each of the considered video sequences. The average system capacity is the same as in the previous scenarios.

- 20 video users: In this scenario the average input traffic rate is increased to 17.5 Mbps (5.833 bits/sec/Hz) as compared to the average system capacity of 5.2 Mbps (1.73 bits/sec/Hz). 5 video flows are selected from each of the video sequences. The average system capacity is the same as in the previous scenarios.

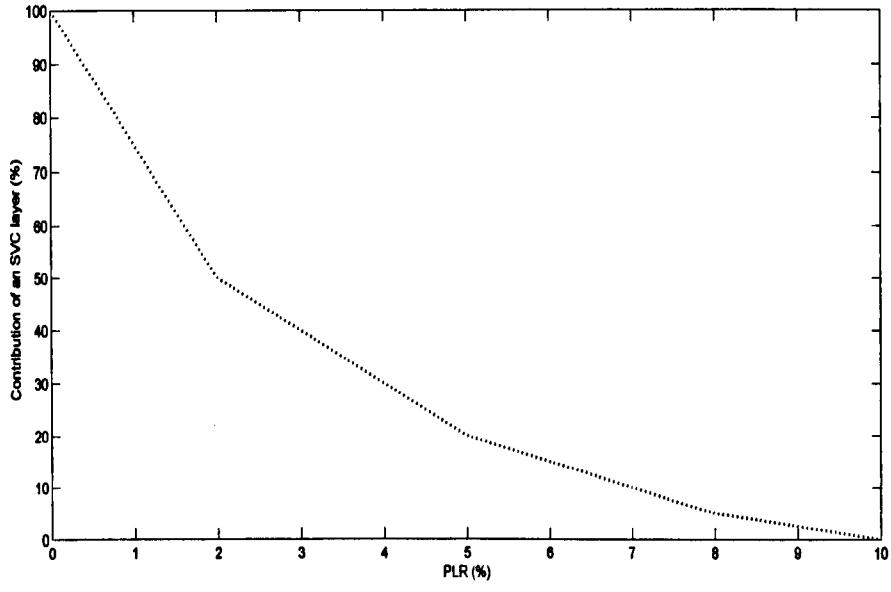


FIGURE 6.18: Contribution towards MOS of one SVC layer at different PLRs.

In each of the considered scenarios, the PLR performance of each video flow (considering the dropped packets due to delay bound violations and blocked packets due to the class based admission control policy) in each of the video layers (base and quality enhancement layers) is analyzed. The resulting MOS (due to the dropped and blocked packets in SVC layer) is also reported for each of the considered scenarios. Furthermore, the admission controller's blocking of priority classes w.r.t time is also reported for each of the considered scenarios. The contribution of an SVC at different PLR is reported in Figure 6.18. According to the figure, the contribution of an SVC layer is 100 % when there are no packet losses. The contribution of an SVC layer decreases to half when the incurred PLR in the layer is 2 %. For instance if the layer contributes 0.5 MOS to the overall video quality, the PLR of 2 % decreases the MOS contribution to 0.25. When the incurred PLR is more than 10 %, the layer contribution towards the overall video quality is zero. It is important to note that the impact of PLR on an SVC layer depends upon the complexity of the video sequence and Figure 6.18 is only an approximation of the contribution of an SVC layer against the incurred PLR.

The mapping of SVC layers into priority classes is shown in Figure 6.19. The total number of priority classes is 12. According to the figure, 13 video layers (1 base and 12 quality layers) are mapped into 12 priority classes. For instance, consider the Ice video sequence (in the left column), the base layer (SVC layer 1) is assigned to priority class 1 and the first quality layer (SVC layer 2) is assigned to priority class 2. SVC layers 3 and 4 are assigned to priority class 5 and the SVC layers 5, 6 and 7 are assigned to

priority class 8 followed by the SVC layer 8 mapping to priority class 9. Layers 9 and 10 are assigned to priority class 10. The last three SVC layers (10, 11 and 12) are assigned to priority class 12.

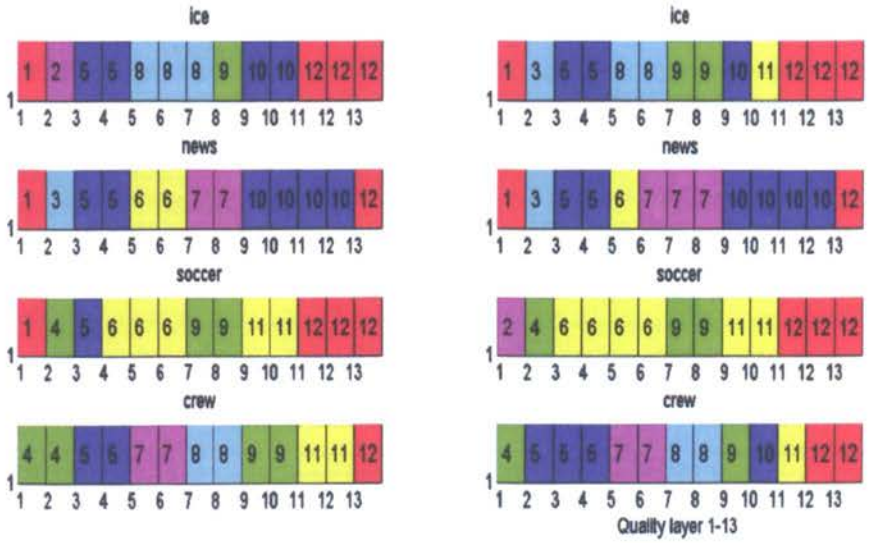


FIGURE 6.19: SVC layers mapping to priority classes for scenario 1 (mapping scheme elaborated by Bo Fu, DOCOMO).

6.7.2 Results and discussion

The performance of all the video flows (Scenario 1) in terms of PLR (considering the dropped packets due to delay bound violations and blocked packets due to the class based admission control policy) is shown in Figure 6.20. The x-axis shows the video layer id starting with the base layer (id 0) and 12 quality layers (id 1 to 12). According to the figure, both the Ice video flows incur very high PLR (more than 15 %) in quality layers 10, 11 and 12. All the packets of the base and 7 quality layers of both the News video flows are scheduled before the delay bound limit. A very low PLR (less than 5 %) occurs for both the News flows in quality layers with id 8, 9, 10 and 11. The last quality layer has a very high PLR (more than 15 %) and is considered as undecodable at the receiver. The blocking of video layers w.r.t time is shown in Figure 6.21. The figure shows the percentage of video flows blocked for each of the priority classes. According to the figure, throughout the simulation period only priority classes 11 and 12 (for all the video flows) are blocked in the time intervals shown on the x-axis; the remaining priority classes are unblocked throughout the simulation period. When we analyze the

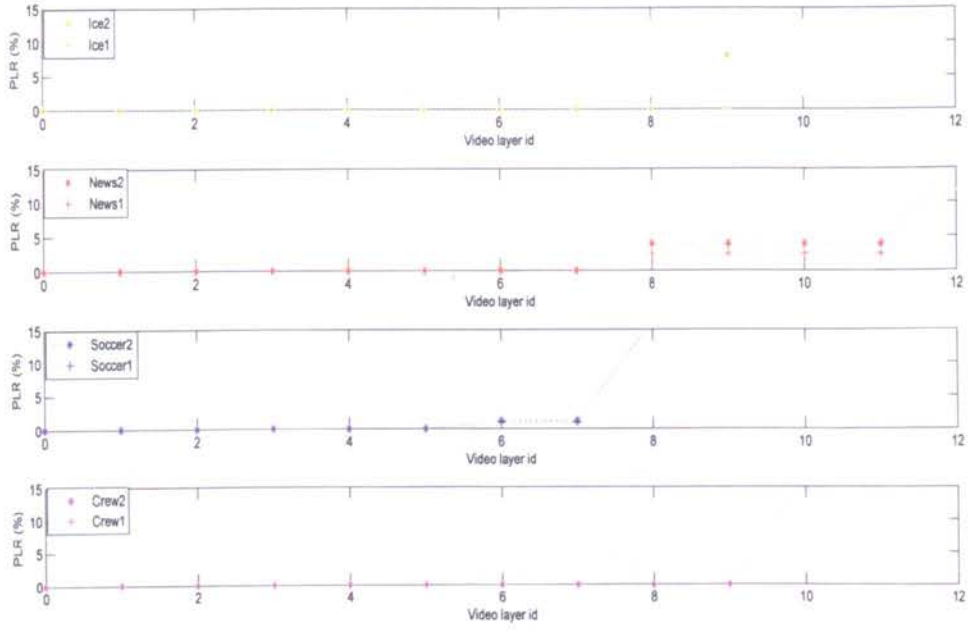


FIGURE 6.20: Scenario 1: PLR (%) in each of the SVC layers for each of the considered video flows. Total system PLR is 34.75 %.

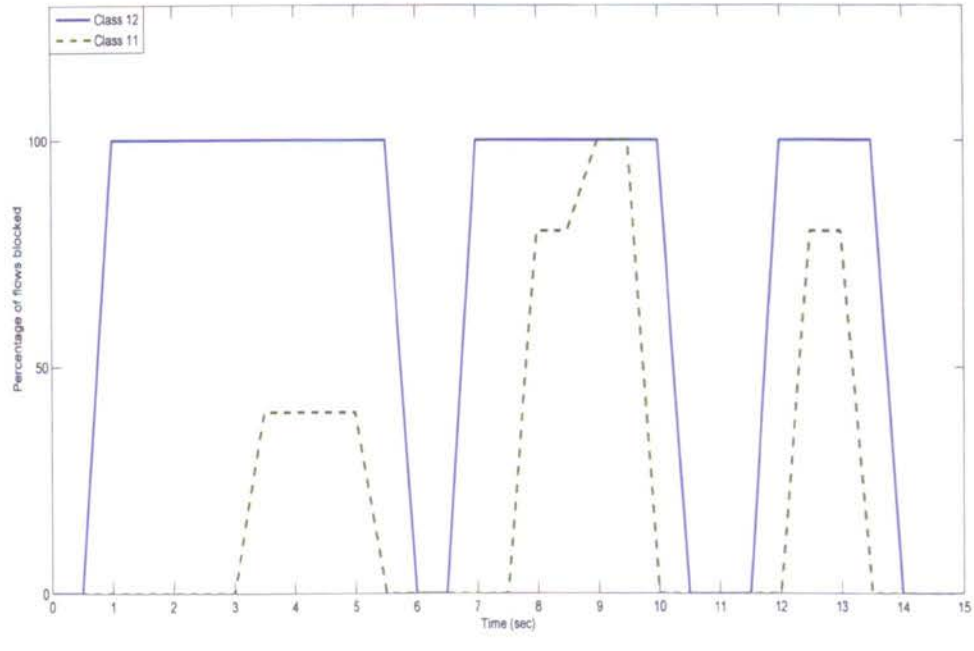


FIGURE 6.21: Scenario 1: Admission controller's blocking of priority classes throughout the simulation period.

SVC layers to priority class mapping in Figure 6.19, the Ice video flows' last 3 quality layers and the News video flows' last quality layer are assigned to priority class 12. Under congestion, the normalized average system delay increases, which decreases the resource allocation probability of the least important priority class. This leads to an increase in the PLR of the least important priority class. The average normalized system delay is calculated over the moving average window. This system delay along with the moving average packet loss ratio of the current least important class is utilized in the priority class (current least important class in the system) blocking of the video flows. According to Figure 6.21, 100 % of the video flows' priority class 12 are blocked, *i.e.*, the normalized averaged system delay increases above the upper threshold of the hysteresis with very high system PLR in priority class 12, which triggers the blocking of the flows. In order to move the arrival rate within the achievable rate region (according to the delay budget constraints), all the flows of priority class 12 are blocked from 1 sec onwards.

After blocking priority class 12, the normalized average system delay does not cross the lower threshold limit (till the 6 sec point in the figure) which prevents the re-admission of the packets of priority class 12. After a 3 second period, there is an increase in the average system delay which increases the delay bound violations of the current least important priority class packets (class 11, since 12 is already blocked). The blocking of the flows of priority class 11 occurs from 3 sec onwards as shown in Figure 6.21. It is important to note that only 40 % of the flows of priority class 11 are blocked. Priority class 11 has packets from 1 Ice, 2 Soccer and 2 Crew video flows a shown in Figure 6.19. Therefore, 40 % of the total flows blocking means that 2 of 5 flows having packets of priority class 11 are blocked. In order to find out which flows to block for priority class 11, the admission control strategy uses the ratio of channel quality and the time-averaged throughput. The flows with lower channel quality and higher time-averaged throughput are blocked. When the normalized system delay crosses the lower threshold of the hysteresis window, the blocked flows' priority classes are re-admitted as shown in the Figure 6.21. A moving average window size equal to 100 scheduling epochs is utilized, after which the re-admission is allowed based on the network system delay as discussed in the admission control design. If the network system delay is below the lower threshold and there are no packet losses, the exponential averaging window equation (6.25) is used to determine the number of flows to readmit. If the average system delay is below the lower threshold for a higher number of window cycles, the re-admission of flows increases exponentially.

According to Figure 6.21, the delay bound violations in the packets of priority classes 11 and 12 incur a higher PLR for the flows having video layers marked with classes 11 and 12. Both the Soccer video flows have 5 video quality layers in priority classes 11 and 12 as shown in Figure 6.19. Therefore, packets of the last 5 quality layers are

dropped with a very high rate (more than 15 %) as shown in Figure 6.20. Similarly, for Crew flows the last 3 video quality layers are marked with priority classes 11 and 12. Therefore, the last 3 quality layers of the Crew flows have a very high PLR. One interesting observation in the results shown in Figure 6.20 is the PLR of above 5 % in the Ice 2 video flow's quality layer with id 9. According to Figure 6.19, one of the Ice video flows has a layer marked with priority class 11. The admission control policy blocked the packets of the priority class based on the proportional fair rule (ratio of channel quality and time-averaged throughput). Since the averaged channel quality of the Ice video flow is good (Figure 6.22) and also the bitrate of this video sequence is the lowest as compared to the other video sequences, therefore this video flow's packets of priority class 11 is blocked for the shortest amount of time. According to Figure 6.21, only the time interval from 9 to 10 seconds, 100 % of the priority class 11 flows are blocked. The remaining period of the time, packets of Ice 2 video flows are scheduled within the delay budget. The MOS value and the associated channel quality of each of the video flows is shown in Figure 6.22. The lower MOS value of the ICE 2 flow is mainly due to the fact that this video flow has a higher PLR (approximately 8 %) in video layer 9 as shown in Figure 6.20. For the MOS calculation, a maximum PLR threshold of 10 % is considered. Above this threshold, the quality layer contribution to the overall MOS is zero. The impact on the MOS for PLR of less than 2 % is low. When the PLR is increased above 2 %, the contribution of the video layer to the overall video quality decrease sharply. News video flow 2 has a higher PLR (approximately 4 %) in the video layers 8, 9, 10 and 11. The main reason for this PLR is the lower channel quality of this video flow. Another main reason of this video's high PLR is that the priority class 10 for this video sequence consists of 4 video quality layers as shown in Figure 6.19, which corresponds to a bitrate of approximately 400 Kbps (when we analyze layers 9 to 12 of the News video sequence in Figure 6.17). Therefore, this video sequence has a higher number of packets in priority class 10. During higher congestion periods, when packet losses occur in priority class 11 as shown in Figure 6.21, the probability of delay bound violations in the neighbouring priority classes also increases. The admission controller reacts to this event by blocking the flows of the least important priority class. However, if the peak rate of a video is much higher, coupled with lower instantaneous channel quality, packet losses occur in the lesser important (neighbouring) priority classes. Thus, the main reasons for a higher PLR in the classes not blocked by admission control policy are:

- Channel quality of the video flow.
- Traffic bitrate assigned to a particular priority class. For instance, the News 1 flow has a better channel quality than the Crew 2 flow and News 2 flow has the same

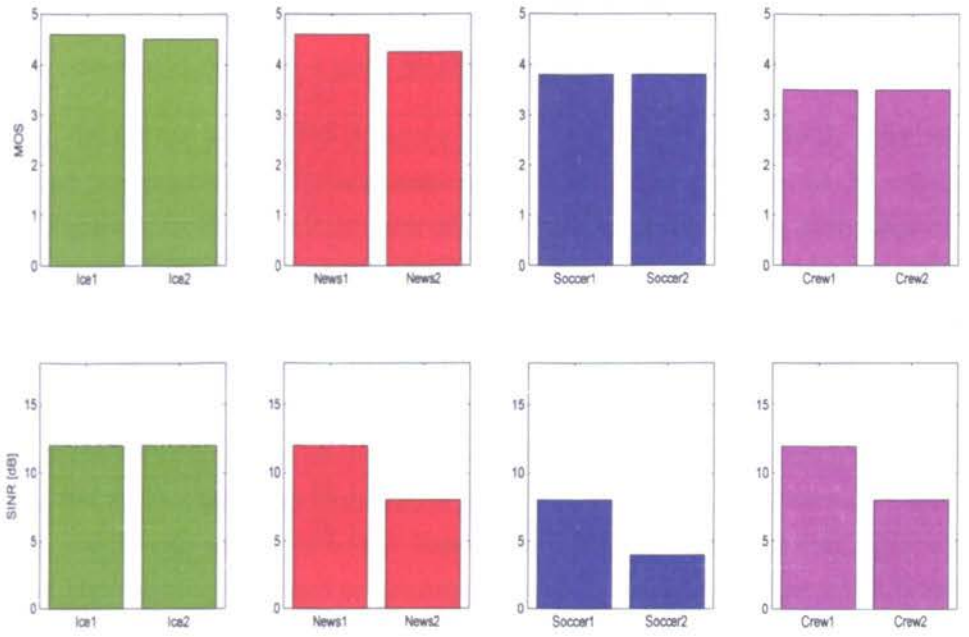


FIGURE 6.22: Video (MOS) and Channel quality (SINR) for each of the video flows in scenario 1. Average MOS per video flow is 4.07.

channel quality as the Crew 2 flow. However, higher PLR occur in the priority class 10 for the News 1 and News 2 video flow mainly because News video flows have four time more traffic bit rate in the priority class 10 as compared to the Crew video flows.

TABLE 6.12: Scenario 1: Packet loss ratio performance of all the video flows for the M-LWDF rule.

Video flows	PLR(%)	Base layer PLR(%)
News 1	5	1.29
News 2	45	14
Ice 1	0	0
Ice 2	0	0
Soccer 1	37	29
Soccer 2	49	45
Crew 1	21	14
Crew 2	38	27

For scenario 1, the performance of the M-LWDF scheduling rule is shown in Table 6.12 where the PLR performance of all the video flows is reported. The table also shows the percentage of delay bound violations in the base layer of each video flow. In this

TABLE 6.13: Scenario 1: PLR (%) in each video layer under the M-LWDF rule.

Video layer id	PLR(%)
0	18.7
1	2.95
2	3.14
3	1.6
4	2.1
5	3.64
6	11
7	11.5
8	23.5
9	20
10	41
11	36
12	37

scenario, the average input arrival rate is 7 Mbps as compared to the system capacity of 5.2 Mbps. There is no rate adaptation policy for the M-LWDF scheduling rule, therefore this scheduling rule requires a proper flow admission control policy which should not increase the arrival rate above the system capacity. The increase in arrival rate above the system capacity incur delay bound violations in the higher rate low channel quality flows (since this rule is derived from the proportional fair strategy). Therefore once the delay sensitive traffic's arrival rate reaches the system capacity, the admission control policy should block further flows from entering the system. Increase in the arrival rate above the system capacity would result in QoS violations. According to Table 6.12, as many as 5 video flows have PLR of more than 10 % in the base layer (note that base layer's PLR is lower than the overall PLR mainly due to the fact that base layer's packets (in a GoP) are streamed first followed by the enhancement layers' packets.). The PLR threshold of the base layer is generally very low, delay bound violation rate exceeding 5 % results in a poor MOS quality. When the system is left to run under high load, this scheduling rule only serves low bitrate good channel quality flows (only 2 Ice and 1 News video flows are served with good quality.). Therefore, all the delay aware scheduling rules must ensure that the arrival rate should not exceed the system capacity, otherwise the QoS performance of the existing users in the network would be violated resulting in an increase in the number of unsatisfied users. The PLR in each of the video layers is given in Table 6.13 which shows that the overall base layer's PLR is as high as 18.7 %. The PPS scheduling rule coupled with the class based rate adaptation policy serves all the video flows with good MOS quality, whereas the delay based scheduling rule can only serve three video flows. In order to increase the number of satisfied users, the delay based scheduling must block the high rate and low channel video quality flows.

For instance, the admission control strategy must block both the Crew or Soccer video flows so that the arrival rate is with in the achievable rate region. Any further increase in the arrival rate would result in a poor MOS performance of all the video flows.

The mapping details for scenario 2 is shown in Figure 6.23. In scenario 2, a further increase in load triggers the admission control strategy for priority classes 10, 11 and 12 as shown in Figure 6.24. The PLR in each of the SVC layers for all the video flows is shown in Figure 6.25. Since higher packet losses occur in classes 10, 11 and 12, therefore Ice video flows' layers 0 to 7 are scheduled with zero PLR and last 5 quality layers marked with priority classes 11 and 12 (Figure 6.23) are dropped. Similar analysis holds for the 6 News video sequences in which the last 5 quality layers have a higher PLR. Now consider the MOS performance of all the 6 News video sequences, the channel quality has not much impact on the MOS performance of the News video flows as shown in Figure 6.26. This is mainly due to the fact that four News flows have a higher traffic bitrate (4 layers with high bitrate) assigned to priority class 10 as shown in Figure 6.23. When the packets of priority class 10 are assigned resources, this increases the average system delay which causes higher packet losses as this priority class for News video flow requires more resources. One News sequence has 4 quality layers marked with priority class 11, therefore layers assigned to this class are also dropped. Only one News sequence has one quality layer (layer with id 8) assigned to the priority class 10 which corresponds to a low traffic intensity. The PLR performance for this News sequence (News 3) is shown in Figure 6.25 where the PLR in the quality layer 8 is approximately 9 %. The MOS performance of all the video flows for this scenario is shown in Figure 6.26.

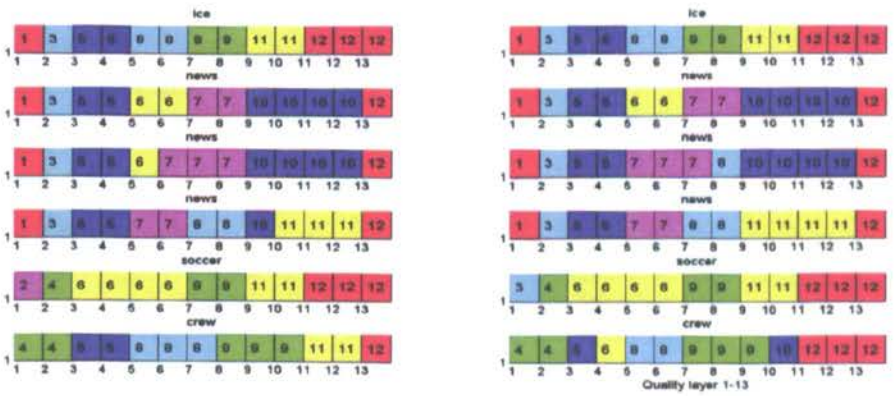


FIGURE 6.23: SVC layers mapping to priority classes for scenario 2 (mapping scheme elaborated by Bo Fu, DOCOMO).

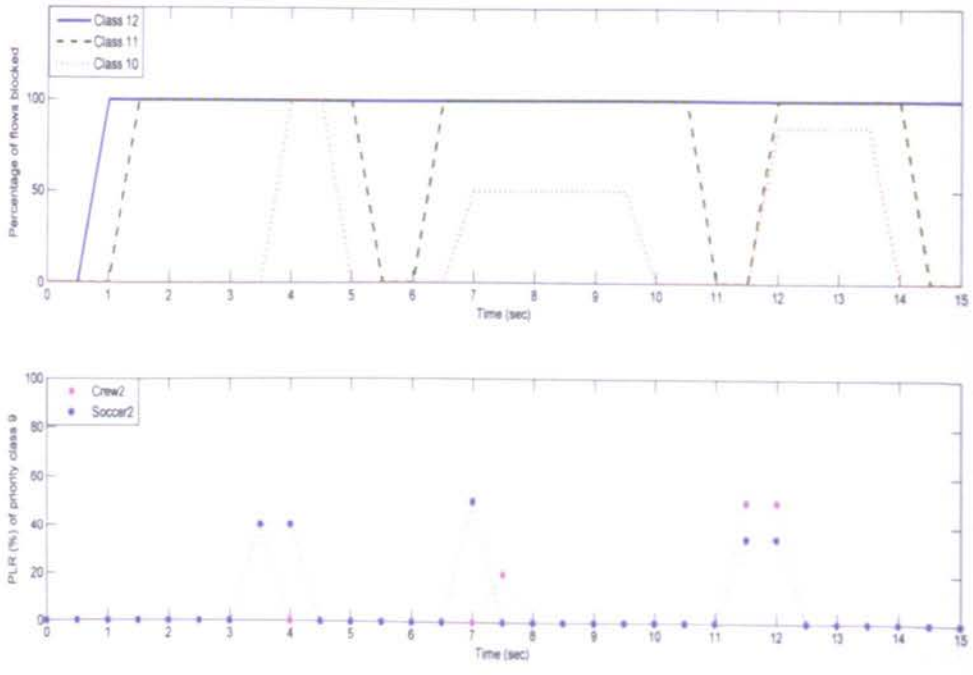


FIGURE 6.24: Scenario 2: Admission controller's blocking of priority classes throughout the simulation period.

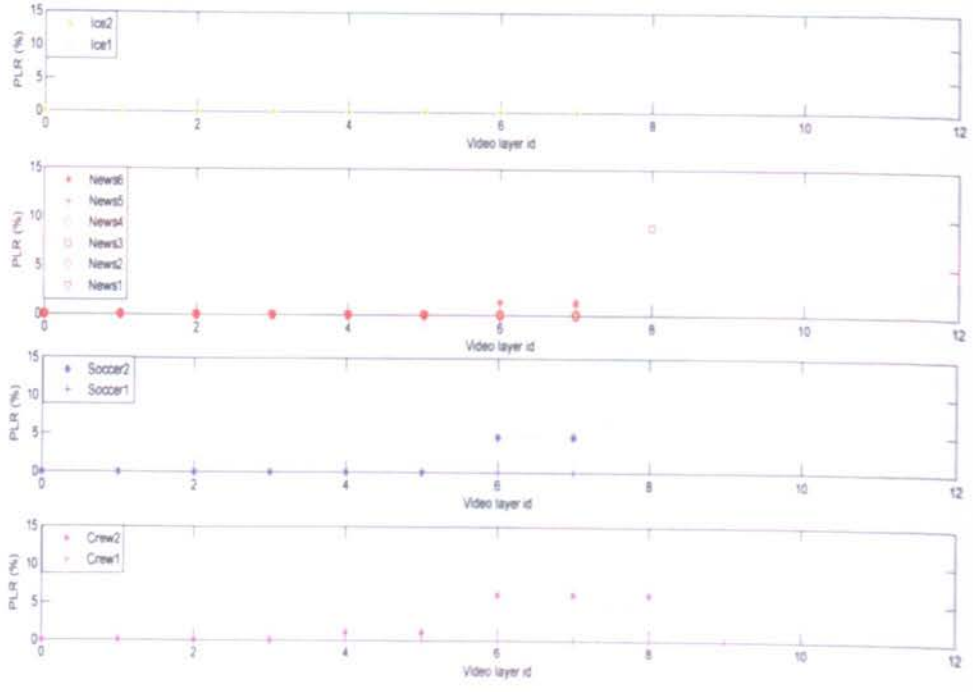


FIGURE 6.25: Scenario 2: PLR (%) in each of the SVC layers for each of the considered video flows. Total system PLR is 47.24 %.

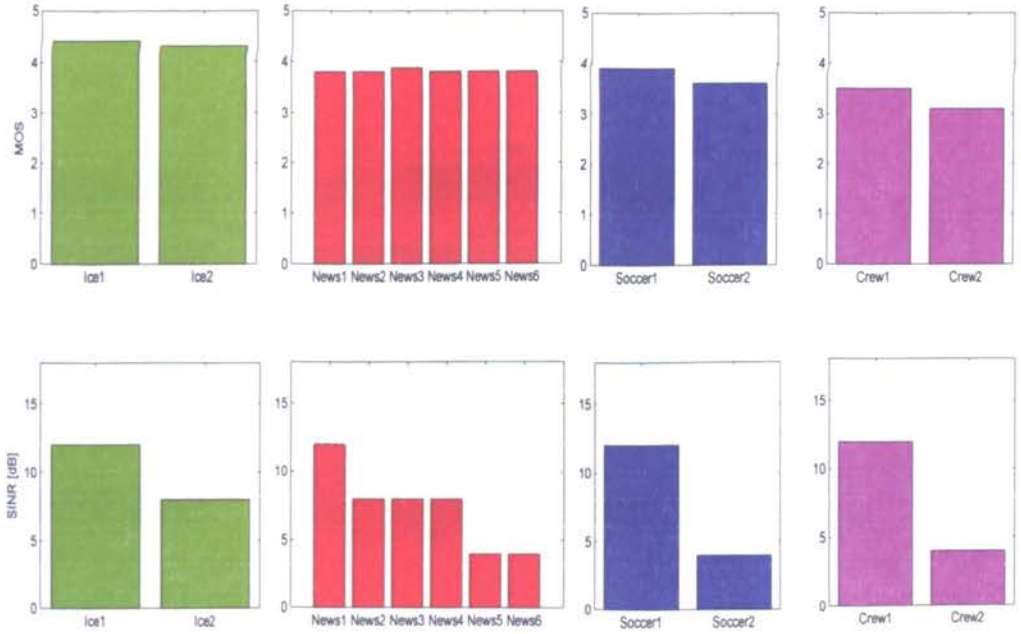


FIGURE 6.26: Video (MOS) and Channel quality (SINR) for each of the video flows in scenario 2. Average MOS per video flow is 3.8.

For the Crew video flows, 4 quality layers are dropped for Crew 2 mainly due to the fact that higher packet losses occur for priority classes 10, 11 and 12, and one of the crew flow's last 4 quality layers are mapped to priority classes 10 and 12 (other crew flow has only 3 quality layers mapped to classes 11 and 12 and no layers to class 10) as shown in Figure 6.23. Crew 2 has also higher PLR in layers 6, 7 and 8, this is mainly due to the poor channel quality of this video flow. When the flows of priority class 10 are blocked (see Figure 6.24) due to higher packet losses for class 10, the delay bound violations occur in the neighboring priority class (class 9) of the poor channel quality video flows (soccer 2 and Crew 2 video flows). The MOS score along with average channel quality for all the video flows in scenario 2 is shown in Figure 6.26.

The mapping details for scenario 3 is shown in Figure 6.27. In this scenario, admission control policy blocks priority classes 9, 10, 11 and 12 during the time intervals shown in Figure 6.28. The performance of both the Ice video flows is approximately same as in the previous scenario. The only difference is that some delay bound violations occur for the Ice 2 video flow's quality layer 6 and 7 due to its lower channel quality.

Both the News video flows have different mapping as shown in Figure 6.27. It is important to note that not all the flows having packets mapped to priority class 9 are blocked. Higher packet losses in priority classes 10, 11 and 12 means that for one News video flow (News2); 5 quality layers are dropped whereas for the other (News 1), only 3

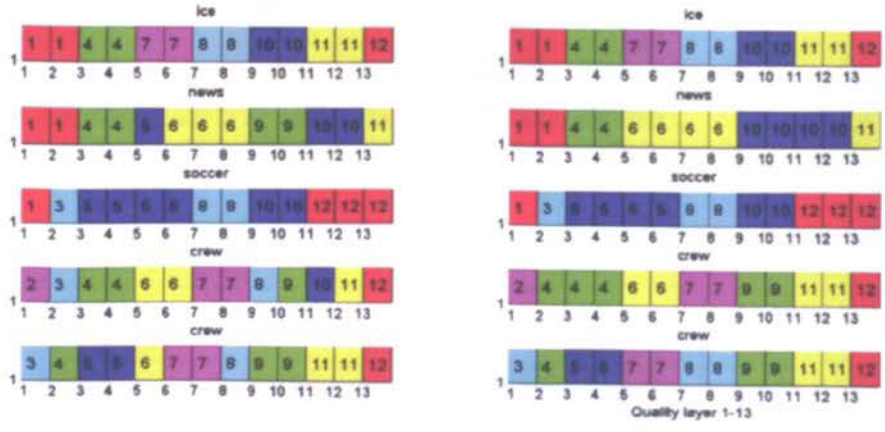


FIGURE 6.27: SVC layers mapping to priority classes for scenario 3 (mapping scheme elaborated with Bo Fu, DOCOMO).

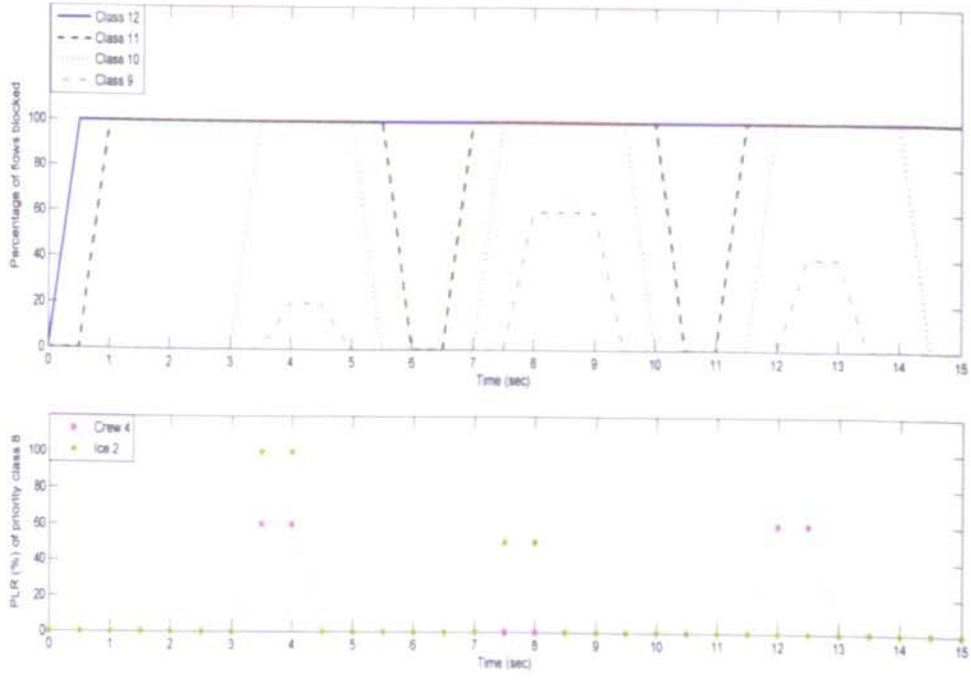


FIGURE 6.28: Scenario 3: Admission controller's blocking of priority classes throughout the simulation period.

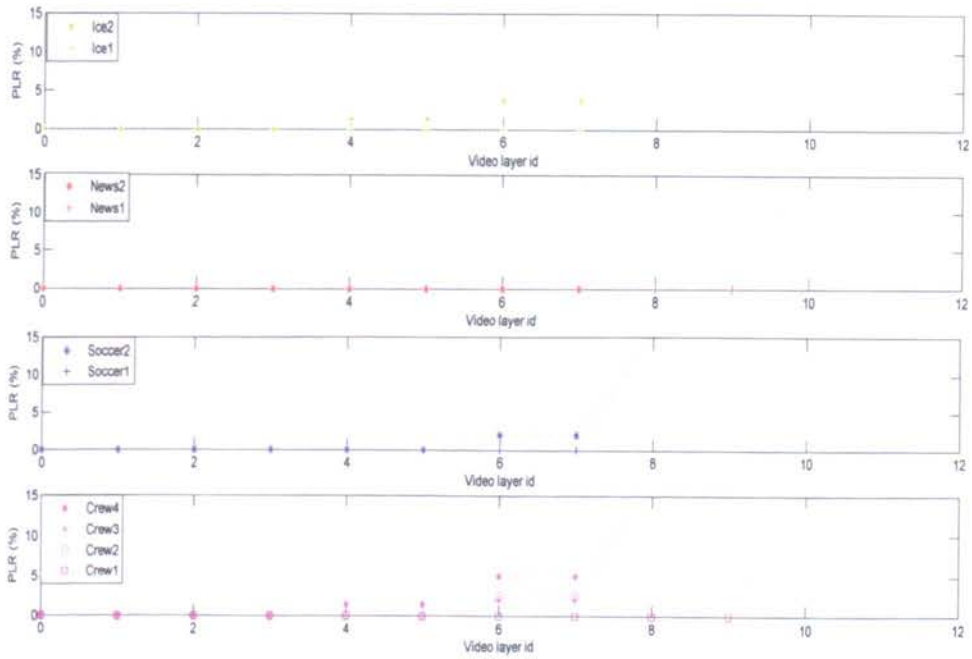


FIGURE 6.29: Scenario 3: PLR (%) in each of the SVC layers for each of the considered video flows. Total system PLR is 47.4 %.

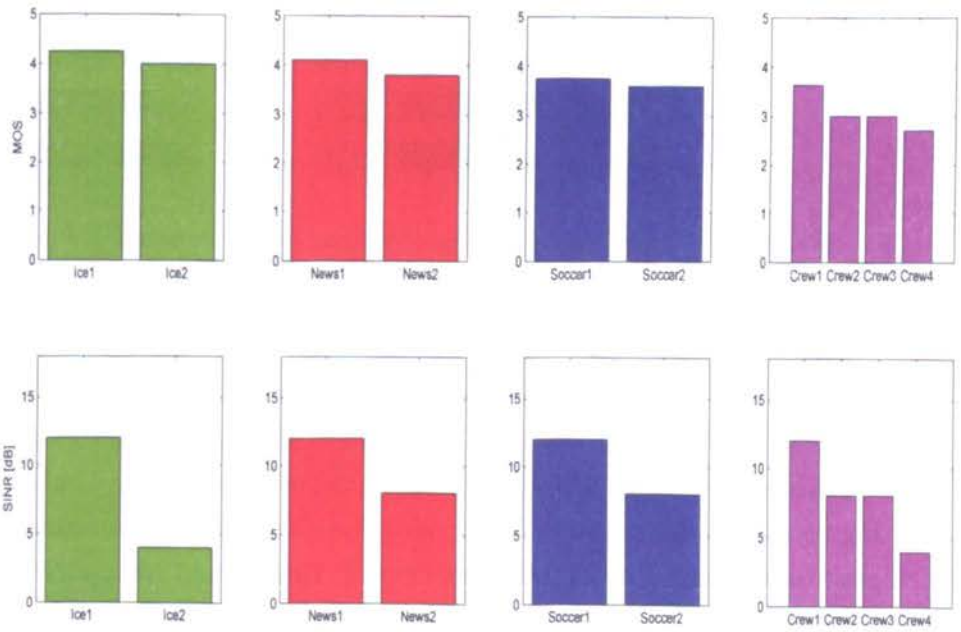


FIGURE 6.30: Video (MOS) and Channel quality (SINR) for each of the video flows in scenario 3. Average MOS per video flow is 3.59.

quality layers are dropped. The News 1 flow (note that this flow has the best channel quality) having quality layers assigned to priority class 9 are never blocked by the admission control strategy. The performance of both the Soccer flows is similar as in the previous scenario except that the PLR for Soccer 2 (Figure 6.29) is improved due to its better channel quality than the previous scenario. For Crew 2, 3 and 4 video flows, same number of video layers are blocked as shown in Figure 6.28. The crew 1 (this flow has the best channel quality same as News 1) flow having quality layers assigned to priority class 9 are never blocked by the admission control strategy. Thus for this video flow, only 3 quality layers (mapped to classes 10, 11 and 12) have a higher PLR. For the other Crew flows, priority classes 9, 10, 11 and 12 are blocked which causes higher PLR in last 5 quality layers for these flows. Crew 4 has the worst channel quality which causes delay bound violations in the priority class 8 as shown in the second sub-figure in Figure 6.28. The MOS score along with average channel quality for all the video flows in scenario 3 is shown in Figure 6.30.

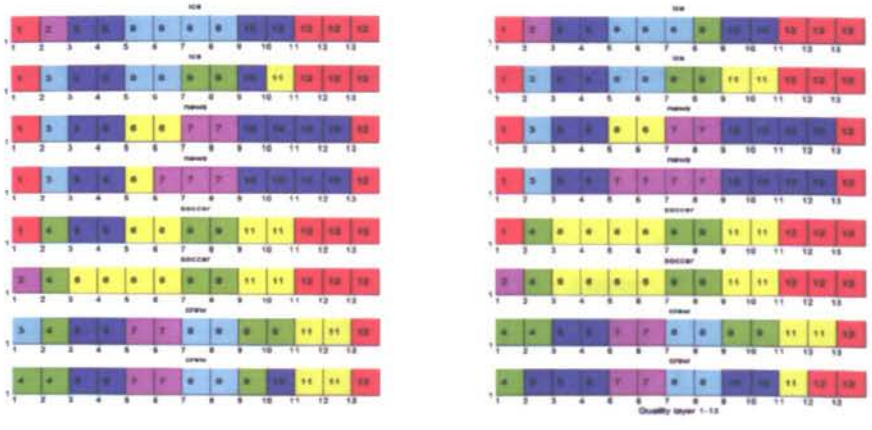


FIGURE 6.31: SVC layers mapping to priority classes for scenario 4 (mapping scheme elaborated by Bo Fu, DOCOMO).

The mapping details for scenario 4 is shown in Figure 6.31. In this scenario, admission control policy blocks priority classes 8, 9, 10, 11 and 12 during the time intervals shown in Figure 6.32. 3 of the 8 (37.5 % as shown in Figure 6.32) video flows having packets of priority class 8 are blocked according to the admission control policy (the policy uses the proportional fair rule to block video flows having packets marked with the current least important priority class). As Crew 2, Crew 4 and Ice 4 video flows have lower channel quality (shown in Figure 6.34, therefore these video flows are blocked (when the scheduling function triggers delay bound violations in priority class 8) from the time interval of 7.5 to 9.5 sec (very higher network congestion period). Video layers assigned to this priority class have a 100 % PLR during this interval. It is important to note

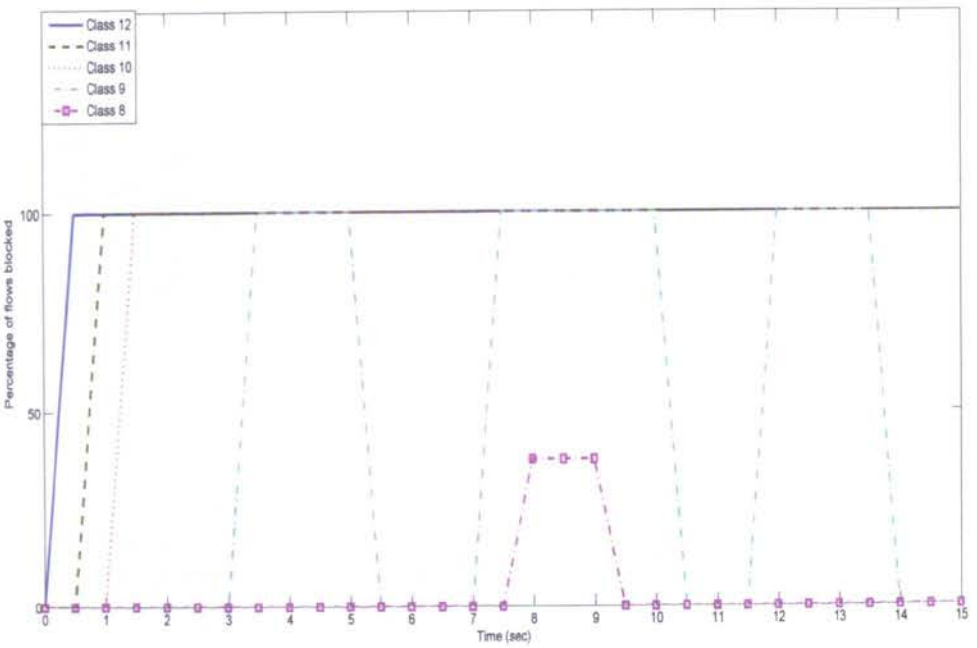


FIGURE 6.32: Scenario 4: Admission controller's blocking of priority classes throughout the simulation period.

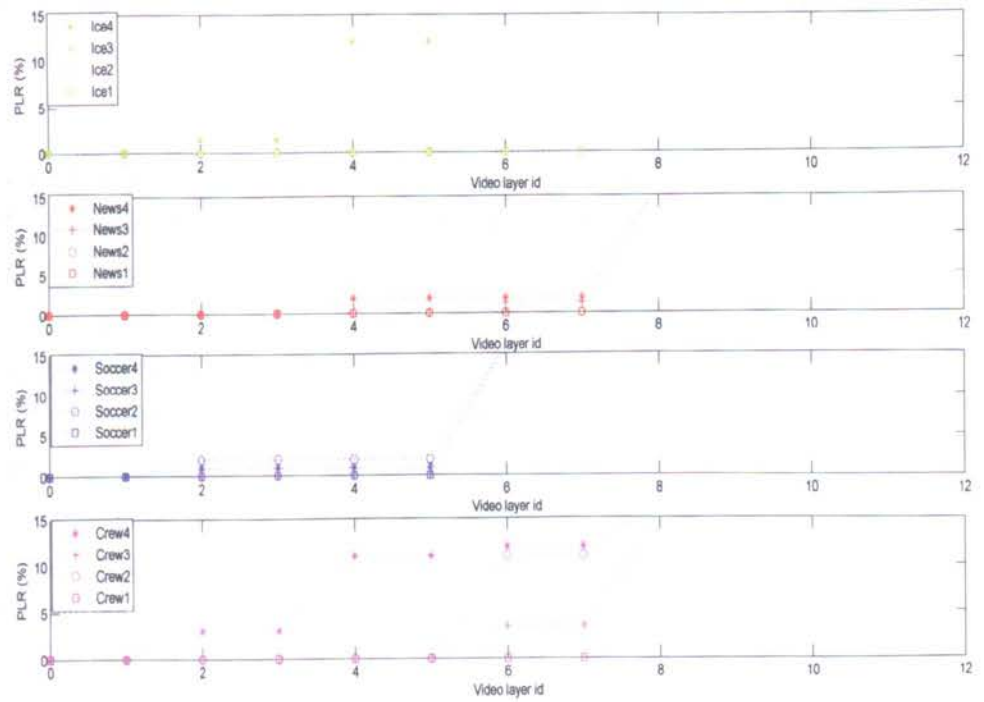


FIGURE 6.33: Scenario 4: PLR (%) in each of the SVC layers for each of the considered video flows. Total system PLR is 62.33 %.

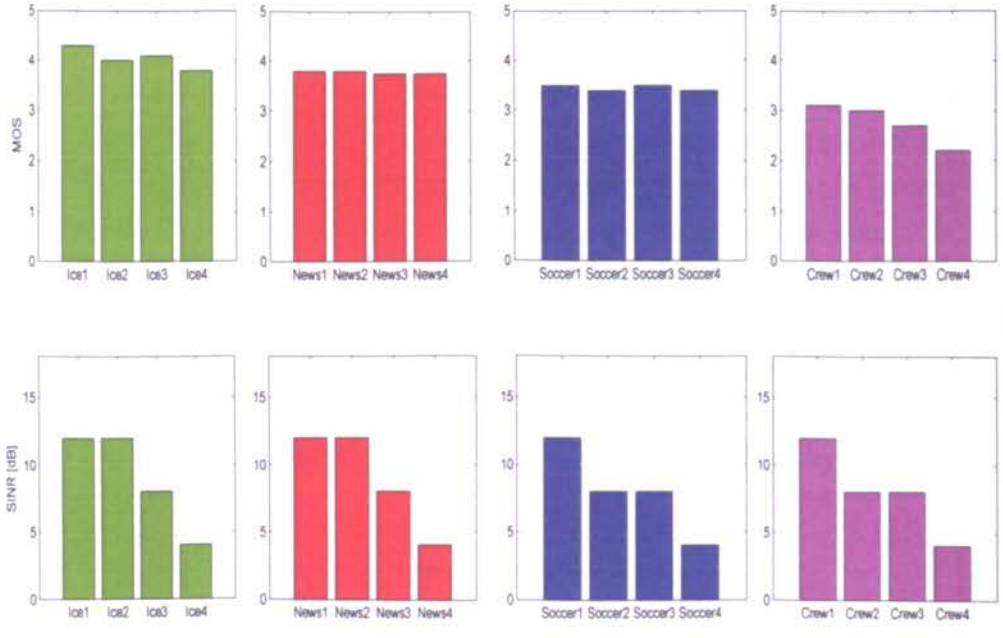


FIGURE 6.34: Video (MOS) and Channel quality (SINR) for each of the video flows in scenario 4. Average MOS per video flow is 3.52.

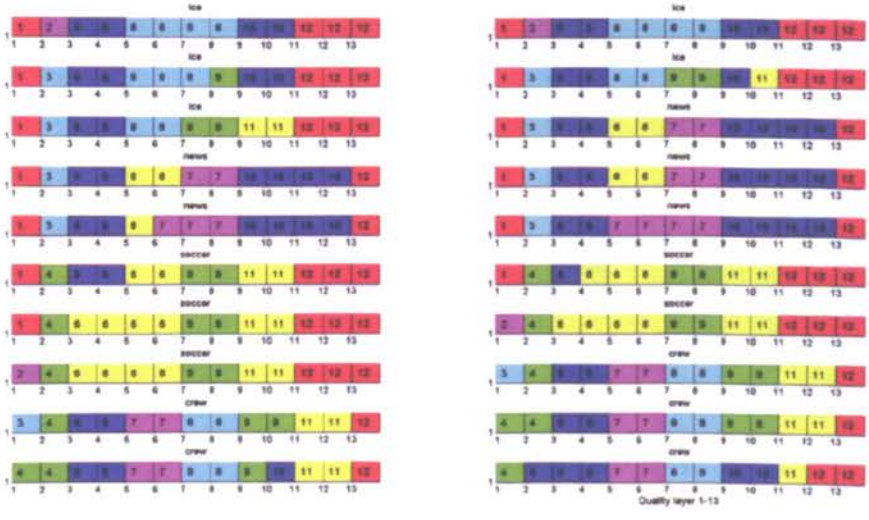


FIGURE 6.35: SVC layers mapping to priority classes for scenario 5 (mapping scheme elaborated by Bo Fu, DOCOMO).

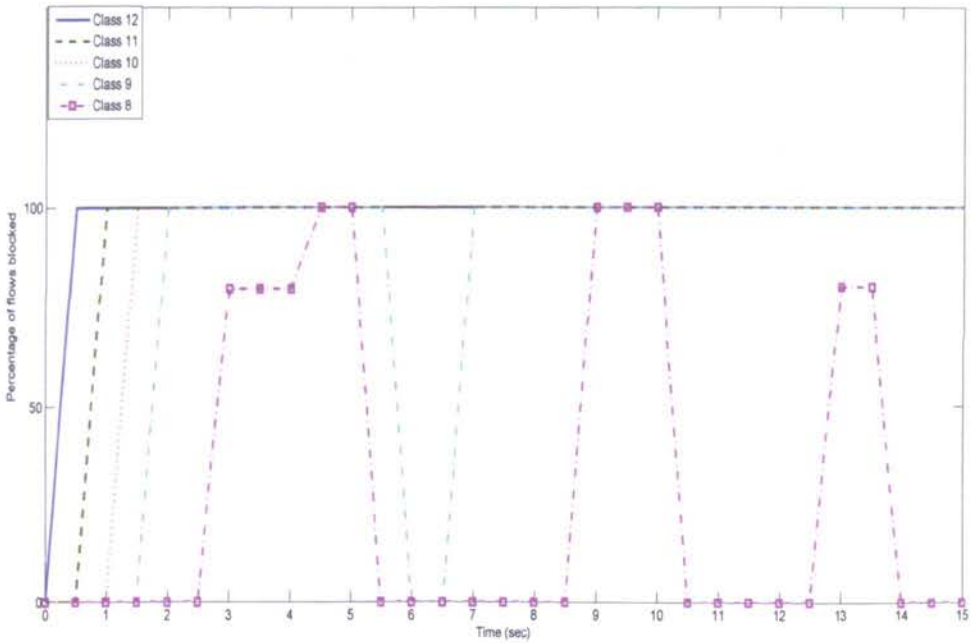


FIGURE 6.36: Scenario 5: Admission controller's blocking of priority classes throughout the simulation period.

that News 4 and Soccer 4 video flows have also lower average channel quality but after the blocking of priority classes 9, 10, 11 and 12 by the admission controller, the least important priority class for these video flows is 6 (for all Soccer flows) and 7 (for all News flows, see the mapping details in Figure 6.31). Therefore, the admission controller blocks Crew 2, Crew 4 and Ice 4 video flows priority class 8 from the time interval of 7.5 to 9.5 sec. The average MOS quality of all the video flows is shown in Figure 6.34. According to the figure, Ice video flows have different average MOS score. Specifically, Ice 1, Ice 2 and Ice 3 each have different quality layers successfully scheduled with 0 PLR as shown in Figure 6.33. This is due to the fact that the mapping of Ice video flows have different number of quality layers mapped to priority classes 9, 10 ,11 and 12 as shown in Figure 6.31. For instance, two of the Ice video flows have 5 quality layers mapped to priority classes 10 and 12 (one of these has one quality layer assigned to class 9 see Figure 6.31), the other two Ice flows have 6 and 7 quality layers assigned to priority classes above 8.

The mapping details for the last scenario is shown in Figure 6.35. In this scenario, admission control policy blocks priority classes 8, 9, 10, 11 and 12 during the time intervals shown in Figure 6.36. In this scenario, the average input traffic rate is approximately 3.4 times more than the average system capacity. The average system PLR is 70.6 %, however the peak PLR can be as high as 80 to 90 %. The PLR in each of the SVC

layers is shown in Figure 6.37. There is a higher PLR in the lower channel quality flows. For instance, Crew 4, Crew5, News 4, News 5, Soccer 4 and Soccer 5 have higher PLR in video layers shown in Figure 6.37. During high peak load periods, the instantaneous system capacity is not enough to schedule packets of higher priority classes of all the video flows. Therefore, the scheduler allocates more resources to good channel video flows. Only the lower channel quality video flows Ice 4 and Ice 5 have zero PLR. When the admission controller blocks the priority class 8, all the Ice video flows have only 1 base and 3 quality layer left (see Figure 6.35). The rate requirements of these remaining layers is as low as 100 Kbps. Hence none of the Ice flows' unblocked priority classes suffer from delay bound violations. The remaining higher rate and lower channel quality video flows suffer from higher PLR. The MOS score of the all the video flows is shown in Figure 6.38.

Crew 3, Crew 5, Soccer 4 and Soccer 5 video flows incur packet losses in the base layer, as shown in Figure 6.37. It is important to note that the base layer of the Crew video flows are either mapped to the third or fourth priority class as shown in Figure 6.35. Under higher traffic load, the probability of delay bound violations in the intermediate priority classes is higher, which results in packet losses in the base layers of the Crew 3 and Crew 5 video flows. The base layers of Soccer 4 and Soccer 5 video flows also suffer from delay bound violations, mainly due to the poor channel quality and a higher rate requirements of these flows (scheduling the base and five quality layers requires the maximum rate (see Figure 6.17) for the Soccer video sequence as compared to other considered video sequences).

The main goal of the mapping algorithm is to maximize the sum MOS of all the video flows. Under higher input traffic rate, the demanding video flows (in terms of the required bit rate to achieve a MOS value) are penalized, thus favoring flows attaining higher video quality at lower bit rate. At the eNodeB, the scheduler favors flows with better channel quality resulting in higher MOS for the flows with better channel quality. The proposed framework results in a considerably reduced cross-layer signaling between the eNodeB and the core network.

6.7.3 Conclusion

With the joint PPS scheduling and admission control policy, the number of video flows with zero PLR for the base layer is 17 (in Scenario 5, only Crew 3, Crew 5 and Soccer 5 have non-zero PLR in the base layer) as compared to the M-LWDF rule which can accommodate only 3 video flows (in Scenario 1, only Ice 1 and Ice 2 have zero PLR in

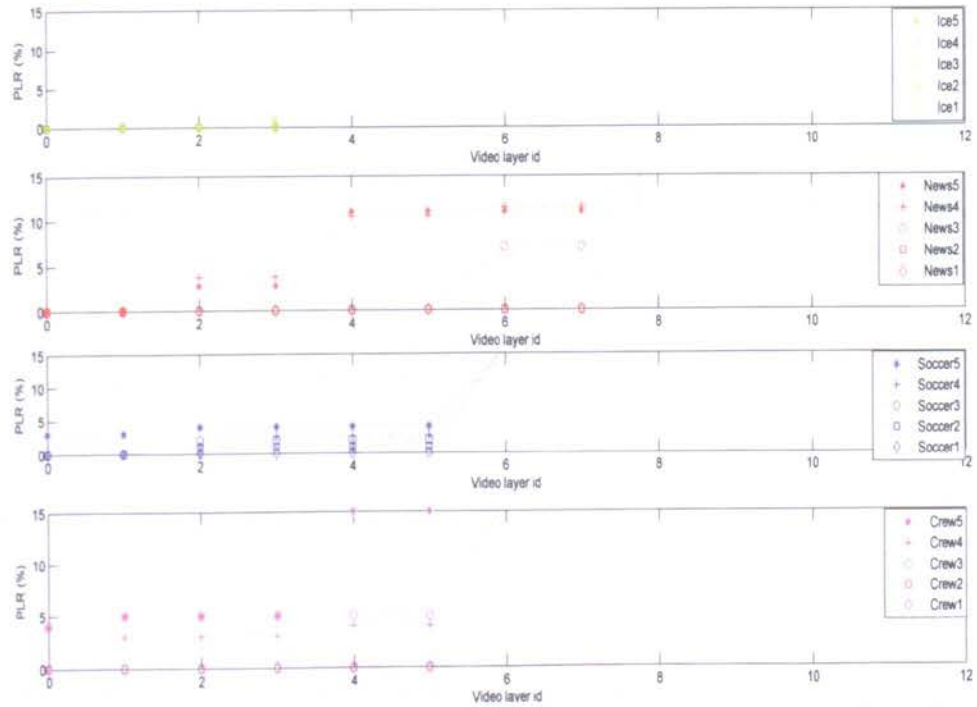


FIGURE 6.37: Scenario 5: PLR (%) in each of the SVC layers for each of the considered video flows. Total system PLR is 70.6 %.

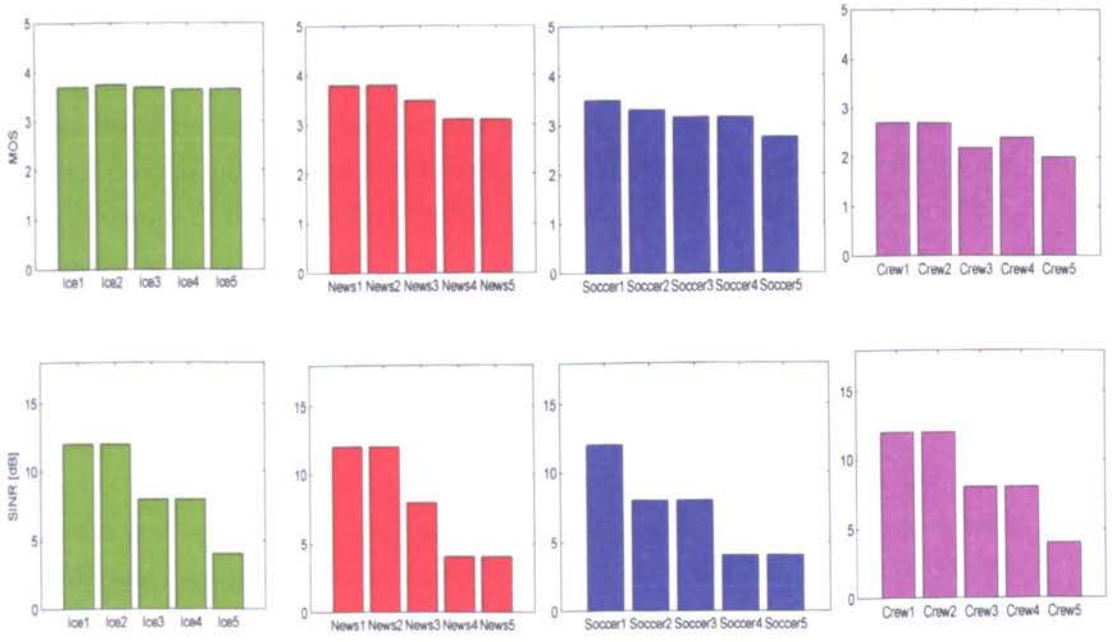


FIGURE 6.38: Video (MOS) and Channel quality (SINR) for each of the video flows in scenario 5. Average MOS per video flow is 3.23.

the base layer). The increase in the system capacity is approximately 5.66 times the considered benchmark rule.

In the subsequent sections the composite scheduling rule, along with the class based admission control policy under delay sensitive and delay tolerant traffic, is analyzed.

6.8 Scheduling rule for composite traffic

In order to consider a delay insensitive traffic class in the scheduling rule, the PPS scheduling rule is modified by defining a special class for best-effort traffic. Higher priority classes consist of important QoE layers for video streaming, as well as stringent delay based VoIP and video conferencing traffic flows. For best-effort traffic, the same priority function (PPS scheduling rule) but with different weight design is proposed. The main goal is to dynamically adjust the weight of the best-effort traffic class based on the QoS performance of the delay sensitive traffic classes.

The priority function depends upon the following four parameters:

- $\chi_{i,\varphi}^{(n)}$ is the channel quality on PRB φ.
- $R_{i,\text{ave}}^{(n)}$ is the time-averaged throughput.
- N_Q is the number of packets currently residing in the buffer of flow i at scheduling instant n.
- $W_i^{(n)}$ is the priority weight of the HoL packet. It is important to note that the proposed weight design depends on the system load. The higher the system load, the higher will be the normalized averaged HoL delay $\overline{A^{(n)}}$, which results in a higher weight for the packets of the most important priority classes. If the normalized averaged HoL delay is low, packets from different priority classes have approximately the same weights. For best-effort traffic, a weight function is defined which varies based on the QoE performance of the video traffic class. The admission control decision on blocking of the lower priority classes is utilized in defining the weight design of the best-effort traffic. Mathematically, the weight design of the composite scheduling rule is:

$$W_i^{(n)} = \begin{cases} \exp\left[(P_i^{(n)})^{\overline{A^{(n)}}} * A_i^{(n)}\right], & \text{if } j \in [0, J_{\max}] \\ C^{P_{j^*}^{(n)}} & \text{if } j \in \text{best} - \text{effort}, C < 1 \end{cases} \tag{6.28}$$

where C is a constant used to control the resource allocation between the delay sensitive and best-effort traffic classes. For instance let us assume a system with 8 QoE based priority classes as shown in Table 6.14. If the current lowest priority class in the blocking matrix is $j^* = 5$ (the admission controller has blocked class 6 and 7 based on the congestion in the system, therefore the current unblocked lowest priority class in the system has a packet priority $P_i^{(n)}$ equal to 3) and considering constant C equals to 0.5, then the priority weight $W_i^{(n)}$ for the best effort traffic

class is 0.5^3. The higher the number of blocked priority classes, the lower is the resource allocation probability for all the best-effort flows.

TABLE 6.14: Priority classes.

Priority Class Index, $j_i^{(n)}$	$P_i^{(n)}$
0	8
1	7
2	6
3	5
4	4
5	3
6	2
7	1

TABLE 6.15: Priority class mapping

Priority class index	**Target MOS**
0	2.5
1	2.75
2	3
3	3.25
4	3.5
5	3.75
6	4
7	4.25

6.8.1 Simulation Scenario

The performance of the PPS strategy combined with the class based admission control is evaluated here for composite traffic. Specifically, video streaming, VoIP and CBR based best-effort traffic are considered in the simulations. The QoE based marking of the video traffic is shown in Table 6.15. For each class a target MOS is assigned and packets of video layers achieving the target MOS are marked with the respective priority. The base layer will be assigned the highest priority, as high MOS values can only be achieved if the base layer is received with no delay bound violation. The class based admission control strategy is applied to the lower 4 priority classes as shown in Figure 6.39. Delay stringent VoIP traffic is assigned to higher priority classes where admission

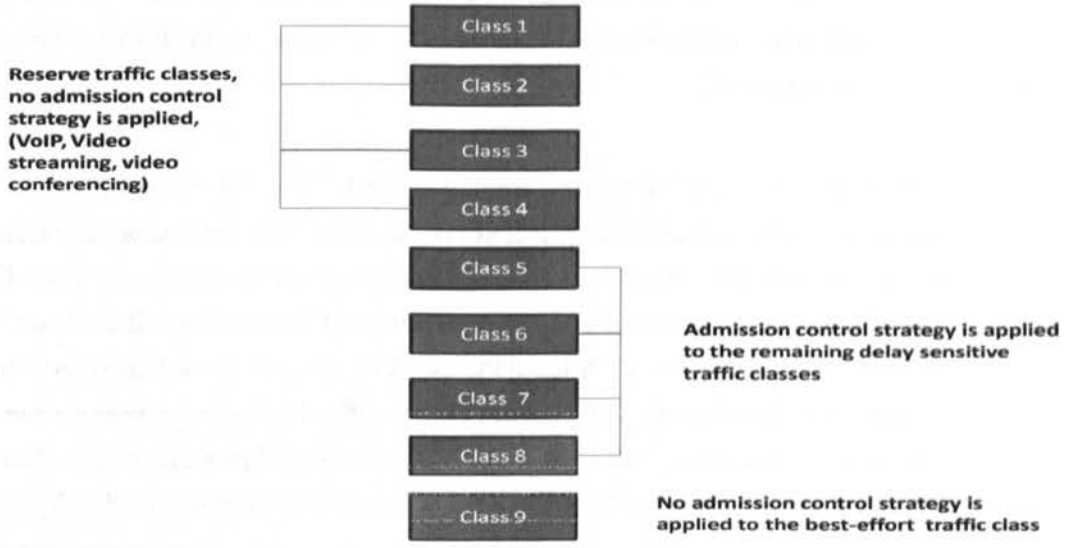


FIGURE 6.39: Priority classes assigned to the real-time and best-effort traffic types.

control strategy is not applied. Specifically, the VoIP flows are distributed equally to the reserved traffic classes.

In order to evaluate the fairness and efficiency performance of the PPS strategy coupled with class based admission control, state-of-the-art delay based schedulers are considered. Therefore, the M-LWDF rule (linear function of the HoL delay), the Log-rule (logarithmic function of the HoL delay) and the exponential rule (exponential function of the HoL delay) are simulated. A network with 18 video flows, 27 VoIP flows and 6 best-effort flows is simulated. All the users are uniformly distributed in the cell. According to this traffic combination with 27 VoIP and 18 video flows, real-time traffic requires an average spectral efficiency of 2.26 bits/sec/Hz (note that this is an average requirement: due to variable video traffic the peak requirement is greater), whereas best-effort traffic requires an average spectral efficiency of 0.72 bits/sec/Hz. The main goal is to analyze the results under very high load. Therefore, the system capacity (in terms of spectral efficiency) is fixed to 2.4 bits/sec/Hz whereas the average spectral efficiency requirement of the incoming traffic is increased to 3.6 bits/sec/Hz. Specifically, three different scenarios are simulated which are as follow:

- Scenario 1 - 27 VoIP flows, 18 video flows and 6 best-effort flows are simulated. The combined average spectral efficiency requirement is 2.73 bits/sec/Hz.
- Scenario 2 - A different traffic mix is simulated by increasing the video flows to 21 and reducing the VoIP flows to 24 flows whereas, the best-effort flows are fixed at 6. The combined average spectral efficiency requirement is 3 bits/sec/Hz.

- Scenario 3 - In the the third scenario, the average spectral efficiency requirement is further increased to 3.7. Specifically, 27 video, 18 VoIP and 6 best-effort flows are simulated.

The packet loss threshold of the base layer (first 2 priority classes) is set to 5 %. If the packet loss rate in the base layer is above the threshold, then none of the enhancement layers are decoded. Real-time traffic is characterized by video and VoIP flows. CBR flows with the data rate of 400 Kbps are selected for the best-effort flows, whereas 64 Kbps VoIP traffic with the threshold packet loss rate of 1 % and maximum delay budget of 100 ms is selected for VoIP users. Video traffic is generated from a trace file, where the average and peak traffic rates are 530 and 1500 kbps respectively. The maximum delay budget for video packets is 200ms, whereas the threshold packet loss rate is 5 %.

TABLE 6.16: Packet loss ratio of each traffic class for each of the considered scheduling rules for scenario 1.

Class	Total number of packets	PPS(C =0.2)	M-LWDF	LOG-RULE	EXP-RULE
1	5645	0	0.0187	0.0898	0.0397
2	6025	0	0.0319	0.1406	0.0404
3	4399	0	0.0380	0.1523	0.0533
4	4292	0	0.0263	0.1137	0.0435
5	4730	0.0036	0.0232	0.0996	0.0545
6	4798	0.0183	0.0292	0.1382	0.0398
7	4981	0.3104	0.0640	0.1600	0.0775
8	5179	0.8389	0.0542	0.1682	0.0919

6.8.2 Results

The PLR performance of the video traffic classes for all the considered scheduling rules is shown in Table 6.16. The video quality performance in terms of MOS is shown in Figure 6.40. According to the table, delay based scheduling rules incur packet losses over all the traffic classes, the worst being the log rule for which the packet loss rate in class 0 is as high as 8.08 %. This scheduling rule prioritizes the video traffic by using the logarithmic function of the HoL delay. Due to high PLR in the base layer video traffic class (class 0 and 1), the MOS performance for the lower channel quality flows is as low as 1.5. According to Figure 6.40, there are 11 video streaming users with a MOS quality of less than 3 (considered as a poor perceived video quality). Among the state-of-the-art scheduling rules, the M-LWDF scheduling rule performs best for the video streaming traffic. The delay based scheduling rules are highly opportunistic and

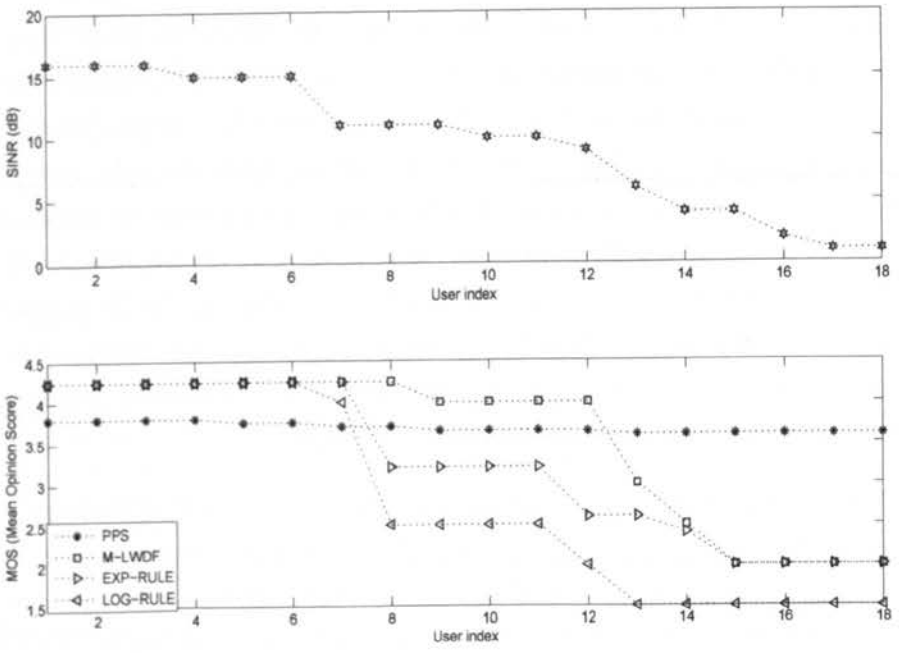


FIGURE 6.40: Performance of the video flows under scenario 1 for different scheduling rules.

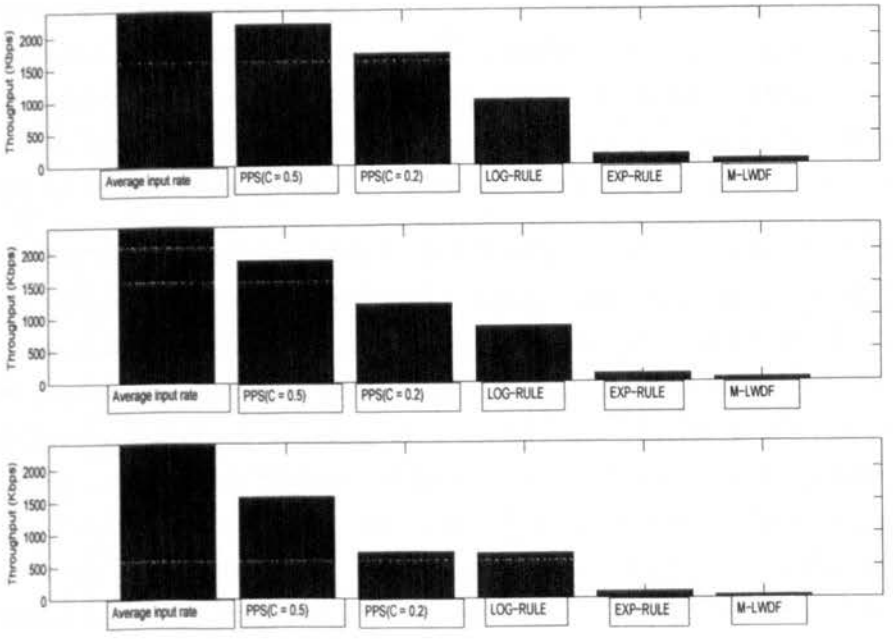


FIGURE 6.41: Performance of the best-effort flows in scenario 1, 2 and 3.

favor the flows with better channel quality, thus reducing video quality performance of the lower channel quality flows. Figure 6.41 shows the performance of all the considered scheduling rules for best-effort traffic. Among the state-of-the-art strategies, the log rule achieves the best performance for best-effort flows, but it is highly inefficient for video flows, as shown in Figure 6.40. It is evident from Figure 6.40 that when the average spectral efficiency requirement of all the flows is higher than the system capacity, the logarithmic, linear and exponential scheduling rules penalize the lower channel quality flows by incurring packet losses in the base layer. The exponential and linear functions of the HoL delay prioritize the video flows thus reducing the throughput performance of the best-effort traffic. The PLR performance of the VoIP flows are not shown here, since for all the scenarios the proposed and all the considered benchmark scheduling rules serve the VoIP flows within the delay bound.

When compared to state-of-the-art scheduling rules, the PPS strategy combined with class based admission control improves the fairness performance for video traffic flows. There is no delay bound violations in the first 4 priority classes, thus all the flows receive the base layer and 2 quality layers with no packet losses. Furthermore, delay bound violations in classes 5 and 6 are below the threshold, thus these contribute to the overall video quality. Under congestion, the exponential increase in the priority of higher priority classes reduces the delay bound violations in the base layer. Furthermore, packet dropping due to the admission control strategy in the lower priority classes fairly distributes the resources between best-effort and video traffic. Packet marking allows rate adaptation by dropping the least important priority classes thus reducing the packet losses in the higher priority classes. The packet marking strategy is based on the concept of max-min fairness, where the goal of the marking is to achieve the same video quality for all the video flows. On the other hand, consideration of the queue size in the scheduling decision increases the resource allocation probability of the best-effort traffic class.

When the number of video traffic flows increases, as stated in scenario 2, the performance of the video flows decreases further for all the delay based scheduling rules. The PLR in the base layer of the video traffic is above the threshold of 5 % as shown in Table 6.17. The video streaming quality of more than half of the video flows is below the MOS scale of 3 as shown in Figure 6.42. Furthermore, for scenario 3 none of the video flows receive a MOS value of 3 as shown in Figure 6.43. At this high load scenario, only VoIP flows receive the prescribed QoS performance whereas none of the video streaming and best-effort flows receive an adequate service. This shows that when a system is left to run into overload situations with no rate adaptation, the PPS scheduler with priority based rate adaptation can significantly increase the capacity of the system. The delay based scheduling rule can only benefit from having more VoIP flows in the system and less users for other traffic types. However, when the average input rate is more than the

achievable rate region, delay based scheduling rules with no rate adaptation degrade the quality of most of the flows thus reducing the system capacity under high loads. The advantage with the PPS strategy is that it reduces the need for regular end-to-end link probing as compared to other cross-layer optimization resource allocation frameworks. The proposed strategy only requires packet marking coupled with prioritized packet scheduling and class based admission control, thus reducing the frequent congestion status signal to the video servers or other traffic engineering modules at the P-GW.

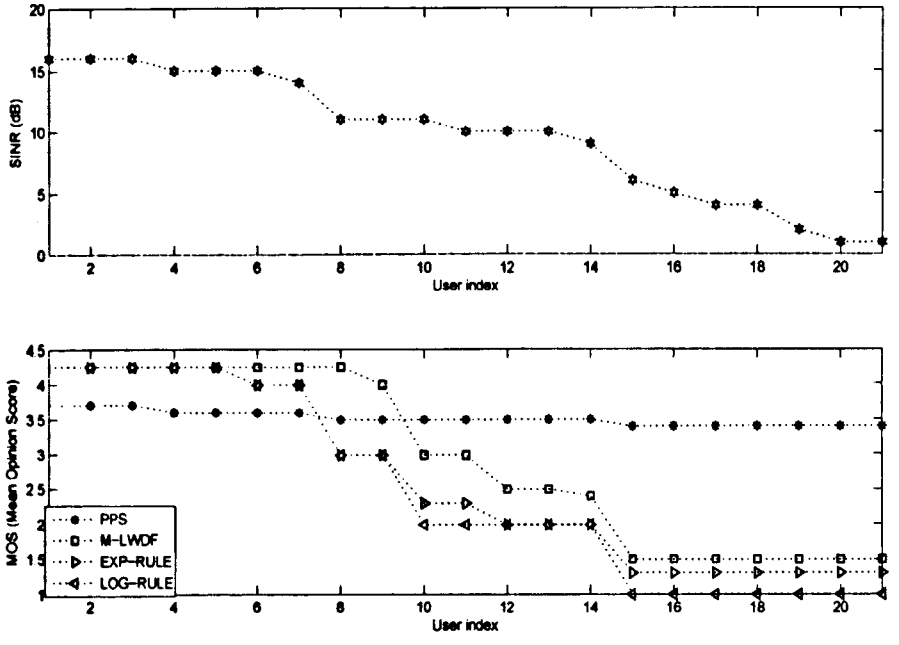


FIGURE 6.42: Performance of the video flows under scenario 2 for different scheduling rules.

TABLE 6.17: Packet loss ratio of each traffic class for each of the considered scheduling rules for scenario 2.

Class	Total number of packets	PPS(C =0.2)	M-LWDF	LOG-RULE	EXP-RULE
1	6616	0	0.0808	0.1421	0.1091
2	7017	0	0.1029	0.2028	0.1160
3	5225	0.0029	0.1082	0.2014	0.1334
4	5063	0.0063	0.0979	0.183	0.1420
5	5367	0.0325	0.0748	0.1543	0.1154
6	5877	0.1422	0.1133	0.1987	0.1339
7	5846	0.5855	0.1440	0.2172	0.1739
8	5746	0.9351	0.1336	0.2203	0.1797

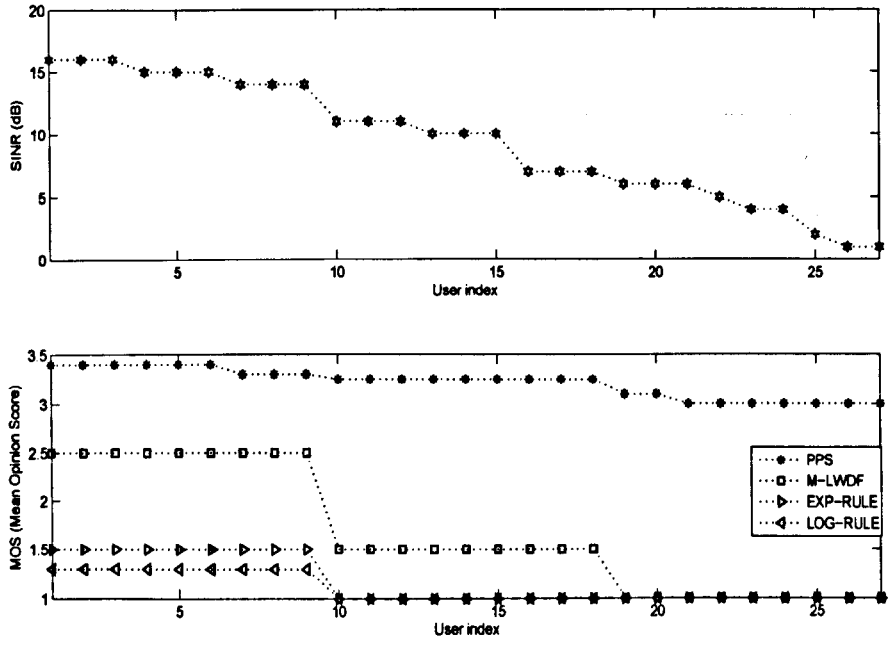


FIGURE 6.43: Performance of the video flows under scenario 3 for different scheduling rules.

TABLE 6.18: Packet loss ratio of each traffic class for each of the considered scheduling rules for scenario 3.

Class	Total number of packets	PPS(C =0.2)	M-LWDF	LOG-RULE	EXP-RULE
1	8406	0	0.1991	0.2569	0.2195
2	8981	0.0001	0.2394	0.3385	0.2601
3	6881	0.0020	0.2606	0.3255	0.3008
4	6789	0.01	0.2811	0.3172	0.3380
5	6763	0.2329	0.2063	0.2553	0.2639
6	7497	0.5237	0.2564	0.3053	0.2910
7	7587	0.9195	0.2823	0.3325	0.3302
8	7165	0.9859	0.2752	0.3377	0.3110

TABLE 6.19: QoS performance (packet loss ratio for delay sensitive traffic and throughput for the best-effort traffic) in scenario 1 for each of the considered traffic classes with different prioritization parameters.

Class	Total number of packets	PPS(C=0.2)	PPS(C=0.3)	PPS(C=0.4)	PPS(C=0.5)
1	5645	0	0	0	0
2	6025	0	0	0	0
3	4399	0	0	0	0
4	4292	0	0	0	0
5	4730	0.0036	0.0038	0.0034	0.0051
6	4798	0.0183	0.0289	0.0288	0.0590
7	4981	0.3104	0.3186	0.4678	0.4830
8	5179	0.8389	0.9140	0.9088	0.9325
9	10000	1738 Kbps	1854 Kbps	2074 Kbps	2194 Kbps

TABLE 6.20: QoS performance (packet loss ratio for delay sensitive traffic and throughput for the best-effort traffic) in scenario 2 for each of the considered traffic classes with different prioritization parameters.

Class	Total number of packets	PPS(C=0.2)	PPS(C=0.3)	PPS(C=0.4)	PPS(C=0.5)
1	6616	0	0	0	0
2	7017	0	0	0	0
3	5225	0.0029	0.0017	0.0005	0
4	5063	0.0063	0.0055	0.00237	0.0001
5	5367	0.0325	0.0317	0.0327	0.0193
6	5877	0.1422	0.1977	0.1896	0.3408
7	5846	0.5855	0.6547	0.7675	0.7750
8	5746	0.9351	0.9418	0.9449	0.9629
9	10000	1205 Kbps	1384 Kbps	1558 Kbps	1883 Kbps

The QoS performance for the best-effort traffic class (in terms of throughput) and MOS based traffic class (in terms of packet loss ratio) are shown in Table 6.19, 6.20 and 6.21 for each of the three scenarios. By increasing the priority weight parameter C for the best-effort traffic class, the throughput performance of the best-effort traffic class is increased at the expense of increased packet loss rate in the lower priority classes of video streaming traffic. When the video traffic flows are increased as reported in scenario 2 and 3, the throughput of the best-effort traffic class is decreased as shown in Table 6.20 and 6.21. The main reason of an improved performance for the best-effort traffic class when compared to the delay based scheduling strategy is the utilization of the same priority function for different traffic types. Traffic prioritization is achieved through the design of the weight function for each traffic type. Under higher congestion, lower

TABLE 6.21: QoS performance (packet loss ratio for delay sensitive traffic and throughput for the best-effort traffic) in scenario 3 for each of the considered traffic classes with different prioritization parameters.

Class	Total number of packets	PPS(C=0.2)	PPS(C=0.3)	PPS(C=0.4)	PPS(C=0.5)
1	8406	0	0	0	0
2	8981	0.0001	0.00022	0.00022	0.00022
3	6881	0.020	0.0023	0.0042	0.0029
4	6789	0.01	0.01	0.0154	0.0187
5	6763	0.2329	0.2468	0.30	0.3755
6	7497	0.5237	0.6395	0.7183	0.8001
7	7587	0.9195	0.9381	0.9309	0.9452
8	7165	0.9859	0.9871	0.9884	0.9894
9	10000	712 Kbps	992 Kbps	1275 Kbps	1577 Kbps

video priority classes are blocked, thus allowing higher priority classes to be scheduled within the delay budget constraints. Furthermore, the utilization of the queue size in the priority function design eliminates resource starvation for the best-effort traffic class.

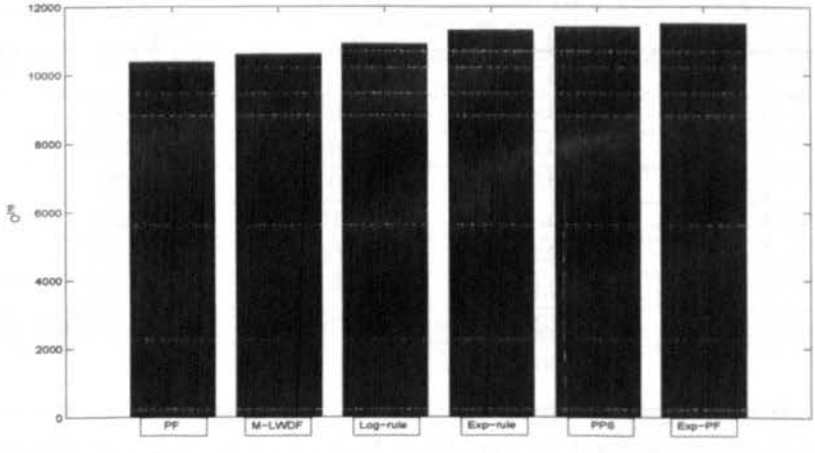


FIGURE 6.44: Number of machine cycles required for each of the considered scheduling rules.

6.9 Complexity analysis

The complexity in terms of the number of machine cycles at each scheduling instant of all the state-of-the-art scheduling rules is given in Table 6.22. According to the table, $O_{\mathrm{PF}}^{(n)}$ is the number of machines cycles required at each scheduling instant. I is the

number of flows and M_{PRB} is the number of the number of PRBs in the system. The complexity of each scheduling rule (100 flows and 100 PRBs) is shown in Figure 6.44.

TABLE 6.22: Number of machine cycle requirements for different scheduling functions.

Scheduling rule	Computation complexity $O^{(n)}$
PF	$I(4 + M_{\text{PRB}})$
M-LWDF	$I(6 + M_{\text{PRB}})$
EXP-PF	$I(15 + M_{\text{PRB}})$
LOG-RULE	$I(9 + M_{\text{PRB}})$
EXP-RULE	$I(13 + M_{\text{PRB}})$
PPS	$I(14 + M_{\text{PRB}})$

Chapter 7

Conclusion

The continued growth of mobile data traffic coupled with the advent of smart phones and tablets has triggered a new era for cellular communications. In order to cope with the continuously growing demand, operators world-wide are enhancing their network infrastructure by adding more capacity and services thus substantially increasing their investments. Mobile networks are evolving to an all-IP infrastructure which leads to a scenario where the volume of voice traffic on mobile networks would be considerably lower than video and data traffic. Furthermore, flat rate billing models are slowly making way for quality based billing models. Network operators must introduce new services and new billing models. The increase in the network capacity can never outspace the increase in mobile data traffic. Therefore, operators are working on the possible approaches of managing the huge increase in mobile traffic. The goal of the operators is to deliver an acceptable level of service quality under all network conditions. The importance of introducing new business models is increasing thus making way for efficient partitioning of network resources. Scheduling has a tremendous impact in implementing a desired operator's policy rules as it plays an important role in the prioritization of a particular traffic type and is the main mechanism for assigning resources.

This thesis proposed several scheduling strategies for delay-sensitive applications over the downlink of LTE networks as different operators may have different business models. The diverse models can range from providing simple QoS (in terms of packet delay, throughput and packet loss rate) to more fine level of service satisfaction by utilizing the content of the video traffic. Furthermore, the problem of traffic prioritization is also addressed by designing a framework which is tunable according to the service demands of a particular traffic type.

First an Opportunistic scheduling strategy is proposed which considers delay and packet loss rate in the scheduling decisions. One of the main goals of the proposed scheduling

rule is to reduce the QoS violations of the flows with relatively lower channel quality. The OPLF rule considers the channel quality iteratively (after assigning a resource) in the scheduling decisions thus fully exploiting the multi-user frequency diversity. In order to design a composite scheduling rule which considers both the real-time and best-effort traffic types, the fuzzy logic framework is utilized. The proposed fuzzy logic based scheduling strategy provides tunable prioritization which can be set to prioritize a particular traffic type. Different service needs of different traffic type are considered by the fuzzy logic controller. The main goal of the fuzzy logic controller is to minimize the QoS violations of the real-time flows while guaranteeing a minimum bandwidth to the best effort traffic. The novel concept of time-averaged channel quality is utilized by the fuzzy controller thus fully exploiting the instantaneous variations in the channel quality of different flows. The fuzzy logic priority scheme provides a considerable improvement in the QoS performance of real-time flows as compared to the benchmark scheduling rules. Furthermore, the performance of the best-effort traffic flows is also improved by efficiently using the robust fuzzy logic priority framework.

In order to efficiently exploit the characteristics of the video traffic, a scheduling function based on the novel concept of frame significance throughput is proposed. This scheduling function considers the importance of a frame in a GOP and thus provides fairness in terms of important video layers rather than bit throughput. This metric can be constructed easily as compared to complex objective video metrics which require complex application layer information. It is important to note that the proposed scheduling strategies are specifically designed for video streaming traffic. They assign a delay budget to the whole GOP as compared to QoS-aware packet scheduling strategies which schedules traffic on packet basis. Results show that the proposed strategy improves the objective video quality of the users with relatively lower channel quality without compromising the video quality of the users having good channel quality. Furthermore, the scheduling rule provides a good trade-off between fairness and efficiency which can be varied through the operators defined parameters.

Next, a composite scheduling rule is proposed which considers the strict delay requirements of the real-time traffic such as the VoIP and video conferencing flows and at the same time provides fine level of service satisfaction by utilizing the content of the video traffic. The proposed joint QoS and QoE strategy (PPS scheduling rule) targets at offering an appropriate trade-off between efficiency and fairness by considering packet importance in addition to the channel quality of each user, average throughput, and Head of Line (HOL) delay. It reduces the need for cross-layer signaling and frequent end-to-end link probing, through the exploitation of application layer packet marking and QoE based mapping of SVC layers into priority classes performed at the P-GW. The main goal of the PPS strategy is to minimize the delay bound violations for the most

important priority classes. Furthermore, admission control policy operating jointly with the PPS scheduler provide rate adaptation at the link layer via class-based admission control on the video layers. The performance of the proposed scheme has been assessed via simulation in comparison with state-of-the art schedulers. Results have been provided for different types of traffic (video, voice over IP, best effort). From the simulation results, following important conclusions are drawn

- when the proposed PPS scheduler is used in conjunction with the admission control strategy, there is a performance gain of a factor 5.66 in terms of system capacity with respect to the M-LWDF rule in the considered high load scenario.
- Different packet marking schemes are considered at the core network. The operators can implement sum-MOS based packet marking, where the goal is to maximize the over all video quality. On the other hand, the operators can also implement a max-min based packet marking, where the goal is to provide the same video quality in terms of MOS to all the video flows.
- The novel class-based admission control policy provides rate adaption at eNodeB thus providing graceful video quality degradation.
- The performance of the best-effort traffic flows is also analyzed in a loaded network with video and VoIP flows. Results show that the composite scheduling function provides service to the best effort flows by efficiently exploiting the higher delay tolerance property of the best-effort traffic.
- The complexity analysis performed highlighted that the complexity of the proposed scheduling strategy is comparable with existing non packet priority based strategies.

The scheduling strategies proposed in this thesis can serve as a basis for further research. The following paragraph briefly discusses the interesting subjects for future work in the area.

The considered video packet marking is based on H.264 SVC. Mobile devices running different video codecs require a common packet marking algorithm at the core network which can accommodate different video codecs. The network operators may employ Deep Packet Inspection (DPI) at the gateway of the core network to get information which can be useful in developing a content aware packet marking algorithms. Packet marking based on general video information such as distortion, motion vectors can be useful in accommodating different video codecs. Furthermore, the setting of the delay bound for packets of different video flows with different scalability levels is another challenge that has not been addressed yet.

References

[1] Cisco visual networking index: Global mobile data traffic forecast update 2009-2014. White Paper, February 2010.

[2] Cisco visual networking index: Global mobile data traffic forecast update 2011-2016. White Paper, February 2012.

[3] C. Mehlfhrer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp. Simulating the long term evolution physical layer. In *European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, UK, 2009.

[4] C. Mehlfhrer, J. C. Ikuno, M. imko, S. Schwarz, M. Wrulich, and M. Rupp. The Vienna LTE simulators - Enabling reproducibility in wireless communications research. *EURASIP Journal on Advances in Signal Processing*, 2011(29), July. 2011.

[5] 3GPP-TS 23.203 V11.7.0. Technical specification, policy and charging control architecture (release 11).

[6] E. Dahlman, S. Parkvall, J. Skold, and P. Beming. *3G Evolution: IISPA and LTE for Mobile Broadband.* Academic Press, 2007.

[7] J. Klaue, B. Rathke, and A. Wolisz. Evalvid - a framework for video transmission and quality evaluation. In *Computer Performance Evaluation / TOOLS'03*, Urbana, IL, USA, 2003.

[8] N. Khan, M. G. Martini, Z. Bharucha, and G. Auer. Opportunistic Packet Loss Fair Scheduling for Delay-Sensitive Applications over LTE Systems. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Paris, France, April. 2012.

[9] N. Khan, M. G. Martini, and Z. Bharucha. Quality-aware Fair Downlink scheduling for scalable video transmission over LTE systems. In *13th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cesme, Turkey, June. 2012.

[10] N. Khan, M. G. Martini, and D. Staehle. Opportunistic QoS-Aware Fair Downlink Scheduling for Delay Sensitive Applications using Fuzzy Reactive and Proactive Controllers. In *IEEE Vehicular Technology Conference (VTC)*, Las Vegas, USA, Sept. 2013.

[11] N. Khan, M. G. Martini, and D. Staehle. Opportunistic QoS-Aware Fair Downlink Scheduling for Delay Sensitive Applications using Fuzzy Reactive and Proactive Controllers. In *IEEE Vehicular Technology Conference (VTC)*, Las Vegas, USA, Sept. 2013.

[12] C. T. E. R. Hewage, M. G. Martini, and N. Khan. 3D medical video transmission over 4G networks. In *International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, Barcelona, Spain, Oct. 2011.

[13] B. S. Tsybakov. File transmission over wireless fast fading downlink. *IEEE Transactions on Information Theory*, 48:2323–2337, 2002.

[14] 3GPP - TS 36.323 Packet Data Convergence Protocol (PDCP) specification (Release 10). 2011.

[15] 3GPP - TS 36.322 Radio Link Control (RLC); protocol specification (Release 10). 2011.

[16] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose. Modelling TCP Reno performance, "a simple model and its empirical validation". *IEEE Transactions on Networking*, 8(2):133–145, 2000.

[17] 3GPP - TS 36.331 Radio Resource Control (RRC); protocol specification (Release 10). 2011.

[18] 3GPP - TS 36.331 Medium Access Control (MAC); protocol specification (Release 10). 2011.

[19] S. Tsai L. Wan and M. Almgren. A fading-insensitve performance metric for a unified link quality model. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Las Vegas, USA.

[20] X. He, K. Niu, Z. He, and J. Lin. Link layer abstraction in MIMO-OFDM system. In *International Workshop on Cross Layer Design (IWCLD)*, Jinan, China, Sept. 2007.

[21] N. Kolehmainen, J. Puttonen, P. Kela, T. Ristaniemi, T. Henttonen, and M. Moisio. Channel quality indication reporting schemes for UTRAN long term evolution downlink. In *IEEE Vehicular Technology Conference (VTC)*, Singapore, May. 2008.

[22] Technical Specification Group RAN, E-UTRA; physical channels and modulation, 3GPP Tech. Rep. TS 36.211 Version 8.7.0. 2009.

[23] Technical Specification Group RAN, E-UTRA; multiplexing and channel coding, 3GPP Tech. Rep. TS 36.212. 2009.

[24] Technical Specification Group RAN, E-UTRA; physical layer procedures, 3GPP Tech. Rep. TS 36.213. 2009.

[25] J. Kim, A. Ashikhmin, A. van Wijngaarden, E. Soljanin, and N. Gopalakrishnan. On efficient link error prediction based on convex metrics. In *IEEE Vehicular Technology Conference (VTC)*, Los Angeles, CA, USA, Sept. 2004.

[26] S. Tsai and A. Soong. Effective-SNR mapping for modeling frame error rates in multiple-state channels. *3GPP2, Tech. Rep. 3GPP2-C30-20030429-010*, 2003.

[27] T. Okumura, E. Ohmore, and K. Fukuda. Field strength and its variability in vhf and uhf land mobile service. *Rev. Elec. Commun. Lab*, pages 825–873, 1968.

[28] M. Hata. Empirical formula for propagation kiss ub kabd nibuke radui services. *IEEE Transactions on Vehicular Technology*, pages 317–325, 1980.

[29] M. Wrulich and M. Rupp. Computationally efficient MIMO HSDPA system-level modeling. *EURASIP Journal on Wireless Communications and Networking 2009*, 2009.

[30] M. Wrulich and M. Rupp. Performance and modeling of LTE H-ARQ. In *International ITG Workshop on Smart Antennas (WSA)*, Berlin, Germany, 2009.

[31] D. Skoutas, D. Komnakos, D. Vouyioukas, and A. Rouskas. Enhanced dedicated channel scheduling optimization in WCDMA. In *European Wireless Conference (EW)*, Prague, Czech Republic, 2008.

[32] S. Schwarz, C. Mehlfhrer, and M. Rupp. Low complexity approximate maximum throughput scheduling for LTE. In *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2010.

[33] R. V. Rasmussen and M. A. Trick. Round robin scheduling - a survey. Technical report, European Journal of Operational Research, 2006.

[34] A. Jalali, R. Padovani, and R. Pankaj. Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System. In *IEEE Vehicular Technology Conference (VTC)*, Tokyo, Japan, May. 2000.

[35] H. Kim and Y. Han. A Proportional Fair Scheduling for Multicarrier Transmission Systems. *IEEE Communications Letters*, 9(3), 2005.

[36] F. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8, 1997.

[37] R. Jain, D. Chiu, and W. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *Tech. Rep. TR-301*, 1984.

[38] http://trace.eas.asu.edu/. H.264/AVC and SVC video trace library, .

[39] P. Seeling, M. Reisslein, and B. Kulapala. Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial. *IEEE Communications Surveys & Tutorials*, 6(3):58–78, 2004.

[40] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar. Providing Quality of Service over a Shared Wireless Link. *IEEE Communications Magazine*, 39(2):150–154, Feb. 2001.

[41] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijaykumar, and P. Whiting. CDMA data QoS scheduling on the forward link with variable channel conditions. *Bell Labs Technical Memo No. 10009626-000404-05TM*, 2000.

[42] H. A. M. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand. Performance of Well Known Packet Scheduling Algorithms in Downlink 3GPP LTE System. In *IEEE Malaysia International Conference on Communications (MICC)*, Kuala Lumpur, Malaysia, Dec. 2009.

[43] J. H. Rhee, J. M. Holtzman, and D. K. Kim. Performance Analysis of the Adaptive EXP/PF Channel Scheduler in an AMC/TDM System. *IEEE Communications Letters*, 8(8):497–499, Aug. 2004.

[44] J. H. Rhee, J. M. Holtzman, and D. K. Kim. Scheduling of Real and Non-real Time Services: Adaptive EXP/PF Algorithm. In *IEEE Vehicular Technology Conference (VTC)*, Jeju Island, Republic of Korea, April. 2003.

[45] A. K. F. Khattab and K. M. F. Elsayed. Opportunistic Scheduling of Delay Sensitive Traffic in OFDMA-based Wireless Networks. In *International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Buffalo-NY, USA, June. 2006.

[46] K. Sandrasegaran, H. A. M. Ramli, and R. Basukala. Delay-Prioritized Scheduling (DPS) for Real Time Traffic in 3GPP LTE System . In *IEEE Wireless Communications and Networking Conference (WCNC)*, Sydney, Australia, April. 2010.

[47] S. Shin and B. H. Ryu. Packet Loss Fair Scheduling Scheme for Real-Time Traffic in OFDMA Systems. *ETRI Journal*, 26(5):391–396, 2004.

[48] C. Koksal, H. Kassab, and H. Balakrishnan. An Analysis of Short Term Fairness in Wireless Media Access Protocol. In *ACM SIGMETRICS*, California, USA, June. 2000.

[49] Tech. rep. m.1225: Guidelines for evaluation of radio transmission technologies for IMT-2000. *Geneva, Switzerland*, 1997.

[50] Technical Specification Group GSM/EDGE Radio Access Network, Radio transmission and reception, Annex C.3 Propagation Models, 3GPP, Tech. Rep. TS 05.05 V.8.20.0 (Release 1999).

[51] 1xev-dv evaluation methodology- addendum (v6). *3GPP2 WG5 Evaluation AHG.*, 2001.

[52] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch. Multicarrier OFDM with Adaptive Subcarrier, Bit, and Power Allocation. *IEEE Journal on Selected Areas in Communications*, 17(10):1747–1758, Oct 1999.

[53] J. Jang and K. B. Lee. Transmit Power Adaptation for Multiuser OFDM Systems. *IEEE Journal on Selected Areas in Communications*, 21(2):171–178, Feb. 2003.

[54] W. Rhee and J. M. Cioffi. Increasing in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation. In *IEEE Vehicular Technology Conference (VTC)*, California, USA, May. 2000.

[55] J. Jang and K. B. Lee. Adaptive Resource Allocation in Multiuser OFDM Systems with Proportional Fairness. *IEEE Trans. on Wireless Communications*, 21(2):171-178, Feb. 2003.

[56] J. Gross, H. Karl, F. Fitzek, and A. Wolisz. Comparison of heuristic and optimal subcarrier assignment algorithms. In *International Conference on Wireless Networks (ICWN)*, Las Vegas, USA, June. 2003.

[57] J. Gross, J. Klaue, H. Karl, and A. Wolisz. Subcarrier allocation for variable bit rate video streams in wireless OFDM systems. In *IEEE Vehicular Technology Conference (VTC)*, Florida, USA, October. 2003.

[58] A. K. F. Khattab and K. M. F. Elsayed. Opportunistic Subcarrier Management for Delay Sensitive Traffic in OFDMA-based Wireless Multimedia Networks. In *IST Mobile and Wireless Communications Summit*, Dresden, Germany, June. 2005.

[59] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda. Downlink packet scheduling in LTE cellular networks: Key design issues and a survey. *IEEE Communications Surveys and Tutorials*, pages 1–23, 2012.

[60] B. Sadiq, R. Madan, and A. Sampath. Downlink scheduling for multiclass traffic in LTE. *EURASIP Journal on Wireless Communications and Networking*, 2009, 2009.

[61] D. Liu and Y. II. Lee. An efficient scheduling discipline for packet switching networks using Earliest Deadline First Round Robin. In *International Conference on Computer Communications and Networks(ICCCN)*, Dallas, USA, Oct. 2003.

[62] J.M. Holtzman. Asymptotic analysis of proportional fair algorithm. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, San Diego, California, USA, Oct. 2001.

[63] G. Fodor, A. Furuskar, P. Skillermark, and J. Yang. On the impact of uplink scheduling on intercell interference variation in MIMO OFDM systems. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Budapest, Hungary, April. 2009.

[64] K.C. Beh, S. Armour, and A. Doufexi. Joint time-frequency domain proportional fair scheduler with HARQ for 3GPP LTE systems. In *IEEE Vehicular Technology Conference (VTC)*, Calgary, AB, Canada, Sept. 2008.

[65] Jan Jantzen. Tutorial on fuzzy logic. Technical report, 1998.

[66] R. G. Garroppo, S. Giordano, D. Iacono, and L. Tavanti. Game theory and time utility functions for a radio aware scheduling algorithm for WiMAX networks. *Wireless networks*, 17(6):1441–1469, 2011.

[67] R. G. Garroppo, S. Giordano, and D. Iacono. Radio-aware scheduler for WiMAX systems based on time-utility function and game theory. In *IEEE Global Communications Conference (GLOBECOM)*, Hawaii, USA, Nov. 2009.

[68] S. Ali and M. Zeeshan. A utility based resource allocation scheme with delay scheduler for LTE service-class support. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Paris, France, April. 2012.

[69] S. Ali, M. Zeeshan, and A. Naveed. A capacity and minimum guarantee-based service class-oriented scheduler for lte networks. *Eurasip Journal on Wireless Communication and Networking*, 2013, 2013.

[70] M. Iturralde, A. Wei, T. Yahiya, and A. Beylot. Performance Study of Multimedia Services Using Virtual Token Mechanism for Resource Allocation in LTE Networks. In *IEEE Vehicular Technology Conference (VTC)*, San Francisco, USA, Sept. 2011.

[71] J. Shin, J. W. Kim, and C. C. Jay Kuo. Quality-of-service mapping mechanism for packet video in differentiated services network. *IEEE Transactions on Multimedia*, 3(2):219–231, 2001.

[72] 3GPP, Tech. Specif. Group Services and System Aspects - Policy and charging control architecture (Release 9), 3GPP TS 23.203. 2009.

[73] P. Seeling and M. Reisslein. Video transport evaluation with H.264 video traces. *IEEE Communications Surveys and Tutorials, in print*, 14(4):1142–1165, 2012. Traces available at `trace.eas.asu.edu`.

[74] S. Schwarz, C. Mehlfuhrer, and M. Rupp. Throughput Maximizing Multiuser Scheduling with Adjustable Fairness. In *IEEE International Conference on Communications(ICC)*, Kyoto, Japan, June. 2011.

[75] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, 2007.

[76] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

[77] R. S. Tupelly, J. Zhang, and E. K. P. Chong. Opportunistic scheduling for streaming video in wireless networks. In *Conference on Information Sciences and Systems*, Johns Hopkins University, Baltimore, Md, USA, 2003.

[78] G. Liebl, H. Jenkac, T. Stockhammer, and C. Buchner. Radio link buffer management and scheduling for wireless video streaming. *Telecommunication Systems*, pages 255–277, 2005.

[79] G. Liebl, M. Kalman, and B. Girod. Deadline-aware scheduling for wireless video streaming. In *IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, Netherlands, July. 2005.

[80] M. van der Schaar, Y. Andreopoulos, and Zhiping Hu. Optimized scalable video streaming over IEEE 802.11a/e HCCA wireless networks under delay constraints. *IEEE Trans. on Mobile Computing*, 5(6):755–768, 2006.

[81] Xin Ji, Jianwei Huang, Mung Chiang, Gauthier Lafruit, and Francky Catthoor. Scheduling and resource allocation for SVC streaming over OFDM downlink systems. *IEEE Trans. on Circuits and Systems*, 19(10):1549–1555, 2009.

[82] L. Choi, W. Kellerer, and E. Steinbach. On cross-layer design for streaming video delivery in multiuser wireless environments. *Eurasip Journal on Wireless Communication and Networking*, 2006:1–10, 2006.

[83] E. Rasmusen. *Games and information: An introduction to game theory.* Wiley-blackwell, 2007.

[84] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[85] H. Boche and M. Schubert. Nash bargaining and proportional fairness for wireless systems. *IEEE Transactions on Networking*, 17(5):1453–1466, 2009.

[86] http://trace.eas.asu.edu/. H.264/AVC and SVC video trace library, .

[87] P. Li, H. Zhang, B. Zhao, and S. Rangarajan. Scalable video multicast with adaptive modulation and coding in broadband wireless data systems. *IEEE Transactions on Networking*, 20(1):57–68, Feb. 2012.

[88] N. Freris, C. Hsu, J. Singh, and X. Zhu. Distortion-aware scalable video streaming to multi-network clients. *IEEE Transactions on Networking*, 21(2), June. 2012.

[89] X. Liu, E. K. P. Chong, and N. B. Shroff. Transmission Scheduling for Efficient Wireless Utilization. In *IEEE Conference on Computer Communications (INFOCOM)*, Anchorage, Alaska, USA, April. 2001.

[90] V. Tsibonis, L. Georgiadis, and L. Tassiulas. Information theory information for throughput maximization. *IEEE Transactions on Information Theory*, 50(11): 2566–2582, 2004.

[91] A. Eryilmaz and R. Srikant. Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. In *IEEE Conference on Computer Communications (INFOCOM)*, Miami, USA, March. 2005.

[92] N. Nasser, B. Al-Manthari, and H. Hassanein. A performance comparison of class-based scheduling algorithms in future UMTS access. In *IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Phoenix, USA, April. 2005.

[93] G. Manfredi, P. Annese, and U. Spagnolini. A channel aware scheduling algorithm for HSDPA system. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Berlin, Germany, Sep. 2005.

[94] M. G. Martini and V. Tralli. Video quality based adaptive wireless video streaming to multiple users. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Las Vegas, NV, April. 2008.

[95] G. Liebl, T. Stockhammer, C. Buchner, and A. Klein. Radio link buffer management and scheduling for video streaming over wireless shared channels. In *Packet Video Workshop*, Irvine, CA, USA, December 2004.

[96] G. Liebl, H. Jenkac, T. Stockhammer, and C. Buchner. Radio link buffer management and scheduling for wireless video streaming. *Telecommunication Systems, Springer Science & Business Media B.V.*, 30/1-3:255–277, November 2009.

[97] P. V. Pahalawatta, R. Berry, T. N. Pappas, and A. K. Katsaggelos. Content-aware resource allocation and packet scheduling for video transmission over wireless networks,. *IEEE Journal on Selected Areas in Communications (JSAC)*, 25 (4):749–759, 2007.

[98] P. Sudame and B. R. Badrinath. On providing support for protocol adaptation in mobile wireless networks. Technical report, Mobile Networks and Applications, 1997.

[99] Q. Wang and M. A. Abu-Rgheff. Cross layer signaling for next-generation wireless systems. In *IEEE Wireless Communications and Networking Conference (WCNC)*, New Orleans, USA, March. 2000.

[100] W. Kellerer, L. U. Choi, and E Steinbach. Cross-layer adaptation for optimized B3G service provisioning. In *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Yokosuka, Japan, Oct. 2003.

[101] H. Jiang, W. Zhuang, and X. Shen. Cross-layer design for resource allocation in 3G wireless networks and beyond. *IEEE Communications Magazine*, 43(12):120-126, Dec. 2005.

[102] R. A. Berry, , and E. Yeh. Cross-layer wireless resource allocation. *IEEE Signal Processing Magazine*, 21(5):59–68, Sept. 2004.

[103] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, 1(1):1401-1415, 2006.

[104] F. Fu and M. van der Schaar. Decomposition Principles and Online Learning in Cross-Layer Optimization for Delay-Sensitive Applications. *IEEE Transactions on signal Processing*, 58(3):1401–1415, Mar. 2010.

[105] T. Fingscheidt, T. Hindelang, R. V. Cox, and N. Seshadri. Joint source-channel (de-)coding for mobile communications. *IEEE Transactions on Communications*, 50(2):200–212, 2002.

[106] T. Breddermann, H. Luders, P. Vary, I. Aktas, and F. Schmidt. Iterative source-channel decoding with cross-layer support for wireless VoIP. In *ITG Conference on Source and Channel Coding (SCC)*, Siegen, Germany, Jan. 2010.

[107] M. G. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon. Content adaptive network aware joint optimization of wireless video transmission. *IEEE Communications Magazine*, 45(3):1–10, 2007.

[108] J. Huusko, J. Vehkaperä, P. Amon, C. Lamy-Bergot, G. Panza, J. Peltola, and M. G. Martini. Cross-layer architecture for scalable video transmission in wireless network. *Signal Processing: Image Communication*, 22(3):317–330, 2007.

[109] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer. MOS-Based Multiuser Multiapplication Cross-Layer Optimization for Mobile Multimedia Communication. *Advances in Multimedia*, 2007.

[110] A. Saul. Wireless resource allocation with perceived quality fairness. In *IEEE Annual Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov 2008.

[111] B. J. Kim. A network service providing wireless channel information for adaptive mobile applications. i. proposal. In *IEEE International Conference on Communications (ICC)*, Helsinki, Finland, June 2001.

[112] L. A. Larzon, U. Bodin, and O. Schelen. Hints and notifications [for wireless links]. In *IEEE Wireless Communications and Networking Conference (WCNC)*, Orlando, Florida, USA, March 2002.

[113] A. Takacs, A. Kovacs, I. Godor, F. Kalleitner, H. Brand, M. Ek, T. Stefansson, and F. Sjoberg. The layer-independent descriptor concept. *Journal of Computers*, 1(2):23–32, 2006.

[114] S. Thakolsri, S. Cokbulan, D. Jurca, Z. Despotovic, and W. Kellerer. QoE-driven cross-layer optimization in wireless networks addressing system efficiency and utility fairness. In *IEEE Workshop on Multimedia Communications and services (GLOBECOM Workshops)*, Houston, USA, Dec. 2011.

[115] S. Thakolsri, W. Kellerer, S. Khan, and E. Steinbach. Application-driven cross layer optimization in wireless networks. In *Seminar on Service Quality Evaluation in Wireless Netowrks supported by COST 290*, Stuttgart, Germany, June. 2007.

[116] S. Thakolsri, W. Kellerer, S. Khan, and E. Steinbach. QoE-driven cross-layer optimization for high speed downlink packet access. *Journal of Communications*, 4(9):669–680, 2009.

[117] S. Thakolsri, W. Kellerer, and E. Steinbach. Application-driven cross layer optimization for wireless networks using MOS-based utility functions. In *International Conference on Communications and Networking in China (ChinaCOM)*, Xi An, China, Aug. 2009.

[118] M. Li, Z. Chen, and Y. Tan. Scalable video transmission over multiuser MIMO-OFDM systems. In *International Conference on Communications and Networking in China (ChinaCOM)*, Beijing, China, Aug. 2010.

[119] M. Li, Z. Chen, and Y. Tan. Joint packet prioritization and QoS mapping for SVC over wlans. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, March. 2010.

[120] M. Li, Z. Chen, S. Chuah, and Y. Tan. Efficient packet scheduling for scalable video delivery to mobile clients. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, Paris, France, June. 2010.

[121] J. B. Martens and L. Meesters. Image dissimilarity. *Signal Processing*, 70, 1998.

[122] Z. Wang and A. C. Bovic. A universal image quality index. *IEEE Signal Processing Letters*, 9, 2002.

[123] Z. Wang, A. C. Bovic, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 2004.

[124] M. Pinson and S. Wolf. A New Standardized Method for Objectively Measuring Video Quality. *IEEE Transactions ON Broadcasting*, 50(3), 2004.

[125] S. Winkler and P. Mohandas. The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. *IEEE Transactions ON Broadcasting*, 54(3), 2008.