Optimization of Rateless Code Based Video Multicast

by

Ali Bakhshali

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Applied Science

Queen's University
Kingston, Ontario, Canada
December 2011

Abstract

Multimedia services have become one of the major demands in wireless systems. As a result of growing demands for media services, traffic in wireless networks are increasing. Hence, optimization of multimedia delivery systems to efficiently consume the valuable transmission resources in wireless networks has gained a lot of interest. Raptor codes, with linear encoding and decoding time complexity are one branch of fountain codes (also known as rateless codes) which have found their ways in many recent communication standards as application layer forward error correcting (FEC) codes. Various attempts have been made in order to adapt these codes to wireless channels with their time varying nature. When multimedia delivery is targeted, some other issues such as delay should also be considered. Moreover, in multicast solutions, the system has to address demands of multiple clients.

In this thesis, we investigate some optimization scenarios for wireless multimedia multicast systems wherein clients with heterogeneous channels and media quality demands subscribe to a video program. The video program is assumed as a multilayer source with possible spatial, temporal and fidelity layers. The point of optimization under various systems is to provide the clients of different quality constraints with their demanded services while imposing the minimum network/client cost (e.g. delay, power consumption, outage probability) or maximize the provided utility to the

clients while considering their heterogeneous capabilities. To achieve these tasks, we study outage probability which serves as a measure to quantify the reliability of a service in a client's side. Packetized rateless multimedia multicast (PRMM) with few optimization criteria regarding the experienced delay in clients are studied and analytical solutions are obtained. A new optimization framework for rateless multimedia multicast is proposed in which, the provided utility to heterogeneous clients are maximized with respect to the clients channel and their quality demands. Application of this optimization in a rateless multimedia multicast system wherein the utility is defined based on perceptual quality experience of clients is also investigated.

Acknowledgments

First of all, I would like to express my grateful thanks to my supervisor Prof. Wai-Yip Geoffrey Chan for his constructive comments and valuable suggestions through out this work. I would like to extend my gratitude to Prof. Steven Blostein, Dr. Yu Cao, Mr. Saeed Akhavan Astaneh and Mr. Maziar Sanjabi for their precious feedbacks regards to the materials presented in this thesis as well as all the students in the multimedia coding and communications laboratory (MC^2L) for the thoughtful discussions. I also thank Prof. Fady Alajaji and Prof. Tamás Linder for serving in my thesis committee.

I also acknowledge Queen's university for the financial supports I received while completing my degree.

I am really thankful to my friends for their kindness, and continuous supports and encouragements during the past two years, specially Mr. Ali Moallem, Mr. Shervin Shahidi, Mr. Behnam Behinaein and many others who were beside me in my joys and sorrows. My life in Kingston could not have been this much exciting without them.

My special thanks goes to my parents Mohammad and Fatemeh and my sister Leili, whom I am indebted forever. Their love, understanding and support are the main reasons behind my success in every stage of my life.

Dedicated to my father, Mohammad Bakhshali (1951-2010), who was not alive to see my graduation. May his soul be happy of my achievements.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP 3rd Generation Partnership Program

AP Access Point

BEC Binary Erasure Channel

BER Bit Error Rate

CSI Channel State Information

EEP Equal Error Protection

EWF Expanding Window Fountain

FEC Forward Error Correcting

HD High Definition

IPTV Internet Protocol Television

LDPC Low Density Parity Check

LT Luby Transform

MDS Minimum Distane Separable

MEC Memoryless Erasure Channel

MMBS Multimedia Broadcast/Multicast Services

MMSE Minimum Mean Square Error

MOS Mean Opinion Score

MSE Mean Square Error

MTCO Multimedia Transcoder & Channel Optimizer

PC Pearson Correlation

PDA Personal Digital Assistant

PET Priority Encoding Transmission

PLR Packet Loss Rate

PRMM Packetized Rateless Multimedia Multicast

PSNR Peak Signal to Noise Ratio

QoS Quality of Service

RC Reception Coefficient

RS Reed-Solomon

RSD Robust Soliton Distribution

SCM Simplified Channel Model

SD Standard Definition

SVC Scalable Video Coding

UEP Unequal Error Protection

ULP Unequal Loss Protection

VoD Video on Demand

VQEG Video Quality Experts Group

WLAN Wireless Local Area Network

WWAN Wireless Wide Area Network

Chapter 1

Introduction

1.1 Motivation

Demand for better multimedia services is growing extensively. Each year, advancement in device technology enables the implementation of algorithms with higher computational complexity in both media servers and clients. This enables communication systems to introduce new services to the clients, which were not practically possible before. Displaying media streams such as live TV has nowadays become one of the most basic characteristics of smart phones, personal digital assistants (PDA) and portable computers. As a direct result of these advancements, new varieties of video-on-demand (VoD) networks and Internet-protocol-television (IPTV) based systems are commercialized every year. These services increase the network traffic. Hence, service providers are continuously increasing their network capacity as well as optimizing their transmission schemes.

Multicast solutions are widely used in the design of multimedia delivery systems. In contrast to their unicast counterpart, the multicast solutions utilize transmission

resources more efficiently. However, incorporating multicast solutions into multimedia delivery is a challenging task. A major challenge is to design a system in which clients with heterogeneous channel conditions and quality demands are addressed efficiently. This becomes even more challenging in wireless systems since the channel condition has a time varying nature. This requires adaptive optimization of transmission resources with respect to the channel state information (CSI) of the clients.

To address different quality demands, scalable media coders such as scalable video extension of H.264 [1] and Microsoft embedded audio coder (EAC) [2] can be utilized to encode the media contents. Scalable media coding encapsulates various adaptations of a processed media into a single stream. This stream can be decoded partially or entirely by different clients according to their quality demand. Thus, scalable media coding combined with multicasting significantly reduces the transmission bandwidth and enables progressive reconstruction of the media at the clients. For instance, the bitstream generated by a scalable media coder contains various slices of data, referred to as layers. Each layer —built on top of the preceding layers— yields additional spatial or temporal resolution or enhances the quality of the reconstructed media within the spatial-temporal resolution of the preceding layer. Therefore, higher layers can only take part in the reconstruction process provided that all the preceded layers are successfully decoded.

To guarantee a robust reconstruction process at the clients, proper forward error correcting (FEC) codes must be implemented prior to the transmission of the generated stream. FEC codes offer immunity against channel errors at the cost of adding redundant information. This reduces the good-put of transmission while increasing the reliability of the service. Conventional FEC codes, such as Reed-Solomon (RS) codes [3], are usually designed for specific loss rates. Hence, conventional FEC codes

that merely accommodate the client with the worst loss rate may not be efficient. In contrast to the conventional FEC codes, fountain codes [4] do not exhibit a fixed code rate and therefore can accommodate all the clients with various loss rates. Fountain codes are widely known as rateless codes since the code rate is not preselected in the transmitter. A fountain encoder generates a potentially infinite number of symbols from the input data. The generated symbols are such that the receiver can recover the input data from any subset of these symbols provided that its size is slightly larger than that of the original data. Raptor codes are the first known class of fountain codes with linear encoding and decoding time complexity. Shortly after being introduced, they found their ways into different standards in multimedia services such as DVB-H [5] and the third generation partnership program (3GPP) Multimedia Broadcast/Multicast Services (MBMS) [6]. They are used as an application layer coding scheme in order to provide packet level protection on top of the physical layer FEC codes.

Since in scalable video coding (SVC) multiple layers are decoded progressively, the lower layers of the transmitted source must be better protected against channel errors under any proposed channel coding scheme. This is referred to as unequal error protection (UEP). Different frameworks have been proposed to use conventional FEC codes for multimedia multicast [7–9]. Priority encoding transmission (PET) [7] is one of the first schemes with UEP property. Inspired by that, Mohr et al. [8] proposed a PET-packetization scheme for transmitting compressed images over noisy channels. Many approaches based on rate-distortion optimization, including Mohr's scheme, are studied in [9].

UEP design using rateless codes has been addressed in various studies over the past decade. One approach is to alter the degree distribution of the fountain codes

(see Section 2.3). For example, UEP is achieved by non-uniform selection of source symbols such that more important part of the source is more likely to be selected for transmission [10]. In [11], windows with expanding sizes are used such that the symbols that lie inside each window have equal probability to be selected. Windows are nested and expanding in size such that the smaller window is entirely included in the larger window. The idea of expanding window fountain (EWF) codes is combined with SVC in a multicast scenario in [12]. Choosing a proper degree distribution is perhaps the most important task in rateless code design. It is well known that the characteristics of rateless codes, e.g. small overhead, vary considerably with changes in the degree distribution. Therefore, modifications of the degree distribution must be carefully studied when such methods are used. Instead, some researchers aimed to achieve UEP by modifying the pre-code rates in a raptor code design [13]. However, this approach faces some practical concerns since dynamic modification of precode rates requires multiple encoders and decoders (in the transmitters and receivers, respectively), which may not be possible. Another approach is to design UEP schemes based on standardized fountain codes. Sheng et al. [14] proposed a new multicast transmission scheme based on Mohr et al.'s [8] PET-packetization and using rateless codes. The design allows heterogeneous clients to join and leave the multicast at random times. Cao et al. [15] proposed another multicast scheme where different media layers are encoded separately using fountain codes. The fountain symbols from different layers are optimally interleaved in order to minimize the transmission delay among various quality groups.

Although significant contributions have been made in the context of scalable video coding, fountain codes, and efficient multicast transmission schemes, there are few notable works that jointly employ these techniques to design multimedia delivery

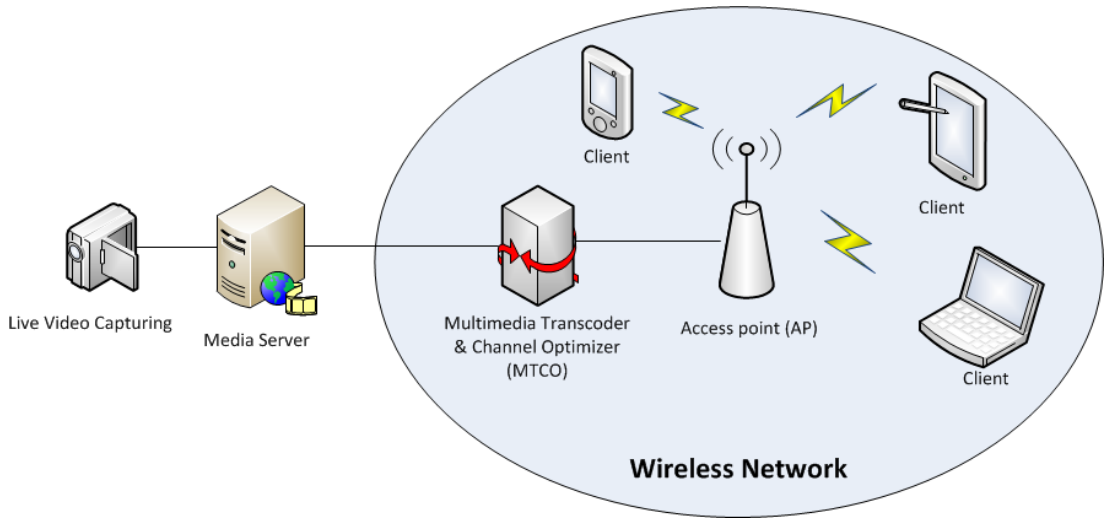


Figure 1.1: Multimedia multicast over a wireless network

systems wherein heterogeneity of the clients is properly addressed.

1.2 Multimedia Multicast in Wireless Networks

Fig. 1.1 illustrates a simple architecture for multimedia multicast implemented over a wireless network which can be a wireless local area network (WLAN) or a wireless wide area network (WWAN). A media server captures the stream of a video camera which is denoted as a sequence of group of pictures (GoP). Using a proper video encoding technology (e.g., H.264), the captured GoPs are encoded prior to transmission to clients and media relays in the network. SVC technology is highly suitable for multimedia multicast. SVC-capable encoders not only compress the video signal and reduce the imposed traffic to the network, but also yield the ability of parsing the stream to low quality streams without the need to decode and re-encode the bitstream

in media relays. Multimedia transcoder and channel optimizer (MTCO) is one of the relays that have subscribed to the bitstream of the media server. Typically, the media server is located outside the WLAN, e.g., Internet. Hence, the MTCO can be seen as a gateway connecting the outer network (Internet) to the WLAN. This gateway/relay is able to transcode the video into other formats and utilize proper packetization and channel coding schemes to adapt the video based on the constraints and demands of WLAN clients. MTCO generated packets are delivered to a wireless access point (AP), which is responsible to handle the required signaling between the MTCO and clients according to the utilized wireless transport protocol. Clients try to recover their preferred adaptation of the video program by decoding the packets that contribute to their demanded quality level.

The proposed wireless multimedia multicast architecture embeds various challenges and concerns. Clients that subscribe to the video multicast may have heterogeneous terminal capability. The terminal capability can indicate the resolution of the terminal display, the maximum power that the client terminal can allocate to video decoding task due to the internal power management policy, or a combination of both. Unlike wired networks, the underlying channel in WLAN is error prone due to fading and channel interference. Consequently, a client may not be able to correctly recover the transmitted packets of those timeslots wherein the corresponding channel was suffering from deep fades. This highlights the fact that the time varying nature of the clients's channel —that manifests itself as the burst errors— should be properly addressed by the physical layer, e.g., by using data interleavers and FEC coding applied in physical layer. In fact the main objective of an intelligently designed MTCO is to ensure that a specific goal is optimized by 1) analyzing the clients, their demands and their corresponding channels by the means of CSI and controlled feedbacks and

2) effectively use the the available tools to increase the performance. These tools can include but not limit to application layer FEC codes, various packetization structures and proper choice of the transmission rates from different layers of the multilayer bitstream. In fact the main challenge is to find a suitable optimization goal according to the network structure, it's limiting factors, and important aspects of the service. Having a proper optimization goal in hand, the best operating point can be obtained accordingly. One strategy for operating point selection is to maximize the received video quality for the receiver with the worst channel condition. However, with such a strategy, other receivers with better channel conditions unnecessarily experience a lower quality. Another strategy is to select an operating point that maximizes the average received video quality across all receivers. This strategy on the other hand ignores the individual fairness among receivers, e.g., some receivers may be overly penalized. It is crucial to design an operating point selection strategy that can achieve a high overall system performance, while guaranteeing fairness among receivers. Timing is an important aspect of video multicast. Unlike data communication, the video packets are prioritized in time, meaning that their significance is highly related to their arrival time. Failure to recover enough packets of certain quality layers during the transmission of current GoP can result in service outage. Outage or quality degradation of the service can lead to client dissatisfaction. Therefore, monitoring and controlling the outage occurrences is crucial in order to supply a reliable service. As a result, reliability factor should be integrated into the optimization scheme as a client constraint and/or a part of the optimization goal. Because of the time varying nature of the clients' channel in wireless networks, and since clients may join or leave the multicast continuously, any optimization has to be repeatedly performed

according to the most recent state of the network. Therefore, there is a great tendency towards optimization scenarios with low computational complexity that can yield solutions with close-to-optimal performances.

In this thesis we consider few multimedia multicast optimization scenarios for wireless networks similar to Fig. 1.1. The multimedia multicast is implemented on top of the underlying physical layer of the network. Throughout this thesis, the term "server" refers to MTCO which is in charge of optimizing the transmission resources in wireless multimedia multicast. Furthermore, we assume that the MTCO and the actual media server are connected using an error-free communication link. Hence, the reliability of multimedia delivery is only a function of the multicast scheme and the embedded randomness in clients channel and rateless codes. In Chapter 3, the outage probability —which is a measure of the service reliability— is introduced and proper ways of monitoring that in server are investigated. Rateless codes are utilized as the application layer FEC codes in various multicast schemes discussed in this thesis. In Chapter 4, a recently proposed rateless multimedia multicast scheme with the ability to serve heterogeneous clients is described and few optimization goals regarding the experienced delay by clients is addressed. Clients are assumed to have strict quality demands that must be perfectly addressed. In Chapter 5 another multicast scheme is proposed with the intention to maximize the provided utility (good-put) to the multicast crowd. In scalable video coding, the encoder can embed layers of different spatial, temporal and granular resolutions in the output bitstream. However, various combination of these layers may affect networks differently under certain assumptions on client terminal constraints and server bandwidth. Using the proposed scheme In Chapter 5, we conduct a research to study whether a SVC structure that incorporates various scalability layers can lead to any significant increase in the perceived quality

experience compared to when only a single type of scalability is considered. Various scenarios regarding this study are constructed and optimized in Chapter 6.

1.3 Thesis Contributions

The contributions of this thesis to address some of the problems in multimedia multicast are as follows:

- Because of random behavior of error events in wireless channels, the delivered service has a probabilistic nature. The quality of experience is not only measured by the quality of the reconstructed media but also by the continuity of the service over a large time period. Outage probability is a measure of reliability of a service over time. This measure is not only a function of the client channel but also depends on the characteristics of the fountain encoding/decoding process in rateless systems. In Chapter 3, we investigate outage probability in the broad context of graph codes while focusing on fountain codes. We derive a novel exact closed-form solution for the outage probability, which leads to formulations with lower computational complexity.

- Furthermore, we propose a novel model that can explicitly relate the total number of the transmitted symbols to the outage probability of clients in memoryless erasure channels and vise versa. The proposed model is highly accurate and has a simple structure. Therefore, this model is highly suitable for cross-layer optimization.

- A multimedia multicast scheme proposed by Sheng et al. [14] is briefly discussed in Chapter 4. Within this framework, we formulate the problem for packetized

rateless multimedia multicast with min-max regret and min-max delay criteria. Regret coefficients reflect the cost of the additional delay that users in each quality class suffer as a result of joining the multicast compared to the multiple unicast systems. In min-max regret optimization, we aim to minimize the cost of these additional delays (denoted as regret coefficients) encountered by heterogeneous clients. The min-max delay criterion aims to minimize the maximum amount of delay induced. By reformulating this problems, we derived convex optimizations. Furthermore, analytical solutions for these problems are also obtained. The results are compared to the result of the average delay minimization scheme proposed in [16].

- Heterogeneous users with various channel condition and quality demand may join the multicast. The server may not be able to provide all the users with their highest quality demand due to the bandwidth limitation. Some users might be denied any level of service because of their extremely poor channel while some other clients might be assigned to a level of service lower than their demands. Previous works have focused on resource allocation in systems wherein users are not flexible with their demands. In Chapter 5, we demonstrate a novel resource allocation optimization that dynamically assigns users to various quality layers of the media based on their channel quality and terminal capability in order to maximize the utility of the entire multicast. An advantageous feature of the proposed scheme is that its complexity is independent of the number of clients. Furthermore, we obtain a convex formulation by some simplifying assumptions and approximations in exchange for negligible inefficiency in utilizing the transmission resources.

- Scalable video coding (SVC) aims to encode the video into a bitstream of multiple layers. These layers can be a combination of spatial, temporal, or granularity layers built upon each other. The bitstream can be obtained by configuring the encoder to embed only different layers of one type of scalability (Single-scalable video coding) or to include layers of two or all three types of scalability in the video (Multi-scalable video coding). In Chapter 6, we investigate the potential superiority of multi-scalable video coding compared to single scalable approach in delivering the maximum perceptual quality experience among clients using the proposed rateless multimedia multicast in Chapter 5. Results in this chapter confirms the benefits of multi-scalable video coding under certain conditions. This example application also reveals the power of the proposed optimization in considering different types of source scalability as well as heterogeneity in both clients' terminal capability and channel quality in maximizing the perceived quality experience.

Chapter 2

Preliminaries

2.1 Introduction

In this chapter we provide a brief introduction on some of the main concepts referred to throughout this thesis. Most of the contributions in this thesis are in the context of applications of fountain codes in multimedia delivery in which packet erasure channels are frequently considered in cross-layer design. Therefore, most of this chapter is devoted to introducing these concepts as well as a general overview of erasure channel codes including conventional Reed-Solomon codes as well as the fountain codes. Moreover, video quality measurement is also discussed briefly in this chapter.

2.2 Binary Erasure Channel

The binary erasure channel (BEC) is one of the simplest non-trivial channel models imaginable. According to Fig. 2.1, it can only transmit zero and one as its input, and the output is either the same as the input or erased. The emergence of the Internet

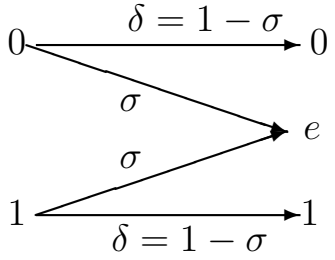


Figure 2.1: Binary erasure channel (BEC) with erasure probability σ.

promoted the erasure channel into the class of real-world channels. Buffer overflow in networks and exceedingly weak received signals in wireless systems can be modeled as erasure events. The BEC channel is fully defined by erasure probability σ. The probability of receiving each input bit without erasure at the output is $\delta = 1 - \sigma$. δ is denoted as reception coefficient and is widely used throughout this thesis. The capacity of the BEC is shown to be [17]

$$\mathrm{C_{BEC}} = 1 - \sigma = \delta \qquad \text{bits/channel use} \; . \tag{2.1}$$

Considering transmission of packets over an erasure channel, we can generalize the concept of BEC to packet erasure channel (PEC) in which packets are independently erased by probability σ_P —also denoted as packet loss rate (PLR). It is well understood that σ_P is a function of FEC codes applied in the physical layer. Since packets are initially processed in the physical layer by means of appropriate FEC coding, the FEC decoder in the physical layer can indicate whether the erroneous bits in the packet are correctable or the packet should be declared as corrupted or lost. σ_P can be decreased by using FEC codes with lower code rates.

2.3 Erasure Channel Codes

Automatic repeat request (ARQ) is a conventional error-control method in data communication that helps to establish reliable data transmission over an unreliable channel. In this method, the client has to inform the server whether previously sent packets are correctly recovered or not. This client-to-server communication utilizes a separate communication link, known as feedback channel. The feedback channel is assumed to be perfect in the sense that no erasure or error event can take place in it. The receiver might send back messages to acknowledge each received packet or request some specific missing packets to be retransmitted by the server. Neglecting the problem of establishing a perfect feedback channel and the cost it entails, retransmission protocols have the advantage of theoretically providing reliable communication provided the erasure probability σ is less than unity. However, if σ is large, the number of feedback messages grows extensively. The deficiency of retransmission protocols is especially evident in the case of broadcast channels. In these channels, the sender broadcasts a set of packets to many receivers, and each receiver receives a random fraction of the packets. Retransmission leads to high redundancy since retransmitted packets which might be useful for one receiver may be correctly received by other clients. Furthermore, when the available reception bandwidth of the server is limited, a large amount of feedback messages can cause feedback implosion.

According to Shannon [18], one should be able to exploit the capacity of the forward erasure channel, whether or not a feedback channel exists. This motivation led to design of different families of erasure codes in the past 60 years.

2.3.1 Reed-Solomon Codes

The classic block codes for erasure correction are Reed-Solomon (RS) codes [3]. An (N, K) Reed-Solomon code over a Galois field of size $q = 2^l$, $GF(2^l)$, has the ideal property that the original K source symbols can be recovered if any K of the N transmitted symbols are received. By appending $T = N - K$ check symbols to the K source symbols, a RS code can detect any combination of up to T erroneous symbols, and correct up to $\lfloor T/2 \rfloor$ symbols. However, RS codes are not practical for large K and N. Standard implementations of encoding and decoding have a cost of $O(K(N-K)\log_2 N)$ symbol operations. Furthermore, with a Reed-Solomon code as with any block code, one must estimate the erasure probability σ and choose the proper code rate $R = K/N$ before transmission. A designed code for erasure probability σ_1 is useless when $\sigma > \sigma_1$.

2.3.2 Fountain Codes

Digital fountain codes are an important family of sparse-graph codes for channels with erasures. In coding theory, fountain codes (also known as rateless codes) are a class of erasure codes with the property that a potentially limitless sequence of encoding symbols can be generated from a given set of source symbols. The original source symbols can ideally be recovered from any subset of the encoded symbols of a size slightly larger than the number of source symbols [19]. As indicated by its name, a rateless code does not have a fixed rate. The rate is determined on the fly by the time at which the receiver decodes the message.

LT (Luby transform) codes [20] are the first practical realization of fountain codes which were invented by Luby in 1998. An erasure code is optimal if the original K

source symbols can be recovered from any K encoding symbols. In fact, LT code and its successors are near optimal since the receiver needs to receive $K(1+\varepsilon)$ symbols to recover the source symbols. $\varepsilon > 0$ denotes the reception overhead which is less than 5% in well designed codes. The overhead is decreased as K is increased. These codes are universal because they are near-optimal for every erasure channel. i.e., regardless of channel erasure error, these codes can be implemented and achieve an information transmission rate close to the channel capacity in (2.1).

Raptor codes [21] are an important family of fountain codes obtained by precoding the information symbols prior to the application of Luby transform. The precode can be itself a concatenated code. Various precoding schemes with combinations of low density parity check (LDPC) codes [22] and Hamming codes have been proposed and standardized. Similar to LT codes, Raptor codes were also shown to be capacity achieving for the BEC channel. Raptor are also known for having a linear-time encoding and decoding complexity [23]. Hence, unlike RS codes, raptor codes are realizable for large K.

LT coding is the common and the most important stage in many fountain coding schemes. Therefore, we briefly describe the encoding and decoding process in this section. Further details and discussions can be obtained in [20].

LT Encoding

The following steps describe how the fountain symbols are generated using Luby transform [20].

1. We first randomly choose the degree D of the encoding symbol according to the chosen degree distribution $\Omega(x)$.

2. Then we choose a random set of D distinct input symbols uniformly from K input symbols. These symbols are "'neighbors"' of the encoding symbol.

3. The value of the encoding symbol is simply the XOR of the values of the D neighbors.

As mentioned before, the degree distribution $\Omega(x)$ plays a crucial role in the code performance. Code overhead is mainly a function of this distribution. Different distributions such as *ideal soliton distribution* and *robust soliton distribution (RSD)* [20] have been proposed. One can represent the encoding process as a bipartite graph, with K vertices on one side of the graph representing input symbols, and one vertex for each encoding symbol on the other side. There is an edge from each encoding symbol to each one of its D neighboring input symbols. Each encoding symbol, which in practice can consist of many encoded bits, each coding a different disjoint chunk of the source bits, is loaded into a separate packet. Each packet contains the value of the encoded symbol. Note that each packet contains $\lceil D \log_2 K \rceil$ bits for the neighbor list (also known as side information). However, many minor modification to this process have been proposed to minimize the amount of side information in these codes. For example, authors in [24] proposed *limited randomness LT coding* which incorporates a specific type of random number generator that reduces the amount of the side information to $\lceil 2 \log_2 K \rceil$ bits per each encoded symbol. We can show that if the codebooks of encoder and decoder are synchronized —i.e., the the encoder generates the encoded symbols using a pseudo-random number generator and the decoder is informed about the seed in advance— then, the amount of the side information is reduced to $\lceil \log_2 N \rceil$ bits per symbol in which N is the number of different transmitted packets considered in the codebook.

LT Decoding

The decoder constructs a bipartite graph using the received encoding symbols, from which it will try to recover the input symbols which are the neighbors of the encoding symbols in the graph. Let the set of received encoding symbols which have not been fully processed be $\mathcal{C}$ (initially $\mathcal{C}$ contains all M packets). We repeatedly choose an encoding symbol c_i, $i = 1, \ldots, |\mathcal{C}|$ from $\mathcal{C}$ and do the following, until none of the c_i's in $\mathcal{C}$ introduces any deletion of edges or recovery of symbols, or until all input symbols are recovered.

1. If c_i no longer has any unrecovered neighbors, remove it from $\mathcal{C}$.

2. If c_i is adjacent to exactly one input symbol we say that input symbol is covered by c_i. We easily recover that input symbol, since its value is the same as c_i. Now c_i has finished being processed, and we remove it from $\mathcal{C}$.

3. Otherwise consider each neighbor of c_i in turn. If it has already been recovered, then XOR its value into the value of the c_i and remove the edge connecting that neighbor to c_i. Since c_i still has unrecovered neighbors, it remains in $\mathcal{C}$.

The decoding process is successful if we can recover all of the input symbols from the M encoded symbols in hand. Otherwise, the decoding process has failed. We may wait to receive more encoded symbols, add them to $\mathcal{C}$, and continue the decoding process. By making M large, the probability of decoding success approaches 1.

2.4 Video Quality Measures

A video signal can be represented as a series of still images, also known as frames. When the frames are being exposed to human eyes rapidly the small differences

among adjacent frames create the illusion of moving objects and scenes in the human brain. The human perceptual effect of a video viewing is complex to model. Some quality measures only take into account the pixel amplitude distortions of a degraded video signal with respect to the clean original video, such as the mean square error (MSE) and peak signal to noise ratio (PSNR) measures. Unfortunately these measures do not accurately reflect the perceptual quality of the video. Moreover, they are not even effective when comparing video signals of different spatial and temporal resolutions. Researchers have investigated more accurate model that predict the subjective quality of video signals distorted by compression schemes, and of video signals with different spatial, temporal and granularity resolutions. The output of the proposed models are compared with the results of standardized subjective quality measurements on human response. In subjective experiments, a panel of viewers are asked to rate the quality of a set of video clips. The quality by itself is an ill-defined concept being interpreted differently among individuals due to personal interests and expectations. Subjective experiments attempt to minimize these factors through precise instructions, training, and a controlled environment. The average rating over all viewers for a given clip is also known as the mean opinion score (MOS). Although MOS obtained from subjective experiments provides the most accurate video quality ratings, subjective experiments are not practical for in-service real-time monitoring applications. Such applications would rely on objective quality metrics, in despite of the fact that they may be less accurate than subjective measurement. An objective metric that demonstrates high correlation with MOS is known to be useful in video quality assessment. Different models such a SSIM [25], VQM [26], PEVQ [27] have been proposed for video quality measurement. Some metrics are being standardized in the course of the multimedia test phase in ITU-T Rec. J.247

by Video Quality Experts Group (VQEG) [28].

There are different ways that metrics can be applied to measure quality of a video under test, namely, in the presence of the reference video clip and without the reference clip. A single-ended method is used when there is no reference video clip available whereas a double-ended method additionally utilizes the reference clip in the process of calculating an objective score. The main challenge in single-ended quality measurement is how to detect and assess distortions in the test video. The human brain can be viewed as possessing a model which is able to detect and assess distortions in the video content. The double-ended method itself can be applied in two ways, using the reference clip or using only features extracted from the reference clip —also known as reduced reference in quality measurement.

Bit error rate (BER) and packet loss rate (PLR) are simple metrics used to quantify transmission errors. These metrics help to realize the reliability of a communication links within the network. However, they are only relevant for data links, where every bit and packet is treated as equally important, but not for video delivery. The main reason is that these metrics are designed to characterize data fidelity without taking into account the content, i.e., the visual importance of different packets (or different bits) are not considered in data-link reliability metrics. Depending on the coding and the nature of the transmitted packets, the same number of lost packets can have different perceptual effects on the video content. Nevertheless, QoS metrics such as PSNR have been widely used in the context of video quality measurement, mainly, due to the simplicity of the MSE based optimizations and the lack of objective quality measurement standards in the past.

New objective metrics are being developed considering different factors. The main approach is to obtain hybrid metrics such that different aspects of quality are assessed

and pooled to estimate the overall quality [29]. Hybrid metrics can take into account the compression method as well as predicting the effects of distortions on the regions of the video which are more important according to the human visual system. The readers are encouraged to refer to [29] for further details and explanations.

Chapter 3

Outage Probability

3.1 Introduction

In this chapter we formulate outage probability as a function of channel erasure rate, total number of transmitted symbols and code performance. After explaining difficulties in direct calculation of outage probability in Section 3.2, we describe a couple of models that approximate the outage probability in Section 3.3. Without introducing any approximation regards to the original formulation of outage probability, a simple closed-form equation for outage probability is obtained in Section 3.4. Although the closed-form formulation has a lower complexity compared to original formulation, there is still a need for models with low structural complexity in which different parameters can be expressed as an explicit function of other variables. These models are highly suitable for cross-layer optimizations. To contribute in this area, a novel model with high accuracy and yet negligible computational complexity is presented in Section 3.5.

3.2 Outage Probability of Fountain Codes

Probability of failure in decoding a code of length S after receiving M symbols is represented by $P_f(M, S)$. For minimum-distance separable (MDS) codes with hard decision decoders, $P_f(M, S)$ is a binary function that returns 1 when $M < S$ and 0 otherwise. Therefore, there is no randomness in the state of decoding for any number of distinct recovered symbols for such codes. However, many branches of codes are not MDS and for many commonly used non-MDS codes (specially sparse graph codes, e.g., fountain and LDPC codes) with belief propagation decoder, the state of decoding may not be fully deterministic. In the context of rateless codes, Luby et al. [30] demonstrated that we may approximate $P_f(M, S)$ using,

$$P_f(M,S) = \begin{cases} 1 & \text{if } M \leq S \\ a \times b^{M-S} & \text{if } M > S \end{cases} \quad . \tag{3.1}$$

Here, S is the number of encoded information symbols and M represents the number of correctly received symbols in receiver. $0 < a$ and $0 < b < 1$ are model parameters and may vary for different codes. For example, $a = 0.85$ and $b = 0.567$ are derived for the rateless codes utilized in [30].

Outage probability P_{out} is the probability that a receiver fails to decode a code of length S, given a certain amount of symbols N has been transmitted from the server. It is clear that not only the number of transmitted symbols, but also characteristics of the clients' channel affects the outage probability.

$$P_{out} = \mathbb{E}_{M,\delta}[P_f(M,S)] \quad , \tag{3.2}$$

$0 \leq \delta \leq 1$ is the channel transmission coefficient which is the probability that a

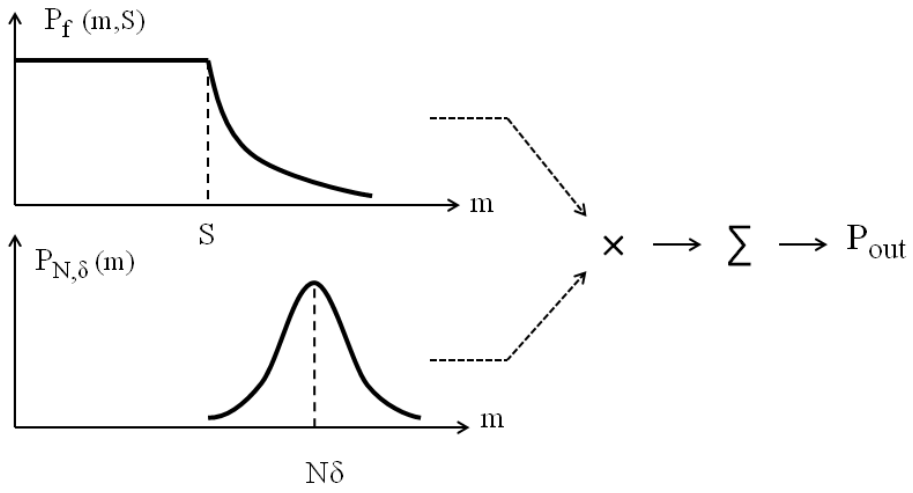


Figure 3.1: Outage probability as a function of code performance and channel quality

transmitted symbol is not erased. Throughout this thesis, erasure channel may also be characterized by its erasure rate σ which is related to δ according to $\sigma = 1 - \delta$. $\mathbb{E}_{M,\delta}[.]$ is the expectation operator with regards to the channel transmission coefficient δ and the number of correctly recovered symbols M. By assuming that the variations of a client channel during the transmission period is negligible and the erasures are independent and identically distributed (i.i.d.), we conclude that the number of received symbols M has binomial distribution. Therefore, the outage probability can be expressed by

$$P_{out} = \mathbb{E}_M[P_f(M,S)] = \sum_{m=1}^{N} P_{N,\delta}(m) \times P_f(m,S) \quad . \tag{3.3}$$

N is the total number of transmitted packets and $P_{N,\delta}(m)$ is the probability mass

function (PMF) of binomial distribution with parameters N and δ.

$$P_{N,\delta}(m) = \binom{N}{m}\delta^m(1-\delta)^{N-m} \tag{3.4}$$

A graphical representation of outage probability is illustrated in Fig. 3.1. Substituting $P_f(M,S)$ from (3.1) into (3.3) yields:

$$\begin{aligned} P_{out} &= \sum_{m=1}^{S}\binom{N}{m}\delta^m(1-\delta)^{N-m} + \sum_{m=S+1}^{N}\binom{N}{m}\delta^m(1-\delta)^{N-m}\times ab^{m-S} \\ &= F_{N,\delta}(S) + \sum_{m=S+1}^{N}\binom{N}{m}\delta^m(1-\delta)^{N-m}ab^{m-S}\ . \end{aligned} \tag{3.5}$$

$F_{N,\delta}(.)$ is cumulative distribution function (CDF) of binomial distribution with parameters N and δ. In spite of fairly simple analytical derivation of outage probability, calculation of that using (3.5) can be a computationally complex task, specially for large S and N. This requires calculation of a large summation with the complexity of $O(N^2)$ multiplications. Moreover, terms in the summation tend to decrease exponentially after some point and taking the summation over these terms can lead to relatively high errors, specially when the actual outage probability is very small. Throughout this chapter, we discuss previous attempts to approximate outage probability as well as our own contributions in this area.

3.3 Previous Approximations of Outage Probability

One approach to simplify the outage probability is to approximate the memoryless erasure channel (MEC) with a less complicated model. Simplified channel model

(SCM) described in [15] is a good candidate for the task in hand. In SCM, some of the effects of erasure events on distribution of received symbols are neglected. Therefore, only variations in the expected number of received symbols as a function of the channel transmission coefficient δ is considered, i.e.,

$$P_{out} = \mathbb{E}_{M,\delta}[P_f(M,S)] \approx P_f(\mathbb{E}_{N,\delta}[M],S) \quad . \tag{3.6}$$

The received symbols have a binomial distribution with mean $\mathbb{E}_{N,\delta}[M] = N\delta$. Hence, the approximated outage probability is obtained from

$$P_{out}^{sim} = P_f(\mathbb{E}_{N,\delta}[M],S) = P_e(N\delta,S) = \begin{cases} 1 & \text{if } N \leq \frac{S}{\delta} \\ ab^{N\times\delta - S} & \text{if } N > \frac{S}{\delta} \end{cases} \quad . \tag{3.7}$$

The accuracy of this model —which is more often denoted as simple model throughout this thesis— highly depends on δ. When δ is large, the PMF of the binomial distribution is concentrated around $N\delta$. By decreasing the channel transmission coefficient δ PMF is broadened and yields lower magnitudes . Consequently, this model fails to provide an accurate estimation of outage probability for channels with high erasure rates since the approximation in (3.6) is not accurate anymore. Nevertheless, this model is highly suitable for cross-layer design since the proposed function is a convex function of N for $N > S/\delta$, its derivation are easy to obtain, and different variables can be expressed as an explicit function of others by simple rearrangements.

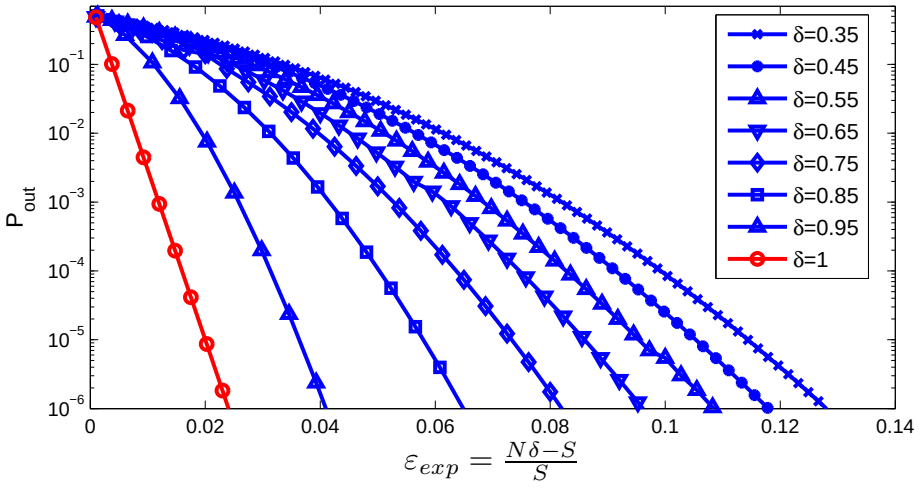
In [15], normal approximation, also known as DeMoivre–Laplace theorem [31], is utilized in order to obtain an accurate model for outage probability with less computational complexity compared to the direct form (3.5). According to equation (7)

in [15] and by considering $h = \min(S + 8, N)$, their approximation of outage probability P_{out}^{Cao} is

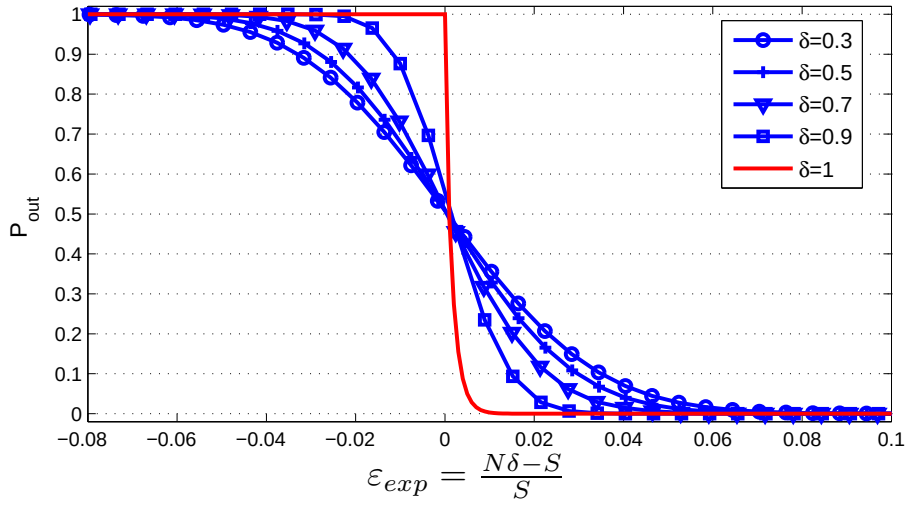
$$1 - P_{out}^{Cao} \approx \sum_{x=S}^{h} (1 - ab^{x-S}) \frac{1}{\sqrt{2\pi N\delta(1-\delta)}} \exp \frac{-(x - N\delta)^2}{2N\delta(1-\delta)} + (1 - ab^8)(Q\left(\frac{h - \delta N}{\sqrt{N\delta(1-\delta)}}\right) - Q\left(\frac{N(1-\delta)}{\sqrt{N\delta(1-\delta)}}\right)) \quad , \quad (3.8)$$

wherein $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{x^2}{2}) dx$. For $S+8 < N$, only $\min(9, N-S+1)$ terms take part in the proposed summation and calculation of Q function in two relatively close points are needed. It should be mentioned that accurate calculation of $Q(x)$ function has its own problems and concerns for large x. However, tight approximations of $Q(x)$ function with negligible complexity is available through different works such as [32, 33]. On the other hand, one may found the Cao et al's approximation very hard to utilize in scenarios wherein there are constraints defined over N, e.g., the proposed optimization in Chapter 5 of this thesis. The reason is that N can not be expressed as an explicit function of outage probability and other parameters according to this model.

Fig. 3.2 illustrates the outage probability as a function of expected fractional overhead, $\varepsilon_{exp} = \frac{N\delta - S}{S}$, when original equation for outage probability (3.5) is utilized. ε_{exp} represents the expected reception overhead (after accounting for expected amount of erasures) required for a receiver with channel δ in order to decode the fountain code with outage probability P_{out}. A rateless code with $a = 0.85$ and $b = 0.567$ is assumed in this figure. For receivers with high quality channels ($\delta \approx 1$), corresponding outage probability curves converge to $P_f(.)$ which is an encoder-decoder pair characteristic. Outage probability curves of different channels approximately



(a) Logarithmic scale



(b) Linear scale

Figure 3.2: Outage probability as a function of expected overhead for different channel transmission for $S = 1000$

intercept at $(\varepsilon_{exp}, P_{out}) = (0, 0.5)$ provided that the decay parameter b is not very large ($b < 0.7$ usually meets this criterion). This can be be easily proven by approximating $P_f(.)$ with a binary function when the failure probability has a small transition band. Furthermore, It is easy to prove that *every* outage probability curve with any transmission coefficient δ obtained from P_{out}^{sim} in (3.7) can be represented by the curve associated to $\delta = 1$ in Fig. 3.2. A careful look at (3.7) reveals the fact that the outage probability is in fact a function of $N\delta$. On the other hand, since this approximation accurately matches the original equation for outage probability (3.5) for $\delta = 1$, any difference between outage curves of any δ with the $\delta = 1$ illustrates the error induced by approximating MEC with SCM. This error becomes more significant as δ is decreased.

3.4 Closed-form Representation of Outage Probability

By deeper investigations on (3.5)—which serves as the main equation for outage probability when MEC is considered— we can obtain a closed-form equation with much less computational complexity. Starting from (3.5), the outage probability can be modified to

$$\begin{aligned} P_{out} &= F_{N,\delta}(S) + \sum_{m=S+1}^{N} \binom{N}{m} \delta^m (1-\delta)^{N-m} a b^{m-S} \\ &= F_{N,\delta}(S) + ab^{-S}(1-\delta)^N \sum_{m=S+1}^{N} \binom{N}{m} \left(\frac{\delta b}{1-\delta}\right)^m . \end{aligned} \tag{3.9}$$

Now, we seek for a hypothetical parameter $\hat{\delta}$ such that

$$\frac{\delta b}{1-\delta} = \frac{\hat{\delta}}{1-\hat{\delta}} \quad .$$

For any $0 \leq \delta \leq 1$, this parameter exists and can be obtained from

$$\hat{\delta} = \frac{\delta b}{1-\delta+\delta b} \quad . \tag{3.10}$$

Using $\hat{\delta}$ we step further in finding a closed-form solution for P_{out} by

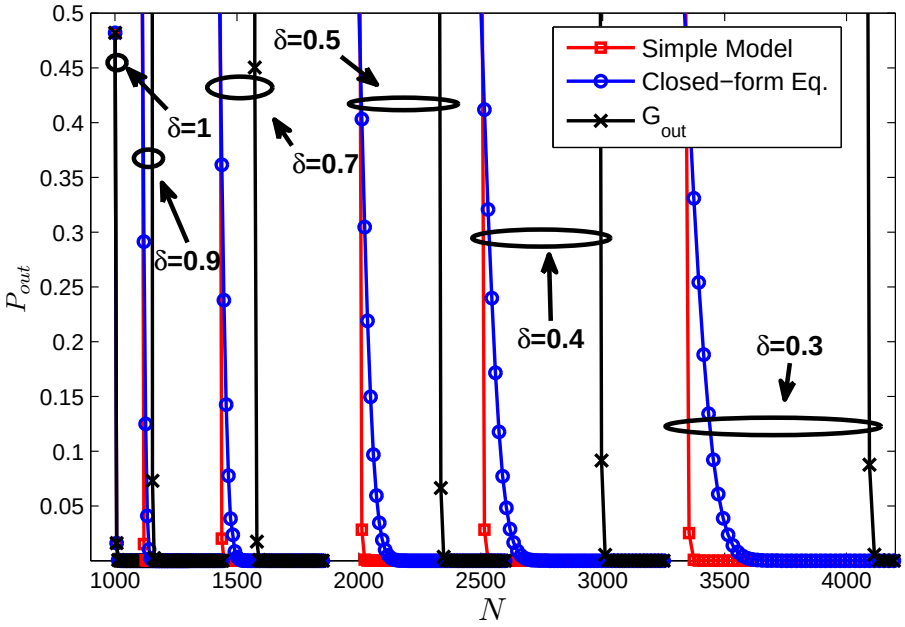
$$\begin{aligned} P_{out} &= F_{N,\delta}(S) + ab^{-S}\left(\frac{1-\delta}{1-\hat{\delta}}\right)^{N} \sum_{m=S+1}^{N} \binom{N}{m} \hat{\delta}^{m}(1-\hat{\delta})^{N-m} \\ &= F_{N,\delta}(S) + G_{out} \times \left(1 - \sum_{m=0}^{S} \binom{N}{m} \hat{\delta}^{m}(1-\hat{\delta})^{N-m}\right) \\ &= F_{N,\delta}(S) + G_{out} \times \left(1 - F_{N,\hat{\delta}}(S)\right) \quad . \end{aligned} \tag{3.11}$$

G_{out} is a function of different parameters according to

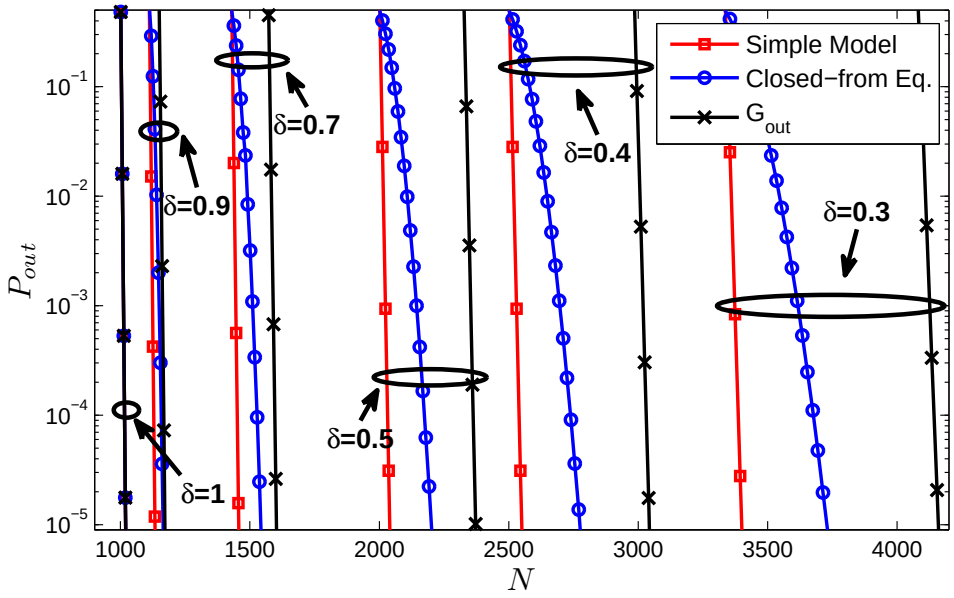
$$G_{out} = ab^{-S}\left(\frac{1-\delta}{1-\hat{\delta}}\right)^{N} = ab^{-S}(1-\delta+\delta b)^{N} \quad . \tag{3.12}$$

Using the closed-form equation (3.11) to obtain the outage probability does not require calculation of any explicit summation. Comparing (3.11) with (3.5), the overall computational complexity of calculating outage probability is reduced from $O(N^2)$ to few multiplications if we use the normal approximation to calculate the binomial CDFs.

Fig. 3.3 presents outage curves associated to simple model (3.7) and accurate model (3.11) for a raptor code of length $S = 1000$. Both methods yield similar results for $P_{out} \approx 0.5$. However, when outage probability is small, simple model returns an



(a) Linear scale



(b) Logarithmic scale

Figure 3.3: Outage probability for (blue) accurate model (3.11), (red) simple model (3.7) and (green) G_{out} (3.12) vs. transmitted bits for $S = 1000$

optimistic estimate for the outage probability. The error associated to simple model is increased as the designated outage probability and channel transmission coefficient are decreased. According to Fig. 3.3, G_{out} is a loose upper-bound for the outage probability.

3.5 Synthetic Model to Approximate Outage Probability

Although the proposed closed-form for outage probability in Section 3.4 is beneficial, it may not be very suitable for cross-layer optimizations wherein structurally simple functions are needed to model the relationship between different parameters. In fact, one of the reasons that the simple model (3.7) is used in cross-layer design is the ability of this model to express different variables as an explicit function of others. For example, the system might need to obtain optimal transmitted symbols N which guarantees an outage probability no more than $\overline{P_{out}}$. Binary search is required to solve for the optimal N when either (3.8) or (3.11) is utilized, while a rough estimate can be easily obtained from (3.7). Simultaneously simple and accurate models are beneficial in utilizing outage probability curves easier and faster in optimization tasks.

By investigating outage probability for various channel conditions and code lengths, we obtained the following function which approximates the outage probability with high accuracy for $10^{-8} < P_{out} < 0.5$.

$$\widehat{P}_{out} = 0.5 \times e^{-A \times (N-B)^H} \quad for \quad N \geq B \tag{3.13}$$

A, B and C are directly obtained from number of information symbols S, total number

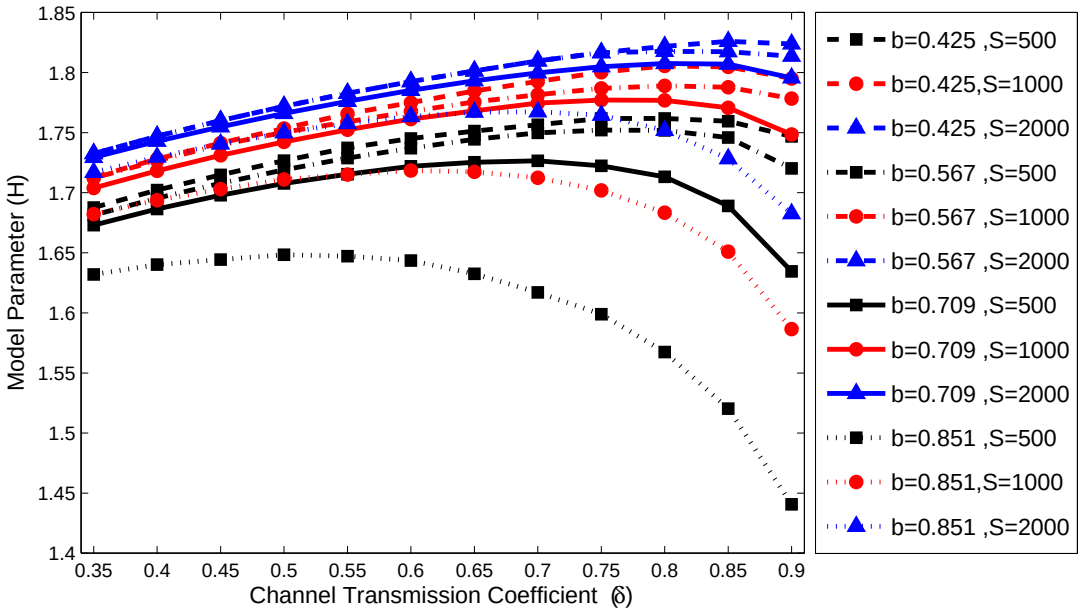


Figure 3.4: H as function of b, S and δ

of transmitted symbols N, channel transmission coefficient δ and fountain code decay factor b (through $t = -\ln(b) > 0$).

$$\begin{cases} A = \frac{\delta}{(1-\delta)S} \\ B = \frac{S}{\delta} \end{cases} \tag{3.14}$$

Most of the operational communication systems aim to deliver services with low outage probability to the clients. Hence, $P_{out} < 0.5$ is barely a limiting factor in real applications.

Optimal H is chosen based on minimum min-square-error (MMSE) criterion. Dependency of H to other variables are illustrated in Fig. 3.4. Although H is a function of code length S, channel coefficient δ and decay factor b, dependency of H to b is not significant when $b < 0.75$. This condition holds for the raptor codes in [30] .

Consequently, the optimal parameter H can be assumed to vary within the range $H \in [1.67\ 1.83]$. Hence we can model H using

$$H = 1.69 + \delta \times [0.125 + 0.04 \log_2(S/1000)] \quad , \qquad b < 0.75. \tag{3.15}$$

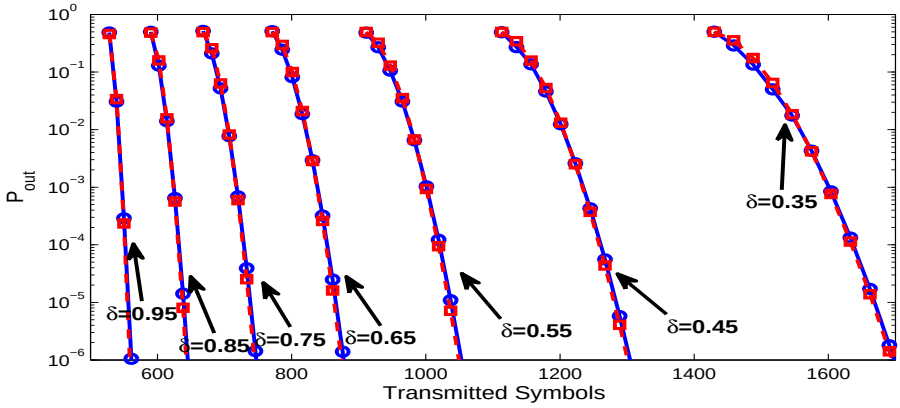
We realized that by considering $\delta = 0.8$ in (3.15), the resulting H still provides us with an accurate approximation of the outage probability for a wide range of δ. For example, Fig. 3.5 illustrates outage probability curves obtained by this function for a raptor code with $a = 0.850$ and $b = 0.567$. The curves associated to the proposed synthetic function perfectly follows the outage probability curves under different parameters settings. Table 5.2 presents the accuracy of the three models described in this report using Pearson correlation (PC) coefficient and MSE. PC between $\mathbf{x} = [x_1, ...x_L]$ and $\mathbf{y} = [y_1, ...y_L]$ can be obtained from

$$\text{PC} = \frac{\sum_{i=1}^{L}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{L}(x_i - \bar{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{L}(y_i - \bar{\mathbf{y}})^2}} \tag{3.16}$$

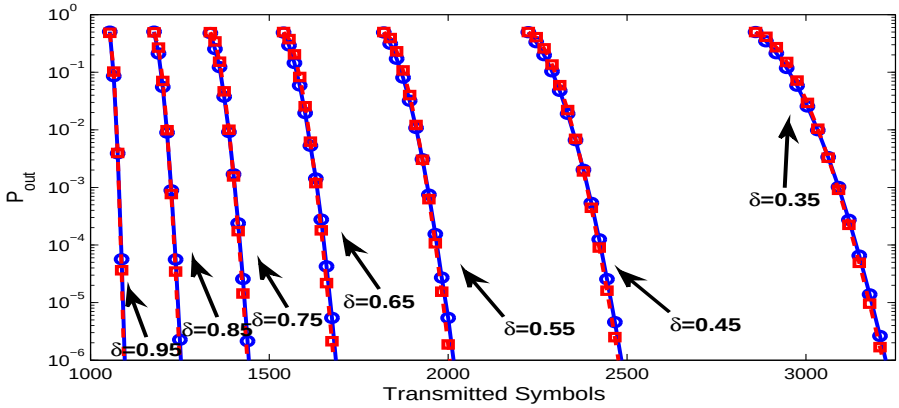
Here, $\bar{\mathbf{x}} = \text{MEAN}(\mathbf{x})$ and $\bar{\mathbf{y}} = \text{MEAN}(\mathbf{y})$. $|\text{PC}| = 1$ when $\mathbf{x}$ and $\mathbf{x}$ are perfectly correlated.

In order to demonstrate the robustness of H with respect to variations of channel parameter δ, the effect of δ on optimal H is neglected by considering $A = 0.5$ and $H = 1.79 + 0.032 \log_2(S/1000)$. Accuracy of different models and approximations are compared to the exact outage probability obtained from (3.11) in the region $10^{-8} \leq P_{out} \leq 0.5$.

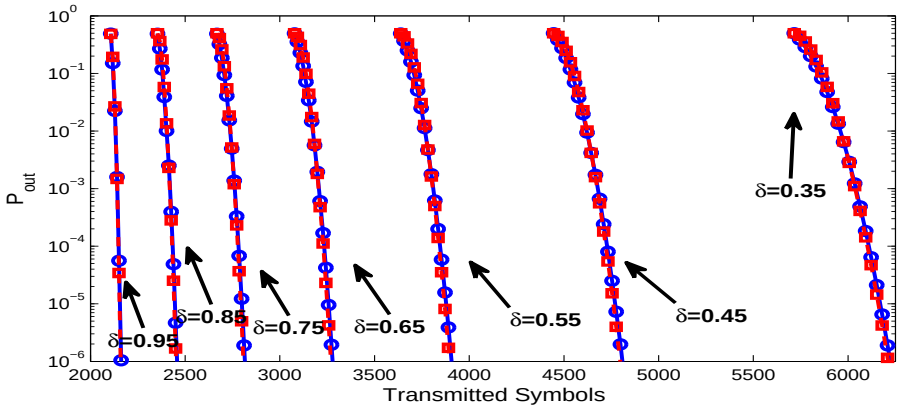
The simple model (3.7) presents very low correlation factor compared to both synthetic model and Cao et al.'s approximation (3.8). Moreover, MSE of the simple



(a) S=500



(b) S=1000



(c) S=2000

Figure 3.5: Comparison of (circles) outage probability and (squares) estimated outage probabilities using synthetic model

Table 3.1: Accuracy of different methods in estimating the outage probability

S	δ	PC			MSE ($\times 10^{-3}$)		
		Simple Model	Cao's App. [15]	Synthetic Model	Simple Model	Cao's App. [15]	Synthetic Model
500	0.3	0.754	1.000	0.963	15.58	0.10	1.27
	0.4	0.774	1.000	0.981	14.99	0.09	0.65
	0.5	0.787	0.999	0.984	15.56	0.08	0.59
	0.6	0.823	1.000	0.993	15.83	0.07	0.25
	0.7	0.872	0.999	0.996	14.07	0.08	0.17
	0.8	0.917	0.999	0.997	13.44	0.11	0.15
	0.9	0.972	0.995	0.995	8.80	0.23	0.10
1,000	0.3	0.658	1.000	0.972	17.91	0.09	1.04
	0.4	0.678	1.000	0.988	17.98	0.09	0.51
	0.5	0.704	1.000	0.995	17.65	0.10	0.39
	0.6	0.737	1.000	0.998	15.73	0.10	0.47
	0.7	0.790	1.000	0.999	14.38	0.09	0.70
	0.8	0.840	1.000	0.999	16.64	0.08	0.32
	0.9	0.921	0.998	0.999	11.69	0.11	0.22
2,000	0.3	0.566	1.000	0.979	20.24	0.09	0.87
	0.4	0.586	1.000	0.992	19.80	0.09	0.44
	0.5	0.610	1.000	0.997	19.03	0.09	0.43
	0.6	0.642	1.000	0.998	18.26	0.09	0.65
	0.7	0.659	1.000	0.992	16.69	0.10	0.92
	0.8	0.752	1.000	0.997	16.58	0.10	1.32
	0.9	0.758	0.998	0.996	14.91	0.11	0.53
4,000	0.3	0.484	1.000	0.984	21.03	0.08	0.70
	0.4	0.501	1.000	0.994	21.33	0.08	0.41
	0.5	0.523	1.000	0.997	20.77	0.08	0.52
	0.6	0.551	1.000	0.997	20.31	0.08	0.86
	0.7	0.575	1.000	0.991	19.59	0.09	1.25
	0.8	0.655	1.000	0.994	18.87	0.09	1.71
	0.9	0.710	1.000	0.985	16.38	0.11	1.82
Average		**0.707**	**1.000**	**0.991**	**16.93**	**0.09**	**0.69**

model is more than two orders of magnitude higher than the most accurate approximation which is the one proposed by Cao et al. This is another witness of the poor quality of the simple model. On the other hand, the proposed synthetic model yields PC = 0.99 among different experiments. High accuracy, negligible complexity and simple structure of the proposed model suggest it as a good candidate for optimization tasks.

3.6 Conclusions

In this chapter, we described the outage probability which is an important QoS metric in multimedia delivery systems. A closed-from equation for outage probability is obtained with much less computational complexity compared to the primary equation. We proposed a new synthetic model that accurately approximates the outage probability for a wide range of channel transmission coefficient, code length and number of transmitted symbols. The complexity of the new model is in the order of few multiplications which is negligible when compared to the complexity of $O(N^2)$ in direct calculation. Various Applications for the proposed closed-form for outage probability (3.11) and synthetic model (3.13) can be considered. Simple yet accurate models can bring on huge computational efficiency while avoiding complications in formulating high level cross-layer optimization problems. For instance, scenarios in [14, 34] involve multiple constraints on clients' outage probability. Proposed algorithms in those works are based on direct calculation of outage probability using (3.5). Complexity of the algorithms can be reduced using the closed-form outage probability or the proposed synthetic model in this chapter.

Chapter 4

Resource Allocation Optimization in Packetized Rateless Multimedia Multicast

4.1 Introduction

In this chapter we investigate few optimization problems for "packetized rateless multimedia multicast (PRMM)" [14]. In PRMM, limited transmission resources of a multimedia server is managed such that clients of heterogeneous service requirement and channel quality are satisfied with their desired level of quality. Transmitted media is considered as a multilayer source, e.g., a H.264/SVC video. The encoded media is transmitted using fixed-length packets. The server controls the flow of data from different source layers in order to optimize some network constraint, e.g., average delay in clients. Rateless coding is applied as a channel FEC code and clients' channels

are assumed to be packet erasure channels (PEC) with different erasure probabilities. In the original description of PRMM, the clients were assumed to be able to join the multicast asynchronously within a finite window after that the transmission has begun, while still being guaranteed to achieve their target QoS assurance levels. However, in this thesis we only consider a system in which clients are synchronized with the server.

In general, assigning clients to various media layers and/or different QoS can be a function of different parameters. In best-effort optimizations the system aims to maximize a revenue or minimize a cost function by providing clients with service levels which may or may not be their demanded ones. However, in QoS-guaranteed scenarios, the system is required to provide clients with their demanded services. Therefore, under the limitations on transmission resources and channel quality of participating clients, a possible solution may or may not exist. One way to address this problem is not serving users that place too much demand on the system, an approach we examine in the next chapter. When the feasible set—which contains possible transmission settings that guarantee the QoS requirements of all clients—is non-empty, various cost criteria can be employed to select the best solution. For example in wireless networks, the system may seek a solution which minimizes the power consumption or reception delay among clients.

In this chapter, problem formulations of few QoS-guaranteed scenarios in PRMM is obtained while in the next chapter a problem formulation for a best-effort optimization is studied. We derive the PRMM optimization under minmax-regret criterion in which the goal is to minimize the maximum additional delay that users in various classes experience as a result of server sharing the transmission resources among various user classes. Next, we study the optimization of transmission resources of same

system under minmax-delay criterion where the aim is to minimize the maximum reception delay experienced by clients in different quality classes. For the proposed optimizations, the analytical solutions are obtained and verified using a convex optimization software.

4.2 PRMM System Design

Fig. 4.1 illustrates the structures of PRMM system. At the beginning of each transmission period, portion of a multilayer media of J layers (e.g. 16 frames of a H.264/SVC scalable bitstream) is loaded in to a PET structure with L layers. This results in L layers that each contains $K_l, l = 1, ..., L$ source symbols. Layers of this PET structure are separately encoded using fountain codes. During every transmission timeslot, the server generates one symbol from every digital fountain and puts all these symbols into a single packet of length L. When total number of N_{max} packets are transmitted, the current media is replaced with the bitstream corresponding to the next portion of the media and the process is continued. Clients are assumed to have heterogeneous channels and quality demands.

In the proposed PET structure, source symbol stream is partitioned into L vertical layers - also known as packet layers - such that symbols in layers with lower indices are more important than those in layers with higher indices. Design objective in PET packetization is to ensure unequal error protection (UEP) among symbols in bitstream, i.e., more important symbols are more protected against packet loss events in channel. As a result, the first layer with the least depth contains bitstream of the most fundamental layer of the media and last layer with the highest depth is assigned

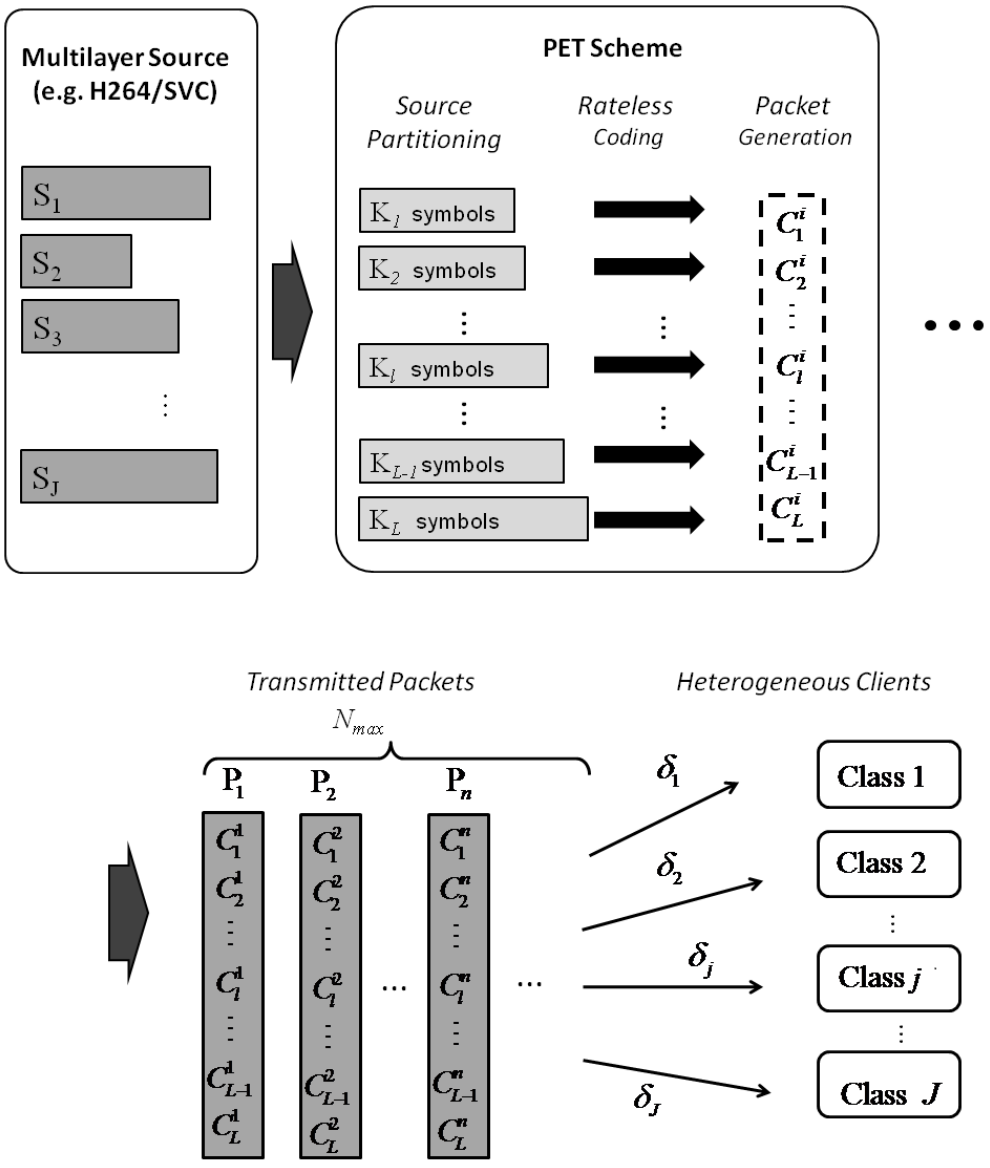


Figure 4.1: Structure of packetized rateless multimedia multicast (PRMM)

to the bitstream of the highest enhancement layer which is used only by clients with the highest quality demand. This objective leads to non-decreasing number of symbols in layers, i.e,

$$K_1 \leq K_2 \leq \cdots \leq K_L \ . \tag{4.1}$$

4.3 Minmax Regret Optimization

4.3.1 Derivation of General formulation

Various fidelity measures can be used to differentiate levels of service according to the transmitted media. Without loss of generality, we use PSNR which is a commonly used quality metric for video. QoS requirement of a user subscribed to class j of source stream who needs to achieve a PSNR greater than or equal to threshold γ_j can be expressed as

$$PSNR_j(K_1, K_2, ..., K_L) \geq \gamma_j \ . \tag{4.2}$$

Here, L is the total number of packet layers and $PSNR_j(K_1, K_2, ..., K_L)$ denotes the PSNR of class j users for a given allocation set $\{K_1, K_2, ..., K_L\}$. Given a non-decreasing function $f(.)$ that represents PSNR value for transmitted video as a function of total number of source symbols, h_j denotes the minimum number of packet layers required for a class j client to achieve the desired PSNR. This can be expressed as

$$\sum_{l=1}^{h_j} K_l \geq f^{-1}(\gamma_j) \ \ , \tag{4.3}$$

where $f^{-1}(.)$ denotes the inverse function of $f(.)$.

Because of non-decreasing number of symbols per layers, the average required transmitted packets for a class j user comes from average number of required packets to recover layer h_j alone and not on layers $l < h_j$. We consider i.i.d. packet erasure events in clients' channels. Therefore, for a user with channel transmission coefficient δ_j (or equivalently erasure rate $\sigma_j = 1 - \delta_j$), the average number of packets should be transmitted by server, N_j, that leads to class j clients satisfaction is given by

$$N_j = \frac{K_{h_j}(1+\epsilon_j)}{\delta_j} \ . \tag{4.4}$$

ϵ_j is the fractional overhead for the rateless code of h_j-th packet layer. Ceiling operator is omitted for analytical simplicity. The fractional overhead is a function of the desired outage probability P_j, code length K_{h_j}, and rateless code characteristic. For example, using the simple outage probability model P_{out}^{sim} described in (3.7), the fractional overheadϵ_j can be obtained from

$$\epsilon_j = \frac{\ln(P_j/a)}{K_{h_j} \ln b} \qquad (P_j \leq a) \ . \tag{4.5}$$

Here, $0 < a, b < 1$ are rateless code parameters described in Chapter 3.3. For well designed rateless codes and in typical applications with $10^{-5} < P_j < 0.01$, $\epsilon_j < 5\%$ when $K_{h_j} > 200$. Channel transmission coefficient of class j users can be modeled by a probability density function (PDF) $f_{\Delta_j}(\delta_j)$. Therefore, the expected number of transmitted packets required for class j users is

$$\begin{aligned} E_{\Delta_j}[N_j] &= K_{h_j}(1+\epsilon) E_{\Delta_j}[1/\delta_j] \\ &= K_{h_j}(1+\epsilon) \int_0^1 \frac{f_{\Delta_j}(x)}{x} dx \\ &= K_{h_j}(1+\epsilon)/\widehat{\delta_j} \end{aligned} \tag{4.6}$$

$E_X[.]$ denotes expectation operation according to random variable X. $\widehat{\delta_j}$ is the effective transmission coefficient for class j users that can be calculated from $f_{\Delta_j}(\delta_j)$.

$$1/\widehat{\delta_j} = \int_0^1 \frac{f_{\Delta_j}(x)}{x} dx$$

Now assume another scenario wherein the server only considers the QoS requirements of the class j users. Therefore, all L layers are filled with the data required for class j users which corresponds to source layers $S_i, i = 1, .., j$. $\mathcal{K}^j = \{K_1^j, K_1^j, ..., K_L^j\}$ denotes the packetization structure for this scheme. Since all packet layers are filled with source symbols required for class j clients, $h_j = L$. Using (4.6) the average number of required packets for class j users in this scenario is denoted by

$$\begin{aligned} E_{\Delta_j}[\hat{N}_j] &= K_L^j(1+\epsilon)E_{\Delta_j}[1/\delta_j] \\ &= K_L^j(1+\epsilon)/\widehat{\delta_j} \end{aligned} \tag{4.7}$$

The difference between the number of the transmitted packets for class j clients in both scenarios defines the "regret" for class j users, i.e.,

$$\begin{aligned} T_j &= w_j(E_{\Delta_j}[N_j] - E_{\Delta_j}[\hat{N}_j]) \\ &= (w_j(1+\epsilon)/\widehat{\delta_j}) \times (K_{h_j} - K_L^j) \\ &= \eta_j(K_{h_j} - K_L^j) \end{aligned} \tag{4.8}$$

Here,

$$\eta_j = w_j(1+\epsilon)/\hat{\delta_j} = w_j(1+\epsilon)\int_0^1 \frac{f_{\Delta_j}(x)}{x} dx$$

and $w_j \geq 0$ are regret weights which help us to adjust the regret penalty associated to different classes. By using (4.6) and (4.8) along with the UEP condition in (4.1),

we obtain the problem formulation for the optimization of PRMM based on minmax-regret criterion.

Problem 4.1:

$$\min_{t,K_1,K_2,...,K_L} t$$

subject to

$$\textit{Minmax Rule :} \quad \eta_j(K_{h_j} - K_L^j) \leq t \;\; j = 1, 2, ..., J$$

$$\textit{UEP Const :} \qquad K_1 \leq K_2 \leq \cdots \leq K_L$$

According to Problem 4.1, the aim is to minimize the maximum of weighted additional delay penalties of different quality classes as a result of moving to multicast scheme from unicast.

4.3.2 Transformation to Convex Optimization

We show that we can obtain a convex optimization by modifying Problem 4.1. As the first step, we introduce the following lemma to continue.

Lemma 4.1: *Among all allocations of S symbols to L_s layers, the optimal allocation that minimizes the layer depth (and hence the associated reception delay) is equally allocating symbols among layers, i.e $K_i = \frac{S}{L_s}, \forall i \in \{1, 2, ..., L_s\}$.*

Proof. Let us consider an arbitrary set $\mathcal{K} = \{K_1, K_2, ..., K_{L_s}\}$ that sums up to S. For any given set of numbers, e.g. $\mathcal{K}$, we know that MEAN($\mathcal{K}$) $\leq$ MAX($\mathcal{K}$) with equality *if and only if* $K_1 = K_2 = ... = K_{L_s}$. Since the reception delay is proportional to MAX($\mathcal{K}$) and MEAN($\mathcal{K}$) $= \frac{\sum_{i=1}^{L_s} K_i}{L_s} = S/L_s$, the minimum is achieved when equality

condition holds, i.e., $K_1 = K_2 = ... = K_{L_s} = \frac{S}{L_s}$. □

By using Lemma 4.1 and after omitting the integer constraints on $K_l, l = 1, ..., L$, we conclude that the optimum allocation that minimizes the partial transmission delay associated to layer j is equally allocating S_j symbols among $l_j = h_j - h_{j-1}$ layers

$$K_i = \frac{S_j}{l_j} \qquad i = h_{j-1} + 1, ..., h_j \text{ and } j = 1, ..., J \tag{4.9}$$

For a PSNR scalable source, S_j can be obtained from

$$S_j = \lceil f^{-1}(\gamma_j) \rceil - \lceil f^{-1}(\gamma_{j-1}) \rceil \qquad \forall j \in \{1, \ldots, J\}, \tag{4.10}$$

wherein we considered $\lceil f^{-1}(\gamma_0) \rceil = 0$. Using (4.9), K_{h_j} can be expressed as

$$K_{h_j} = \frac{S_j}{l_j}. \tag{4.11}$$

By definition, class j users require the bitstream up to layer j to achieve their target QoS. Therefore, the total number of required symbols for users in class j is $R_j = \sum_{k=1}^{j} S_k$. In unicast transmission when only QoS requirement of class j is considered, R_j symbols are equally allocated among all L layers according to Lemma 4.1. Hence K_L^j can be obtained from

$$K_L^j = \frac{R_j}{L}. \tag{4.12}$$

Substituting (4.11) and (4.12) in (4.8), the associated penalty for class j users is

$$\begin{aligned} T_j &= \eta_j(K_{h_j} - K_L^j) = \eta_j(S_j/l_j - R_j/L) \\ &= \alpha_j/l_j - \beta_j, \end{aligned} \tag{4.13}$$

where $\alpha_j = \eta_j S_j$ and $\beta_j = \eta_j R_j / L$ for $j = 1, ..., J$.

Using (4.9), UEP constraints in Problem 4.1 can be transformed into

$$0 \leq l_J/S_J \leq l_{J-1}/S_{J-1} \leq \cdots \leq l_1/S_1. \tag{4.14}$$

Finally, a new formulation for minmax regret optimization is obtained.

Problem 4.2:

$$\min_{\{t, l_1, l_2, ..., l_J\}} t$$

subject to

$$\begin{aligned} \textit{Minmax Rule :} \quad & \alpha_j/l_j - \beta_j \leq t \qquad j = 1, 2, ..., J \\ \textit{UEP Const. :} \quad & 0 \leq l_J/S_J \leq l_{J-1}/S_{J-1} \leq \cdots \leq l_1/S_1 \\ \textit{Packet-length Const. :} \quad & \textstyle\sum_{j=1}^{J} l_j = L \end{aligned}$$

Note that an additional constraint is introduced to the problem to ensure that the number of packet layers remains the same. The objective cost is an affine function and minmax rule constraints are convex in l_j for $l_j > 0$ and affine in t. Furthermore, UEP constraints and equality constraint are both linear. Therefore, Problem 4.2 is a convex optimization and can be solved using any convex optimization tool.

4.3.3 Analytical Solutions

We show that the solution for Problem 4.2 can be obtained analytically. By ignoring the UEP inequality constraints in Problem 4.2, we claim that the optimal solution of a similar problem wherein the UEP constraints are omitted yields the same regret

bound t^* among all user classes. i.e.,

$$\alpha_j/l_j^* - \beta_j = t^* \ . \qquad j = 1, 2, ..., J \tag{4.15}$$

Although the new problem still has other constraints, we refer to it as unconstrained problem throughout this chapter in the sense the UEP constraints are dropped The following lemma justifies the optimality of equal bound solution for the unconstrained problem.

Lemma 4.2: *Consider Problem 4.2 when the UEP constraints are dropped.* $\mathcal{L} = [l_1, l_2, ..., l_J]$ *is the vector of any possible allocation of packet layers to source layers that satisfies* $\sum_{j=1}^{J} l_j = L$. *We define regret bounds* $t_j = \alpha_j/l_j - \beta_j$ *for* $j = 1, ..., J$. $\mathcal{L}^* = [l_1^*, l_2^*, ..., l_J^*]$ *is the optimal solution for this problem if and only if all regret bounds are equal. i.e.* $t_i = t_j \ \forall i, j$.

Proof. Consider a layer allocation set of L available packet layers $\mathcal{L} = [l_1, l_2, ..., l_J]$ which yields unequal regret bounds. Let us assume that $\mathcal{L}$ is the optimal allocation. We gather layers of highest and lowest regret bounds in $\mathcal{L}_m$ and $\mathcal{L}_s$ respectively.

$$\begin{aligned} \mathcal{L}_m &= \arg\max_{l_j}\{t_j | j = 1, ..., J\} \\ \mathcal{L}_s &= \arg\min_{l_j}\{t_j | j = 1, ..., J\} \end{aligned}$$

As a direct result of assuming unequal bounds, $\mathcal{L}_m$ and $\mathcal{L}_s$ are two disjoint sets, each containing a minimum of one element. We can always find small quantities Δ_m and

Δ_s that satisfy $|\mathcal{L}_m|\Delta_m = |\mathcal{L}_s|\Delta_s$ and construct a new layer allocation scheme $\hat{\mathcal{L}}$:

$$\hat{l}_j = \begin{cases} l_j + \Delta_m & \text{if } l_j \in \mathcal{L}_m \\ l_j - \Delta_s & \text{if } l_j \in \mathcal{L}_s \\ l_j & \text{other} \end{cases} \qquad \forall j \tag{4.16}$$

It is easy to prove that $\sum_{j=1}^{J} \hat{l}_j = L$. Because the number of packet layers for source layers in $\mathcal{L}_m$ is increased, we may conclude that their corresponding regret bounds $\alpha_k/l_k - \beta_k = t_k,\ \ l_k \in \mathcal{L}_m$ are decreased. Since layers in $\mathcal{L}_m$ were associated to highest regrets bounds, the maximum regret bound in new allocation scheme is decreased. This implies that the newly constructed layer allocation scheme $\hat{\mathcal{L}}$ achieves a lower upper-bound and therefore, $\mathcal{L}$ could not be optimal. The result is based on the fact that Δ_m and Δ_s are sufficiently small that the new bounds for layers in $\mathcal{L}_s$ do not exceed the maximum bound in $\hat{\mathcal{L}}$ which is always achievable.

The proof for converse part is straight forward. Let us consider $\mathcal{L}^*$ as a layer allocation scheme that yields the equal regret regime. For any other layer allocation scheme, let's say $\mathcal{L}^A$, we define $d_j = l_j^* - l_j^A\ \forall j$. Because the sum of elements in both $\mathcal{L}^*$ and $\mathcal{L}^A$ is L, we have $\sum_{j=1}^{L} d_j = 0$. Consequently, there is at least one negative and one positive element in $d_j, j = 1, ..., J$, otherwise $\mathcal{L}^*$ and $\mathcal{L}^A$ are identical which is against the assumption that they are different. Now consider layer k in $\mathcal{L}^A$ with $d_k > 0$. This is equivalent to $l_k^A < l_k^*$. Therefore, the regret coefficient for layer k in $\mathcal{L}^A$ is higher than its counterpart in $\mathcal{L}^*$, i.e., $t_k^A > t_k^*$. Consequently the maximum regret for $\mathcal{L}^A$ is higher than $\mathcal{L}^*$ which implies that $\mathcal{L}^*$ is the optimal solution for the proposed problem. □

As a direct result from Lemma 4.2, the equal bound solution is applied, i.e.,

$$\alpha_1/l_1^* - \beta_1 = \alpha_j/l_j^* - \beta_j \qquad j = 1, ..., J. \tag{4.17}$$

Hence

$$l_j^* = \frac{\alpha_j l_1^*}{\alpha_1 + (\beta_j - \beta_1) l_1^*} \qquad j = 1, ..., J. \tag{4.18}$$

After applying the equality constraint in Problem (4.2) we have

$$\sum_{j=1}^{J} l_j^* \;\; = \;\; \sum_{j=1}^{J} \frac{\alpha_j l_1^*}{\alpha_1 + (\beta_j - \beta_1) l_1^*} = L. \tag{4.19}$$

l_1^* can be easily obtained by using recursion.

$$l_1^* = L \times \left(\sum_{j=1}^{J} \frac{\alpha_j}{\alpha_1 + (\beta_j - \beta_1) l_1^*} \right)^{-1} \tag{4.20}$$

The acceptable solution satisfies $l_1^* < \min\{\mathcal{D} \cup \{+\infty\}\}$ with

$$\mathcal{D} = \left\{ \frac{\alpha_1}{\beta_1 - \beta_j} |\ \beta_j - \beta_1 < 0, j = 1, ..., J \right\}. \tag{4.21}$$

Having l_1^* in hand, l_j^* for $j \geq 2$ can be obtained from (4.18). Next, the UEP constraints are verified. If none of those constraints is violated, the obtained solution is indeed the optimal solution for Problem 4.2.

Lemma 4.3: *Suppose the solutions of Problem 4.2 without considering the UEP constraints does not satisfy $l_{i-1}/S_{i-1} \leq l_i/S_i$ for a given i. Then the optimal allocation of packet layers to layer i and $i-1$ of source for Problem 4.2 with UEP constraints is such that $l_{i-1}/S_{i-1} = l_i/S_i$.*

Proof. When only one UEP constraint is violated, the proof is evident because otherwise, the equal bound solution in (4.17) should be optimal and hence comply with UEP conditions which is against hypothesis that one UEP constraint is violated. Therefore, the related UEP constraint is active and we have $l_{i-1}/S_{i-1} = l_i/S_i$ [1]. □

Using Lemma 4.3, if the allocation scheme violates the UEP constraint among two consecutive layers of media, l_i and l_{i+1}, i.e., $l_i/S_i < l_{i+1}/S_{i+1}$, one can consider $l_i = l_{i+1}S_i/S_{i+1}$ in equality constraint of Problem 4.2 and apply equal bound criterion in (4.17) without considering layer i,

$$\alpha_1/\hat{l}_1 - \beta_1 = \alpha_j/l_j - \beta_j \qquad j = 1, ..., J \text{ and } j \neq i.$$

Equivalently we have

$$\hat{l}_j = \frac{\alpha_j l_1}{\alpha_1 + (\beta_j - \beta_1)l_1} \qquad j = 1, ..., J \text{ and } j \neq i. \tag{4.22}$$

Considering $l_i = l_{i+1}S_i/S_{i+1} = (S_i/S_{i+1}) \times \frac{\alpha_{i+1}l_1}{\alpha_1+(\beta_{i+1}-\beta_1)l_1}$, summation over l_j yields

$$\sum_{\substack{j=1 \\ j\neq i,i+1}}^{J} \frac{\alpha_j \hat{l}_1}{\alpha_1 + (\beta_j - \beta_1)\hat{l}_1} + \frac{\alpha_{i+1}\hat{l}_1(1 + S_i/S_{i+1})}{\alpha_1 + (\beta_{i+1} - \beta_1)\hat{l}_1} = L. \tag{4.23}$$

By substituting $\alpha_j = S_j\eta_j$ we have

$$\sum_{\substack{j=1 \\ j\neq i,i+1}}^{J} \frac{S_j\eta_j \hat{l}_1}{S_1\eta_1 + (\beta_j - \beta_1)\hat{l}_1} + \frac{(S_i + S_{i+1})\eta_{i+1}\hat{l}_1}{S_1\eta_1 + (\beta_{i+1} - \beta_1)\hat{l}_1} = L. \tag{4.24}$$

The above equation suggests merging layer i and $i+1$ into a single layer with $S_i^{new} = S_i + S_{j+1}$, $\eta_i^{new} = \eta_{i+1}$, $\beta_i^{new} = \beta_{i+1}$ and decreasing the indices of layers with index

[1] For the case in which two or more UEP constraints are violated simultaneously, a complete proof has not been reached yet.

larger that i by one. Equivalently a new problem with $J^{new} = J - 1$ user classes is emerged from the original problem that yields the same optimal solution for any l_j which its corresponding source layer S_j is remained unchanged or just received a new index. The merging process is repeated until no violation of any UEP constraint in the newly constructed problem is observed. Using $\hat{x}$ to show parameter x in the new problem after merging process, the final problem can be expressed as:

Problem 4.3:

$$\min_{\{t,\hat{l}_1,\hat{l}_2,...,\hat{l}_{\hat{J}}\}} t$$

subject to

$$\begin{aligned} \textit{Minmax Rule :} \quad & \hat{\alpha}_j/\hat{l}_j - \hat{\beta}_j \leq t \qquad j = 1, 2, ..., \hat{J} \\ \textit{Packet-length Const. :} \quad & \textstyle\sum_{j=1}^{\hat{J}} \hat{l}_j = L \end{aligned}$$

The solution for Problem 4.3 is obtained after solving the following equation for $\hat{l}_1$ and using (4.22) to calculate $\hat{l}_j$, $j = 2, ..., \hat{J}$ from $\hat{l}_1$.

$$\sum_{j=1}^{\hat{J}} \frac{\hat{\alpha}_j \hat{l}_1}{\hat{\alpha}_1 + (\hat{\beta}_j - \hat{\beta}_1)\hat{l}_1} = L \tag{4.26}$$

In the next step, we obtain the optimal solution of Problem 4.2 which is the main point of interest. Let's consider layer $\hat{j}$ in Problem 4.3 that contains the symbols of layer j in Problem 4.2. Layer $\hat{j}$ can be layer j with a new index or a super layer as a result of merging layer j with its neighboring layers. The optimal allocated packet layers to source layer j can be obtained from

$$l_j = \frac{S_j}{\hat{S}_{\hat{j}}} \hat{l}_{\hat{j}}. \tag{4.27}$$

In another words, if the layer j of the source is not merged in to a new layer then $l_j = \hat{l}_{\hat{j}}$ and if otherwise, l_j receives a potion of $\hat{l}_{\hat{j}}$ such that all layers which are merged in to the layer $\hat{j}$ in the final problem achieve the same l_j/S_j (Hence they are equally protected).

4.4 Minmax Delay Optimization

Another design criterion in PRMM is minimizing the maximum of average delays that clients of various quality classes experience in the multicast. This criterion can be very useful because it gives the system operators an estimate of what would be a good transmission deadline according to the media and distribution of clients channels. Fortunately , the optimal solution for this criterion can be easily obtained using the formulations of min-max regret criterion. According to (4.8) the regret criterion aimed to minimize the maximum difference between the expected delay of class j clients —which is proportional to the total number of required packets to receive— in the multicast compared to unicast, wherein only the QoS requirements of class j users are considered. In Problem 4.2, α_j/l_j represents the average delay for class j clients in multicast scheme while β_j models the experienced delay for the same clients when the server has exclusively optimized the transmission for their convenience. After dropping β_j (i.e, $\beta_j = 0, \forall j$) in Problem 4.2, the new problem aims to optimally allocate available packet layers to various layers of media in order to minimize the maximum expected transmission delay among all user classes in multicast scheme.

Problem 4.4:

$$\min_{\{t,l_1,l_2,\ldots,l_J\}} t$$

subject to

$$\begin{aligned} \textit{Minmax rule :} \quad & \alpha_j/l_j \le t \qquad j = 1, 2, ..., J \qquad (4.28a) \\ \textit{UEP Const. :} \quad & 0 \le l_J/S_J \le l_{J-1}/S_{J-1} \le \cdots \le l_1/S_1 \\ \textit{Packet-length Const. :} \quad & \textstyle\sum_{j=1}^{J} l_j = L \end{aligned}$$

Similar to what we have done in previous section and by using Lemma 4.2, the optimal solution when UEP constraints are omitted is attained by an equal bound regime.

$$\alpha_1/l_1^* = \alpha_2/l_2^* = ... = \alpha_J/l_J^*$$

Considering $\alpha_j = \eta_j S_j$ we have

$$S_j/l_j^* = \eta_{j+1}/\eta_j \times S_{j+1}/l_{j+1}^* \qquad j = 1, ..., J-1. \qquad (4.29)$$

Now let's consider the UEP constraints

$$S_j/l_j^* \le S_{j+1}/l_{j+1}^* \qquad j = 1, ..., J-1. \qquad (4.30)$$

By comparing (4.29) and (4.30) we realize that the optimal solution of unconstrained problem complies with UEP constraint between class j and class $j+1$ when $\eta_{j+1}/\eta_j \le 1$. Otherwise, the solution is somewhere on the boundary of the region defined by UEP condition, i.e.,

$$\hat{l}_j = \begin{cases} \frac{S_j \eta_j}{S_{j+1}\eta_{j+1}} \hat{l}_{j+1} & \text{if } \eta_{j+1}/\eta_j < 1 \\ \frac{S_j}{S_{j+1}} \hat{l}_{j+1} & \text{if } \eta_{j+1}/\eta_j \ge 1 \end{cases} \quad \forall j \in \{1, ..., J-1\} \qquad (4.31)$$

$\hat{l}_j$, $\forall j$ are solutions of Problem 4.2 which includes all possible constraints. When the

UEP condition between any two consecutive class is active, we may reduce the dimensionality of this optimization problem by merging layers with active UEP boundary condition ($\eta_{j+1}/\eta_j > 1$) to a super layer with $S_j^{new} = S_j + S_{j+1}$, $\eta_j^{new} = \eta_{j+1}$, $J^{new} = J-1$ and decreasing the indices of all parameters with index larger than $j+1$ by one. This process is repeated until all remaining η_j are in increasing order. Using $\hat{x}$ to denote the parameters value after the merging process, the final problem can be expressed as :

Problem 4.5:

$$\min_{\{t,\hat{l}_1,\hat{l}_2,...,\hat{l}_{\hat{J}}\}} t$$

Subject to

$$\begin{aligned} \textit{Minmax Rule :} \quad & \hat{\eta}_j \hat{S}_j / \hat{l}_j \leq t \qquad j = 1, 2, ..., J \\ \textit{Packet-length Const. :} \quad & \textstyle\sum_{j=1}^{\hat{J}} \hat{l}_j = L \end{aligned}$$

The UEP constraints are dropped since they are automatically satisfied. Simply, by using (4.29) and applying the packet-length constraint, the optimal packet-layer allocation is obtained.

$$\hat{l}_j^* = \frac{\hat{\eta}_j \hat{S}_j}{\sum_{k=1}^{\hat{J}} \hat{\eta}_k \hat{S}_k}. \tag{4.33}$$

Similar to the minmax-regret optimization, solutions of Problem 4.4 and Problem 4.5 are related to each other according to

$$l_j = \frac{S_j}{\hat{S}_{\hat{j}}} \hat{l}_{\hat{j}}. \tag{4.34}$$

Here, $\hat{j}$ is a source layer in Problem 4.5 with $\hat{S}_{\hat{j}}$ source symbols which contains the

source layer j in Problem 4.4 with size S_j.

4.5 Simulation Results

Following scenario are conducted to investigate the optimality of proposed minmax optimizations. The server transmits a scalable video stream same as what is presented in Table I in [12]. The video sequence consists of one base layer (BL) and fourteen enhancement layers (EL) with total of 3800 symbols where each symbol represents 50 bytes. The total number of layers is $L_{max} = 100$ and overhead for rateless codes is assumed to be $\epsilon = 0.05$. In [16], the authors formulated a an optimization scheme based on PRMM. Given the same system constraints as what we presented in this chapter, their optimization goal was to minimize the expected reception delay amount clients which is expressed by $N_{av} = \sum_{j=1}^{J} \pi_j E_{\Delta_j}[\hat{N}_j]$. Here $E_{\Delta_j}[\hat{N}_j]$ is the expected delay for class j users discussed in (4.7) and $0 < \pi_j < 1$ are the prior weights with $\sum \pi_j = 1$.

Four user classes (J=4) are involved in the simulation. Their effective transmission coefficients $\hat{\delta}_j$, PSNR thresholds γ_j and weights w_j (and π_j in the case of average delay optimization) along with other intermediate variables are illustrated in Table 4.1. Solutions of the unconstrained problems for this multicast system based on the minmax-delay, minmax-regret, and average delay minimization criterion discussed in [16] are obtained after dropping the UEP constraints. The solutions of these problems are also depicted in Table 4.1. The solution of minmax regret criterion for the unconstrained problem complies with the UEP constraints by yielding non-decreasing number of symbols among layers. Therefore, the allocation scheme

Table 4.1: Solution of PRMM with dropped UEP constraints under various criteria

(a) Min-Max Regret

Layer	γ_j	S_j	$\hat{\delta}_j$	w_j	α_j	β_j	l_j	K_{h_j}
1	25.8	400	0.70	1	600.0	6.00	23.03	17.3
2	27.3	300	0.85	1	370.6	8.65	12.90	23.2
3	29.0	455	0.8	1	597.2	15.16	16.96	26.8
4	40.3	2645	0.95	1	2923.4	42.00	47.11	56.2

(b) Min-Max Delay

Layer	γ_j	S_j	$\hat{\delta}_j$	w_j	α_j	η_j	l_j	K_{h_j}
1	25.8	400	0.70	1	600.0	1.50	13.36	29.94
2	27.3	300	0.85	1	370.6	**1.24**	8.25	**36.36**
3	29.0	455	0.8	1	597.2	**1.31**	13.30	**34.22**
4	40.3	2645	0.95	1	2923.4	1.11	65.10	40.63

(c) Min Average Delay

Layer	γ_j	S_j	$\hat{\delta}_j$	π_j	α_j	$\frac{\sqrt{\alpha_j}}{S_j}$	l_j	K_{h_j}
1	25.8	400	0.7	1/4	150	**0.031**	20.03	**20.0**
2	27.3	300	0.85	1/4	92.6	**0.032**	15.75	**19.1**
3	29.0	455	0.8	1/4	149.3	0.027	20.00	22.8
4	40.3	2645	0.95	1/4	730.9	0.010	44.22	59.8

provided in Table 4.1(a) is the ultimate solution under minmax regret criterion .

The UEP constraints between the second and third layers of media are violated ($K_{h2} \not\leq K_{h3}$) under minmax delay criterion. This could be understood beforehand, without the need to solve for the solution, by noting $\eta_3/\eta_2 > 1$ and considering the optimality test provided in (4.31). This suggests that the UEP constraint between second and the third layers are active and those layers are essentially equally protected against erasure events in the optimal solution. Hence, we may merge these layers into a supper layer and construct a new problem similar to Problem 4.5. Also according

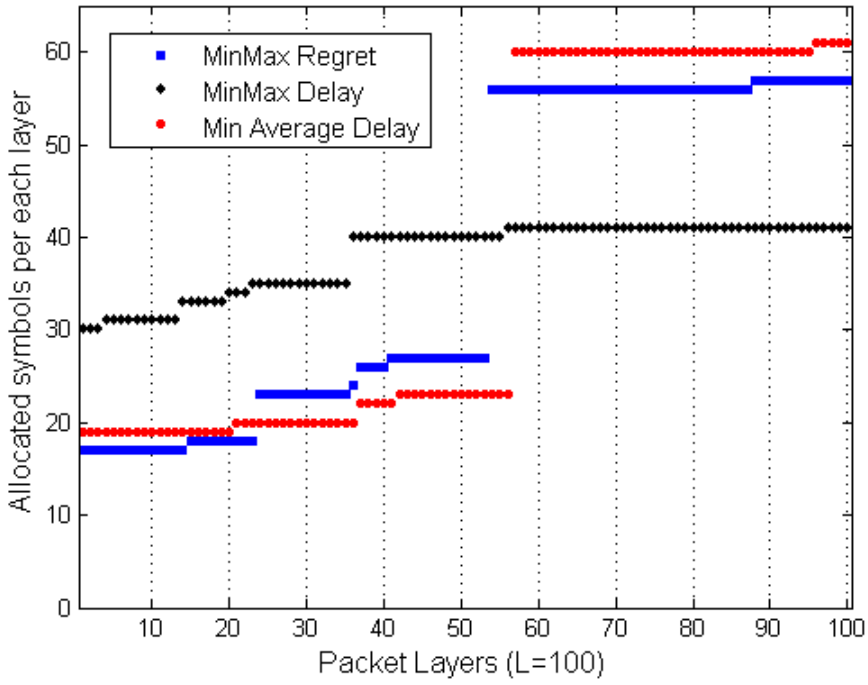


Figure 4.2: Allocation of source symbols among 100 layers in the packetized rateless multimedia multicast

to [16], since $\frac{\sqrt{\alpha_1}}{S_1} \leq \frac{\sqrt{\alpha_2}}{S_2}$ we could expect to observe that UEP constraint between first and second layers of the media to be violated in the unconstrained problem when average delay minimization criterion is considered. Hence, by using a procedure similar to the methods we derived in this chapter, we can merge the corresponding source layers and construct a reduced problem with a source of $J-1$ layers and $J-1$ classes according to Problem 1.2 in [16]. After following the required steps which involve 1) constructing the reduced problems, 2) obtaining the analytical solutions and 3) transforming the results of the reduced problems in to the original problems, the final solutions are derived and depicted in Table 4.2. Different optimization costs are monitored in the solutions of different problems. As we would expect, the optimization cost related to each of the three investigated criteria takes the minimum

in its respective optimization problem (which is indicated by underline). Figure 4.2 illustrates the allocation of symbols to available packet layers under various criteria. Solutions of the minmax regret and average delay criteria demonstrate very similar symbol allocation patterns. This similarity is also prominent by comparing their performances in Table 4.2. Optimization of PRMM based on minmax delay criterion yields a significantly different allocation pattern in contrast with the other two criteria. Symbols are more evenly allocated among layers under this criterion.

To further verify our results, we solved the proposed problems using `CVX` [35]. `CVX` is convex optimization package which can be used to solve for the solutions of convex problems using fast and efficient convex optimization solvers. The solutions obtained by `CVX` perfectly matched the analytical solutions in Table 4.2.

4.6 Conclusions

To select the best criterion to optimize any system one should first look on the limitations and objectives. For a real-time media multicast, there is always a deadline on transmitting every time-segment of the media stream. Therefore, system operators may seek for minimizing the number of transmitted packets while providing Heterogeneous clients with their demanded quality of service to achieve a real-time broadcast more reliably. Hence, the solution of min-max delay criterion appears to fit the design criterion for this example. In this chapter, we derived problem formulations for packetized rateless multimedia multicast based on minmax-regret and minmax-delay criteria. We have shown that the proposed formulations can be easily transformed into convex optimization problems. Furthermore, we managed to obtain the solutions

analytically by introducing a merging process which transforms the problems with active UEP constraints into new problems in which the remaining UEP constraints are not active when the solutions is obtained. Hence, equal regret (delay) criterion is used to find the optimal solutions of minmax-regret (delay) problem. Then, the solutions of the original problems are easily calculated from the solutions of the reduced problem accordingly. Numerical optimizations with `CVX` verified the obtained analytical solutions of the studied optimization criteria in a multicast system with four heterogeneous user-classes.

Table 4.2: Solutions of PRMM optimization under various criteria

(a) Min-Max Regret

Layer	γ_j	S_j	$\hat{\delta}_j$	w_j	$\hat{l}_j$	K_{h_j}	T_j	$E_{\Delta_j}[N_j]$	Max(T_j)	Max($E_{\Delta_j}[N_j]$)	N_{av}
1	25.8	400	0.70	1	23.03	17.37	20.06	26.06			
2	27.3	300	0.85	1	12.91	23.23	20.06	28.71	<u>20.06</u>	62.06	38.01
3	29.0	455	0.80	1	16.96	26.83	20.06	35.22			
4	40.3	2645	0.95	1	47.11	56.15	20.06	62.06			

(b) Min-Max Delay

Layer	γ_j	S_j	$\hat{\delta}_j$	w_j	$\hat{l}_j$	K_{h_j}	T_j	$E_{\Delta_j}[N_j]$	Max(T_j)	Max($E_{\Delta_j}[N_j]$)	N_{av}
1	25.8	400	0.70	1	13.30	30.10	39.14	45.14			
2	27.3	300	0.85	1	8.72	34.40	33.84	42.49	39.14	<u>45.14</u>	44.48
3	29.0	455	0.80	1	13.23	34.40	29.98	45.14			
4	40.3	2645	0.95	1	64.76	40.84	3.14	45.14			

(c) Average Delay

Layer	γ_j	S_j	$\hat{\delta}_j$	π_j	$\hat{l}_j$	K_{h_j}	T_j	$E_{\Delta_j}[N_j]$	Max(T_j)	Max($E_{\Delta_j}[N_j]$)	N_{av}
1	25.8	400	0.70	1/4	20.45	19.56	23.34	29.34			
2	27.3	300	0.85	1/4	15.34	19.56	15.51	24.16	24.11	66.11	<u>37.37</u>
3	29.0	455	0.80	1/4	19.99	22.76	14.72	29.88			
4	40.3	2645	0.95	1/4	44.22	59.81	24.11	66.11			

Chapter 5

Utility Maximization in Rateless Multimedia Multicast with Heterogeneous Clients

5.1 Introduction

Finding optimum user classification and resource allocation algorithms has been a major field of research in broadcast systems. These problems are more complicated when clients demand different adaptations of the media according to their capability and the transmitted media is a multilayer source, e.g., bitstream of a scalable video coder. These systems are optimized either to guarantee a certain level of service utility for the subscribed clients (QoS-guaranteed optimization) or maximize the delivered service utility to the subscribers based on some available transmission resources (Best-effort optimization). Few examples of systems that follow the first type of goal were introduced, analyzed and optimized in previous chapter. In this chapter,

we obtain a a best-effort optimization problem for a multimedia delivery system in which heterogeneous clients with various channel quality and terminal capability are considered. Clients can decode the media up to some layer according to their decoder complexity, power consumption policy, etc. As a result of decoding up to different layers, clients may achieve different utilities. However, the actual experience of any client is also function of its respective wireless channel and the resource allocation in server. For instance, a client may not be able to correctly recover a media layer when few symbols are transmitted from that layer or the packet loss rate (PLR) is high due to fading in wireless channel. Considering a server with limited transmission resources, flow of data from various media layers should be wisely chosen to deliver the maximum utility among clients. This optimal settings in server is obtained by jointly considering the multilayer source, clients' channels, their receiver/media-player capabilities, and their corresponding rate-utility functions.

Many attempts have been made to achieve multimedia multicast systems in which client's and their respective channels are considered in resource allocation optimization. Receiver-driven layered multicast (RLM) [36] is a notable system of this kind wherein enhancement layers of a scalable coded video are added or dropped within different branches of the network to avoid congestion in communication links. Congestion is detected via feedback from clients. The authors in [37] proposed a PET scheme based on multiple description coding to maximize expected fidelity for various loss rates. However, diversity among clients and capabilities of their decoders are not addressed. In another work [38], the server aims to optimize the transmission rates of different layers such that the provided utility among clients is maximized. However, heterogeneity of clients' media playback requirements is not considered.

In this chapter, we describe a novel best-effort optimization framework for rateless coded multimedia multicast. Heterogeneity in client channel quality as well as terminal capability are addressed. This is a distinctive feature of the proposed optimization which was not reported in prior works. The system setup is described in Section 5.2. Section 5.3 is devoted to the problem formulation. After some simplifying assumptions, we obtain a convex formulation for the proposed problem in Section 5.4. According to the numerical results in Section 5.5, we can see that convex formulation provides us with a close to optimal solution compared to the solution of the original formulation in various scenarios. Finally, conclusions are made in Section 5.6.

5.2 System Setup

Fig. 5.1 illustrates the system setup. A media server is responsible to provide various terminal (user device) classes with a multilayer media, e.g., an H.264/SVC encoded stream. Fountain coding [19] is utilized as the channel forward error correcting (FEC) code to protect different layers of the media against erasures due to channel errors.

The multicast clients need to adapt the transmitted bitstream to the media decoding and playback capabilities of their devices even though they aim to enjoy the same content (e.g., same movie or same live stream). Users with high definition (HD) displays naturally demand a HD stream to take full advantage of their receivers while cell-phone users may demand a standard definition (SD) adaptation or lower resolution to suit their screens, memory restrictions and/or power consumption policies. Scalable media standards such as H.264/SVC [1] have enabled generating a universal media stream which can be decoded partially or entirely by different terminals. For encoding, a sequence of video frames is partitioned into consecutive groups of pictures

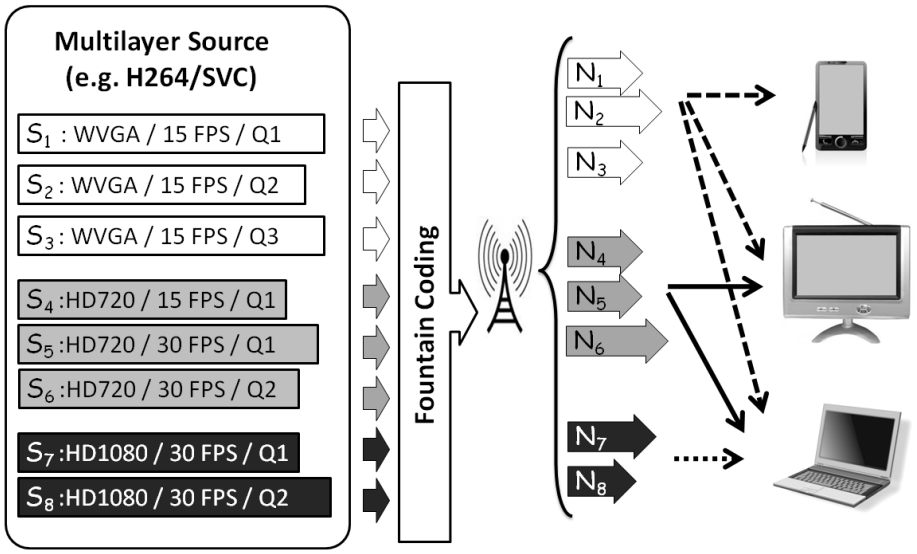


Figure 5.1: System setup for the proposed rateless-based video multicast

(GoPs). Each GoP, which may comprise the frames, let's say, over a one-second interval, is encoded into a scalable bitstream. The generated bitstream embeds J layers with S_j symbols in layer j, $j = 1, \ldots, J$. The fundamental layer of the bitstream (also known as base-layer) is essential to decoding process. Each additional layer may introduce a higher spatial or temporal resolution, or a finer amplitude resolution with less quantization error while maintaining the spatial-temporal resolution of the previous layer. We assume that successful decoding of any layer relies on success in decoding its preceding layers. This implies that media layers with lower indices are generally more important than those with higher indices. Fountain coding [19] is applied to each layer of the bitstream. Each digital fountain can theoretically generate an infinite number of encoded symbols from a layer of source symbols. Successful decoding is ensured probabilistically according to the number of symbols from the same fountain that are successfully recovered in the receiver. For practical considerations, we

assume that N_j symbols are transmitted for the j^{th} layer such that $\sum_{j=1}^{J} N_j = N_{max}$. N_{max} is set by the service provider, e.g., in order not to exceed the time at which transmission of the next GoP must commence. Our problem formulation below is applied separately for each GoP.

Multicast clients are modeled by M classes of media players, each including those who are potentially capable to decode up-to layer h_m, $m = 1, \ldots, M$ of the media. Class ordering is such that h_m is in increasing order. For example in Fig. 5.1, transmission of a source of $J = 8$ layers to $M = 3$ classes of users is considered. The mobile and portable TV clients are able to utilize up to layer $h_1 = 3$ and $h_2 = 6$, respectively, while all 8 layers can be decoded by HD clients ($h_3 = J = 8$).

The clients in each terminal class may have different receiver reception capabilities, e.g., due to having different bandwidths, antenna systems, etc. We model the reception capability of each client by reception coefficient (RC) $0 \leq \delta \leq 1$. By assuming the channel between the server and the client as a memoryless erasure channel (MEC) with i.i.d. erasures, then RC can be denoted as the transmission coefficient of this channel. Recalling our discussion on outage probability in Chapter 3, A client with RC δ_c receives an expected number of transmitted symbols $\delta_c N_{max}$ in the transmission period for one GoP. However, the number of symbols actually received depends on the channel symbol erasure events. The number of decoded source symbols depends on the performance of the fountain code and the decoder. We model the distribution of RCs for the clients in class m by the cumulative distribution function (CDF) $F_{\Delta_m}(\delta_m)$ of their RCs, $m = 1, \ldots, M$.

Each client in class m desires a media playback quality corresponding to a particular (no greater than h_m) layer of the media, as determined by the terminal capability,

software application, and user's choice. The quality of service (QoS) for clients desiring media layer $1 \leq j \leq J$ is expressed in terms of outage probability P_{out}^{j}. While it is conceivable that the clients desiring layer j quality might want different levels of reception assurance, for simplicity we assign one assurance level, in the form of probability $1 - P_{out}^{j}$, to each media layer. P_{out}^{j} are assumed to be increasing with layer index j.

Every additional symbol drawn from a digital fountain ideally improves the probability of decoding. Thus, if N_{max} is allowed to be indefinitely large, all clients with non-zero RC will reach perfectly assured QoS. In general, for a given set of N_j with $\sum_{j=1}^{J} N_j = N_{max}$, we can find a set of minimum RC values $\delta_{min,j}$ such that those clients with RC $\delta_c < \delta_{min,j}$ and desiring layer j media quality will not reach the QoS assurance probability $1 - P_{out}^{j}$. To satisfy all the users who desire layer j quality, we can pick $\delta_{min,j}$ to be equal to the minimum of the RCs of all clients. Doing so might compromise the assurance levels that can be provided to the clients who desire other levels of quality. Hence, careful consideration should be taken in order to choose the best set of $\delta_{min,j}$ respecting to different class channel distributions $F_{\Delta_m}(\delta_m)$ and transmission deadline N_{max} and impact of reconstructing the media up to different layers on the acquired utility of different clients which is itself a function of terminal capabilities. By establishing a proper formulation that properly addresses these issues, the optimal minimum RCs $\delta_{min,j}, 1 \leq j \leq J$, can be determined to maximize an overall utility measure over all the clients, or conversely to minimize an overall dissatisfaction measure.

Since decoding layer j requires successful decoding of all layers $l \leq j$, we can easily

prove that $\delta_{min,j}$ must be in non-decreasing order with respect to layer index j, i.e.,

$$0 < \delta_{min,1} \leq \delta_{min,2} \leq \ldots \leq \delta_{min,J} \leq 1. \tag{5.1}$$

Lemma 5.1: *Any possible set of $\delta_{min,j}, j = 1, ..., J$ which is not in non-decreasing order is not the optimal solution for the proposed multicast system.*

Proof. Let us assume that $\delta_{min,l} < \delta_{min,l-1}$. Therefore, users with $\delta_{min,l} \leq \delta_c < \delta_{min,l-1}$ are provided with a utility associated to layer $l-2$ or bellow that. Suppose we use part of the transmission resources which is currently devoted to layer l of the media to transmit more symbols from layer $l-1$. Hence, $\delta_{min,l-1}$ is reduced and $\delta_{min,l}$ is increased. When the applied changes are small, we can conclude that $\delta_{min,l} < \delta^{new}_{min,l} < \delta^{new}_{min,l-1} < \delta_{min,l-1}$. Consequently, clients with $\delta^{new}_{min,l-1} \leq \delta_c < \delta_{min,l-1}$ are now able to additionally decode layer l and $l-1$. Hence, the provided utility to the system is increased. This modification has no effect on the utility of clients with $\delta_{min,l} < \delta_c < \delta^{new}_{min,l-1}$ since they still can neither decode layer $l-1$ nor layer l. This modification shows that the original set of $\delta_{min,j}, j = 1, ..., J$ was not optimal and the proof is complete. □

5.3 Problem Formulation

Using the client service outage constraints $P^j_{out}, j = 1, ..., J$, and after defining $P^0_{out} \triangleq 0$, we can obtain another set of constraints which comprises the maximum outage probability P_j in decoding a layer j of the source. By considering independency between the decoding of the fountain codes.

$$P_j = 1 - \frac{1 - P^j_{out}}{1 - P^{j-1}_{out}} \ , \qquad j = 1, \ldots, J. \tag{5.2}$$

Lemma 5.2: *Let's assume P_j as in* (5.2)*. If the outage probability in decoding the fountain code of source layer j for the client $\mathcal{C}$ complies with $P_j^{\mathcal{C}} \leq P_j$ for $j = 1, 2, ..., k$, then experienced outage probability in decoding the media up to layer k for this client satisfies $P_{out,\mathcal{C}}^k \leq P_{out}^k$, when we consider independent erasure events on fountain symbols of different source layers.*

Proof. By considering independent erasure events among fountain symbols of different layers along with independence in decoding and encoding processes for different digital fountains, we conclude that recovering the symbols of different fountains are independent from each other. Hence, the outage probability in decoding up to the layer k of the source is:

$$\begin{aligned} P_{out,\mathcal{C}}^k &= 1 - \prod_{j=1}^{k} \text{Prob(layer } j \text{ is perfectly recovered by user } \mathcal{C}) \\ &= 1 - \prod_{j=1}^{k} (1 - P_j^{\mathcal{C}}). \end{aligned}$$

Since $P_j^{\mathcal{C}} \leq P_j$ for $j = 1, 2, ..., k$ and by considering (5.2) we have

$$\begin{aligned} 1 - P_{out,\mathcal{C}}^k &= \prod_{j=1}^{k} (1 - P_j^{\mathcal{C}}) \\ &\geq \prod_{j=1}^{k} (1 - P_j) \\ &= \prod_{j=1}^{k} \frac{1 - P_{out}^j}{1 - P_{out}^{j-1}} = \frac{1 - P_{out}^k}{\cancel{1 - P_{out}^{k-1}}} \times \frac{\cancel{1 - P_{out}^{k-1}}}{\cancel{1 - P_{out}^{k-2}}} \times ... \times \frac{\cancel{1 - P_{out}^{1}}}{1 - P_{out}^0} \\ &= \frac{1 - P_{out}^k}{1 - P_{out}^0} = 1 - P_{out}^k, \end{aligned}$$

or equivalently $P_{out,\mathcal{C}}^k \leq P_{out}^k$. The proof is complete. □

Now, we begin to establish the relationship between the number of transmitted

symbols from each layer N_j to the RC of users in different classes δ_j according to our newly defined outage probability constraints $P_j, j = 1, ..., J$. The benefit of using a simplified channel model (SCM) in lieu of memoryless erasure channels in rateless multimedia multicast was previously described in Chapter 3. SCM leads to lower computational complexity in exchange for less accuracy. According to (3.7), the number of required transmitted symbols from layer j of the media, with S_j media symbols, that guarantees the layer-dependent outage constraint P_j for a user with RC $\delta_{min,j}$ can be well estimated by

$$N_j = \frac{S_j(1+\epsilon_j)}{\delta_{min,j}} \ , \tag{5.3}$$

where ϵ_j is the *additional overhead* needed to achieve the outage probability P_j for decoding the source layer j. Comparing (5.3) with (3.7), ϵ_j can be expressed as

$$\varepsilon_j = \frac{\ln(P_j/a)}{S_j \ln b} \ , \quad P_j \leq a \ , \tag{5.4}$$

where $0 < a, b < 1$ are model parameters, with $a = 0.85$ and $b = 0.567$ for the standardized Raptor codes [21] utilized in [30]. According to (5.4), the overhead is a function of code characteristics, code length, and desired outage probability. The overhead is decreased by setting higher layer-dependent outage probabilities or using codes with smaller decay factor, b.

N_j can be also approximated using the proposed synthetic model for outage probability as described in Chapter 3. By re-arranging the terms in (3.13) we can easily show that

$$N_j = \frac{S_j}{\delta_{min,j}} + \tau_{out,j} \times \sqrt[H]{\frac{1-\delta_{min,j}}{\delta_{min,j}}} \ , \tag{5.5}$$

where $\tau_{out,j}$ is obtained from

$$\tau_{out,j} = \sqrt[H]{-S_j \times \ln\left(P_j/0.5\right)} \qquad (P_j \leq 0.5). \tag{5.6}$$

The total number of transmitted symbols can not exceed N_{max}. Consequently, when simplified model (3.7) is utilized, the maximum transmission constraint is

$$\sum_{j=1}^{J} N_j \leq N_{max} \iff \sum_{j=1}^{J} \frac{S_j(1+\epsilon_j)}{\delta_{min,j}} \leq N_{max}. \tag{5.7a}$$

Similar constraint is obtained when N_j is estimated using the proposed synthetic model in (3.13).

$$\sum_{j=1}^{J} N_j \leq N_{max} \iff \sum_{j=1}^{J} \frac{S_j}{\delta_{min,j}} + \tau_{out,j} \times \sqrt[H]{\frac{1-\delta_{min,j}}{\delta_{min,j}}} \leq N_{max} \tag{5.7b}$$

Let U_m be the guaranteed utility for class m users. This utility is a function of the number of users who are provided with any specific utility level of the media, as well as the amount of utility they each gain, i.e.,

$$U_m = \sum_{j=1}^{h_m} \alpha_{m,j}(1 - F_{\Delta_m}(\delta_{min,j})). \tag{5.8}$$

Here, $\alpha_{m,j}$ is the additional utility gained by a class m client after decoding layer j of the media, provided that all preceding layers are successfully decoded. $1 - F_{\Delta_m}(\delta_{min,j})$ corresponds to the portion of class m clients who can decode layer j with the desired outage probability, due to their RCs $\delta \geq \delta_{min,j}$. Validity of the right side of (5.8) is ensured when $\delta_{min,j}, j = 1, ..., J$ are in non-decreasing order. The overall

utility is obtained from

$$\begin{aligned} U_{total} &= \sum_{m=1}^{M} \pi_m U_m = \sum_{m=1}^{M} \pi_m \sum_{j=1}^{h_m} \alpha_{m,j}(1 - F_{\Delta_m}(\delta_{min,j})) \\ &= \hat{U}_{max} - \sum_{m=1}^{M} \sum_{j=1}^{h_m} \hat{\alpha}_{m,j} F_{\Delta_m}(\delta_{min,j}). \end{aligned} \tag{5.9}$$

Here, $\pi_m > 0$ is a weight reflecting the distribution of users across the classes with $\sum_{m=1}^{M} \pi_m = 1$. These priors can be absorbed into $\alpha_{m,j}$ by defining $\hat{\alpha}_{m,j} = \pi_m \alpha_{m,j}$ and $\hat{U}_{max} = \sum_{m=1}^{M} \sum_{j=1}^{h_m} \hat{\alpha}_{m,j}$. The first term in (5.9) is not a function of minimum RCs $\delta_{min,j}$. Therefore, maximizing the total utility is equivalent to minimizing the absolute value of the negative term. The optimization problem is summarized by

Problem 5.1: *General Formulation*

$$\min_{\{\delta_{min,j}\}_{j=1}^{J}} \quad \sum_{m=1}^{M} \sum_{j=1}^{h_m} \hat{\alpha}_{m,j} F_{\Delta_m}(\delta_{min,j})$$

Subject to

$$\textit{UEP Const. :} \qquad 0 \leq \delta_{min,1} \leq \delta_{min,2} \leq ... \leq \delta_{min,J} \leq 1$$

$$N_{max}\ \textit{Const. :} \begin{cases} \displaystyle\sum_{j=1}^{J} \frac{S_j(1+\epsilon_j)}{\delta_{min,j}} \leq N_{max} & \text{if Simple Model (3.7)} \\ \displaystyle\sum_{j=1}^{J} \frac{S_j}{\delta_{min,j}} + \tau_{out,j} \times \sqrt[H]{\frac{1-\delta_{min,j}}{\delta_{min,j}}} \leq N_{max} & \text{if Synthetic Model (3.13)} \end{cases}$$

$\delta_{min,j}, j = 1, \ldots, J$ are optimization variables. As we can see, Problem 5.1 represents two separate problems which are different from each other based on the model they use to keep track of outage probabilities and enforce the maximum transmission constraint. After obtaining the optimal $\delta_{min,j}$, (5.3) or (5.5) can be used to obtain the number of transmitted symbols from each layer of the source. As a result of this optimization, a HD client with a medium quality channel may not achieve the utility

of decoding the highest layer of the media, since the the minimum RC threshold $\delta_{min,J}$ for the last layer may rise above the client's RC, unless a huge amount of transmission resources is available in server. Similar to that, clients with relatively poor channels are automatically removed from the multicast since their RC is below the minimum RC for the base layer, $\delta_{min,1}$. In contrast to [38] in which clients are individually represented in the optimization, the multicast clients are considered globally through the distributions $F_{\Delta_m}(\delta)$ and associated priors π_m. Consequently, the complexity of the proposed optimization is independent of the number of clients. Client-to-server feedback can be managed to prevent feedback implosion, e.g., by sending a threshold and asking clients with a random number generator output above the threshold to send information about their RC and terminal capability. This threshold can be adjusted adaptively such that the amount of received feedbacks can be handled by the server. Consequently, feedback from all clients is not needed to estimate the distributions of different classes $F_{\Delta_m}(\delta)$ when the size of the multicast is relatively big compared to the available reception bandwidth of the server. However, this might compromise the accuracy of the estimation of class channel distributions $F_{\Delta_j}(\delta_j), j = 1, ..., J$. After performing the proposed optimization, we have:

- *Users of any class with channel quality $\delta < \delta_{min,1}$ are <u>not</u> provided with any guaranteed utility.*

- ***For*** $m = 1 : M$
  For $j = 1 : J$
  Class m users with $\delta_{min,j} \leq \delta < \delta_{min,j+1}$ decode up to layer $k^* = \min\{j, h_m\}$ and achieve the utility $U_{m,k^*} = \sum_{k=1}^{k^*} \alpha_{m,k}$.
  End For *j;*
  End For *m;*

5.4 Convex Formulation

In general, solving Problem 5.1 can be a computationally complex task. The cost function may have multiple extremum points and the solution may not be unique. Moreover if any parameter is changed with respect to previous GoP, the optimal solution for Problem 5.1 should be re-calculated. Therefore, we are facing a trade-off between accuracy and complexity when it comes to finding the solution. Proposed optimization problem is not guaranteed to be convex for every distribution since CDF is not a convex function in general. Moreover, if the synthetic model is utilized to enforce the N_{max} constraint, then the maximum transmission constraint may not define a convex set either. However, after some reasonable simplifications, we can obtain a convex optimization from Problem 5.1. Convex optimization problems usually incorporate faster algorithms for obtaining the solutions. Conversely, the simplifications may severely affect the accuracy of the system such that the obtained solution is not reliable. Therefore, careful considerations should be taken when simplifying problems. In many practical applications, clients with very poor channel ($\delta \ll 1$) are automatically omitted because of their unreliable communication link. On the other hand, transmission resources in wireless systems are valuable, Therefore, they may not be sufficient to provide clients with poor quality channels, even if they were the only clients who were considered in the system. Most of these clients end up being discarded from multicast since optimal $\delta_{min,1}$ becomes higher than their actual RCs, i.e., $\delta_c < \delta_{min,1}$. Therefore, we may focus on clients with medium and high quality channels, who are represented by high δ regions in user's channel quality distribution of each class. Based on this assumption, we approximate $F_{\Delta_m}(\delta)$ with

$$\widetilde{F}_{\Delta_m}(\delta) = \frac{c_1 + c_2\delta^p}{c_1 + c_2} \qquad (p, c_1, c_2 > 0) \tag{5.10}$$

in which p, c_1 and c_2 are parameters. For every distribution, the suitable parameters are obtained such that mean-squared-error (MSE) is minimized over $\delta \geq 0.4$.

Furthermore, the following approximation on N_j is introduced to simplify (5.5) in the case that synthetic model is utilized to enforce the N_{man} constraint.

$$N_j \approx \frac{S_j}{\delta_{min,j}} + \tau_{out,j} \times K(\frac{1}{\delta_{min,j}} - 1) \tag{5.11}$$

Here, $K = 0.83$ minimizes the mean-absolute-error between (5.11) and (5.5) over $0.25 \leq \delta \leq 1$ and $S \in \{500, 1000, 2000, 4000\}$. By applying these approximations and after introducing a parameter transformation $\theta_{max,j} = 1/\delta_{min,j}, \forall j$, we obtain the new optimization problem.

Problem 5.2: *Convex Formulation*

$$\min_{\{\theta_{max,j}\}_{j=1}^{J}} \sum_{m=1}^{M} \sum_{j=1}^{h_m} \hat{\alpha}_{m,j} \widetilde{F}_{\Delta_m}(1/\theta_{max,j})$$

Subject to

$$\textit{UEP Const. :} \quad 1 \leq \theta_{max,J} \leq \theta_{max,J-1} \leq ... \leq \theta_{max,1} < +\infty$$

$$N_{max} \textit{ Const. :} \begin{cases} \sum_{j=1}^{J} S_j(1+\epsilon_j)\theta_{max,j} \leq N_{max} & \text{if Simple Model (3.7)} \\ \sum_{j=1}^{J} (S_j + K\tau_{out,j})\theta_{max,j} \leq N_{max} + \sum_{j=1}^{J} K\tau_{out,j} & \text{if Synthetic Model (3.13)} \end{cases}$$

In the first step to show that Problem 5.2 is convex optimization, we take the

second derivative of the cost function respect to $\theta_{max,j}$.

$$\begin{aligned}\chi &= \frac{\partial^2}{\partial\theta_{max,j}^2}\sum_{m=1}^{M}\sum_{j=1}^{h_m}\hat{\alpha}_{m,j}\widetilde{F}_{\Delta_m}(1/\theta_{max,j}) \\ &= \sum_{m=m'}^{M}\hat{\alpha}_{m,j}\frac{\partial^2}{\partial\theta_{max,j}^2}\left(\frac{c_1+c_2/\theta_{max,j}^p}{c_1+c_2}\right) \\ &= \sum_{m=m'}^{M}\hat{\alpha}_{m,j}\frac{c_2}{c_1+c_2}p(p+1)\theta_{max,j}^{-(p+2)} \qquad (5.12)\end{aligned}$$

Since $c_1, c_2, p > 0$ and $\hat{\alpha}_{m,j} > 0$ and $\theta_{max,j} > 0,\ \forall j$, we conclude that $\chi > 0$. Therefore, the proposed cost function is convex according to the second order condition of convex functions [39]. Also, since the constraints are linear, Problem 5.2 is a convex optimization. If the provided utility of the solution of convex formulation is relatively close to the utility provided by solution of general problem for a wide range of parameters, we may also conclude that the convex formulation is a good approximation of the original problem.

5.5 Numerical Simulations

We consider a scenario with $M = 2$ classes of clients each capable of decoding up to layer $h_m = \{2, 3\}$ of a multilayer video stream with $J = 3$ layers as described in Table 6.1. The video corresponds to 1 sec of *Crew* video sequence being encoded using H.264/SVC and Each symbol represents 50 bytes. Outage probability constraints $P_{out}^j = \{10^{-4}, 5\times10^{-4}, 10^{-3}\}$ are considered in simulations. Truncated Gaussian distributions as depicted in Fig. 5.2 model the RCs of clients in each class. The following utility coefficients $\alpha_{m,j}$ are considered.

Table 5.1: Specification of H.264/SVC video bitstream .

LID	Resolution	FPS	BitRate (kbps)	Y-PSNR	S_j
1	QCIF 176x144	15	150.42	37.3	376
2	CIF 352x288	30	758.34	37.1	1520
3	4CIF 704x576	60	3560.25	37.7	7380

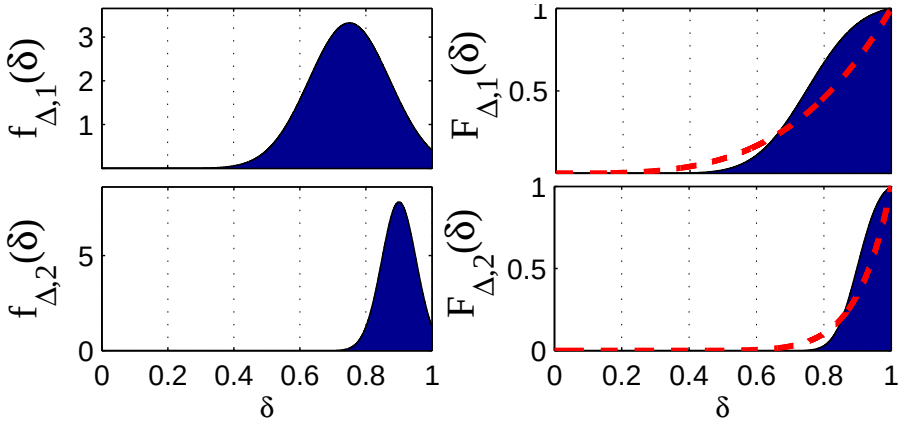


Figure 5.2: (left) PDFs and (right) CDFs of channel distributions for a multicast with 3 classes of users. Approximated distributions are depicted in dash-red

$$[\alpha_{m,j}]_{M\times J} = \begin{bmatrix} 0.35 & 0.52 & 0 \\ 0.31 & 0.50 & 0.19 \end{bmatrix} \tag{5.13}$$

Utilities are normalized by the utility of an HD media player who is able to decode all the media layers. Without the loss of generality, we assumed that a HD terminal gains a lower utility after decoding the first two layers of the media compared to a SD terminal that decodes the same layers. This difference models the inefficiencies incurred by an HD terminal with more sophisticated decoder when it decodes up to the same layer of the media as a SD terminal. Following measures are used to evaluate

the performance of our algorithm.

$$\eta_{EEP}^{CV} \uparrow = \frac{U_{CV} - U_{EEP}}{U_{EEP}}\% \quad , \quad \epsilon^{CV} = \frac{U_{CV}}{U_{OPT}}\% \tag{5.14}$$

Here U_{opt} is the maximum attainable utility and is obtained after searching over all possible allocation schemes in Problem 5.1. This can be done by partitioning the $0 < \delta \leq 1$ to sufficiently small intervals and performing an exhaustive search over $\delta_{min,j}, j = 1, \ldots, J$ to find the best attainable system parameters. Nevertheless, this process can be very computationally complex as the number intervals grows. U_{EEP} is the performance obtained from equal error protection (EEP) solution wherein transmission rates for different layers are equal[1]. Therefore, the EEP solution yields the following resource allocation in server.

$$N_j^{EEP} = \frac{N_{max} \times S_j}{\sum_{k=1}^{J} S_k} \qquad j = 1, ..., J$$

$\eta_{EEP}^{CV} \uparrow$ measures the increase in performance of the multicast system using solution of convex optimization compare to EEP solution while ϵ^{CV} measures the efficiency the solution obtained from solving the convex problem (U_{CV}), respect to maximum attainable performance U_{opt}. All these measures are evaluated by using non-approximated distributions $F_{\Delta_m}(\delta)$. Solution of convex formulation serves as a starting point for a constrained gradient descent (GD) algorithm applied to the original non-convex problem. Similarly, $\eta_{EEP}^{GD} \uparrow$ and ϵ^{GD} are defined to evaluate the performance and efficiency of the solution when GD is applied.

Table 5.2 illustrates the results for different scenarios. In every scenario, N_{max}

[1]In fact, rateless code overhead ε is not fixed according to (5.4). Hence, various layers are not equally protected even-though the transmission rates are equal. However, this issue has negligible impact on EEP solution.

Table 5.2: Simulation results for multiple scenarios with $M = 2$ and $J = 3$.

		Performance (%)		Efficiency (%)	
N_{max}	π_1	$\eta_{EEP}^{CV} \uparrow$	$\eta_{EEP}^{GD} \uparrow$	ϵ^{CV}	ϵ^{GD}
	0.10	950.23	951.22	100.00	100.00
	0.30	1093.08	1105.46	99.86	99.87
11,000	0.50	1373.47	1411.65	97.06	99.57
	0.70	1851.23	1954.04	94.98	99.99
	0.90	2892.49	3097.79	93.58	100.00
	0.10	64.99	67.72	98.37	100.00
	0.30	92.05	95.08	98.44	99.99
12,000	0.50	136.97	142.14	97.85	99.99
	0.70	220.93	230.90	96.98	100.00
	0.90	423.77	446.98	95.76	100.00
	0.10	9.69	11.70	98.20	100.00
	0.30	25.24	28.62	97.37	100.00
13,000	0.50	49.16	53.99	96.86	100.00
	0.70	89.00	95.48	96.68	100.00
	0.90	166.89	174.90	97.09	100.00
Average		**631.61**	**657.85**	**97.40**	**99.96**

constraint is enforced using the simple model for outage probability. By solving the convex problem using a convex optimization toolbox, the solution yields an average efficiency of 97.4%. Using GD algorithm, the efficiency is increased to 99.96%. Since the efficiency of the solution of convex formulation is quite high, GD method can be avoided without significant loss in performance (only 2.56% drop in performance). Moreover, provided utility is more than 6 times higher than the utility of EEP solution. Minimum RCs $\delta_{min,j}$ for different scenarios are demonstrated in Fig. 5.3. $\delta_{min,j}$ obtained from convex formulation are close to those in optimal allocation. By increasing maximum transmitted symbols N_{max}, minimum RCs are reduced. Consequently,

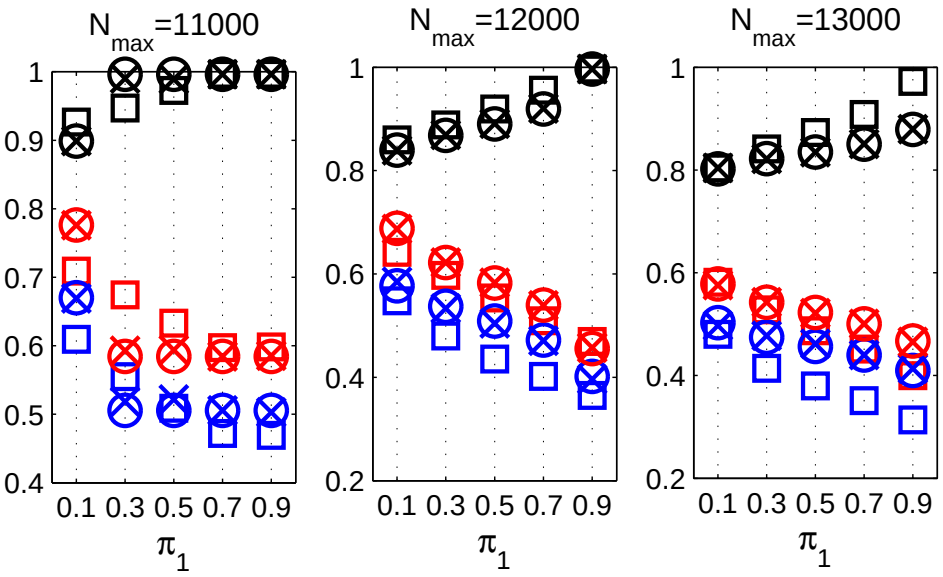


Figure 5.3: $\delta_{min,1}$ (blue) , $\delta_{min,2}$ (red) and $\delta_{min,3}$ (black) obtained using (square) Convex Optimization , (cross) Gradient Descent and (circle) Brute-force search for various settings for prior weights in Table 5.2 ($\pi_2 = 1 - \pi_1$).

service boundaries are extended to clients with less channel quality, whether by providing utility to clients who were previously discarded from multicast or providing more utility to some existing clients since they can decode up to a higher layer now. $\delta_{th,3} \approx 1$ For $N_{max} = 11000$ implies that only HD terminals with perfect channel are able to achieve the highest utility for $\pi_1 \geq 0.3$. For $N_{max} = 13000$, decoding all 3 layers of the bitstream is nearly possible for HD clients with RC $\delta \geq 0.8$.

Fig. 5.4 illustrates the region-of-interest and efficiencies of various resource allocations under both problem formulations in one scenario. The contours of the cost function in original and convex problem are very similar. This implies the fact that the convex formulation is a good approximation of general formulation in the corresponding scenario. Solution for similar scenario is obtained when the N_{max} constraints in

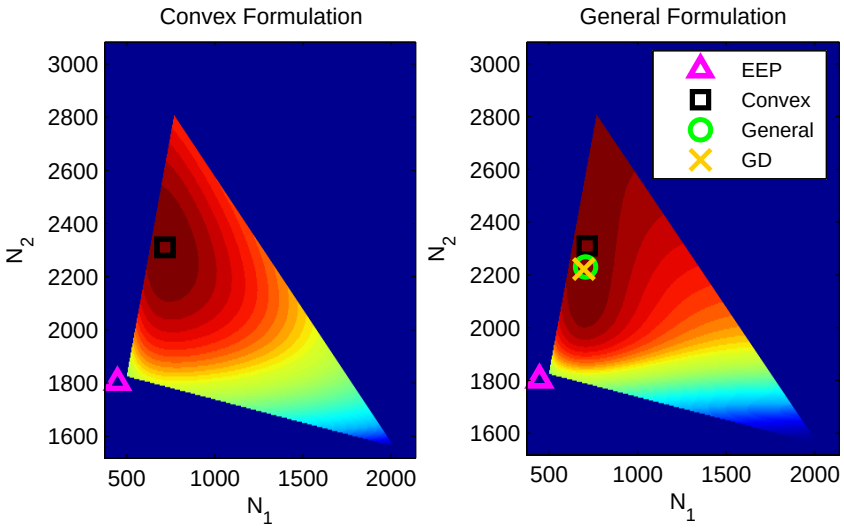


Figure 5.4: Efficiency of provided service for various operating points for $N_{max} = 11000$ and $\pi_1 = 0.2$ under (left) convex and (right) general formulations. ($N_3 = N_{max} - N_1 - N_2$).

both Problem 5.1 and Problem 5.2 are obtained based on the synthetic model. Utilities of different optimization problems as a function of N_{max} are depicted in Fig. 5.5. Note that the utilities are calculated based on the closed-from outage probabilities in (3.11) and non-approximated distributions. Utility of the general formulation is upper-bounded by U_{max}, while itself serves as an upper-bound for the utility of the corresponding convex problem. Depending on the distributions of different classes of media player devices, $\hat{U}_{max}$ may or may not be achievable in different scenarios.

As we expected before, utility of general formulation is higher than the utility corresponding to the convex problem. However, the difference is very insignificant. Furthermore, optimizations that employ the synthetic model yield higher utility than those with simple model. Utilities of different optimizations do not differ more than 2% for any given N_{max}. This justifies the use of simplified channel model as well as the approximations which were made to obtain the convex formulations for this

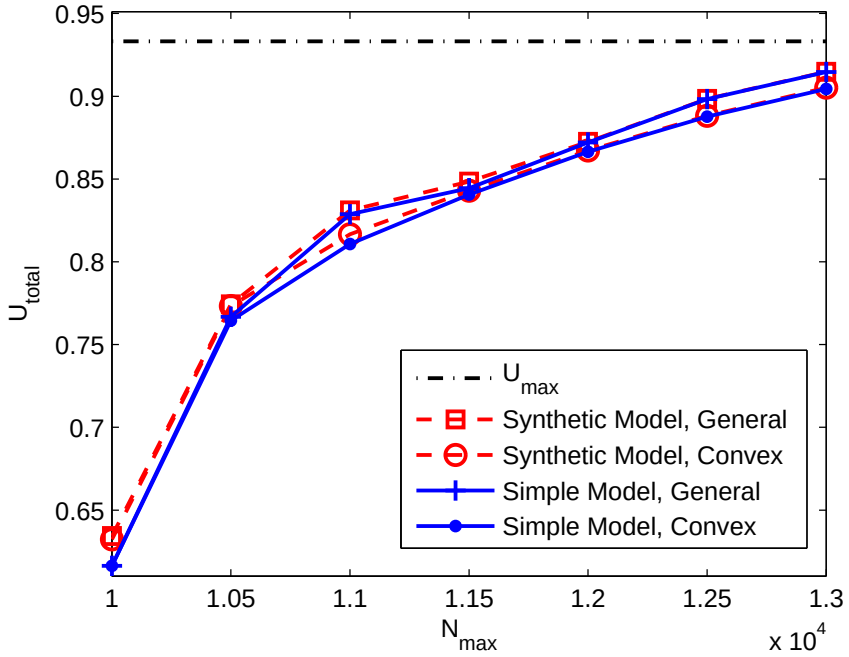


Figure 5.5: Provided utility (U_{total}) for different optimizations and various models ($\pi_1 = \pi_2 = 0.5$).

best-effort optimization[2].

5.6 Conclusions

We derived the formulation for a best-effort utility maximization in rateless multimedia multicast with heterogeneous clients. To solve the general problem, we introduced a two stage algorithm in which a convex problem is formulated by introducing few simplifications. The solution of the convex problem serves as a starting point for a gradient search method over the non-convex problem. The proposed algorithm

[2]Unlike best-effort scenarios, in QoS-guaranteed scenarios, the system has to provide the heterogeneous clients with their demanded reliability factor. Deficiency of any model which is utilized to keep track of the outage probabilities (e.g, simple model in Chapter 3) directly results in violation of those reliability factors.

yields an efficiency of more than 99.9% in average while the solution of the convex formulation alone achieved a 97.4% efficiency. Unlike [38] and similar methods, the complexity of the proposed optimization is independent of the number of clients.

Chapter 6

Maximization of Perceived Quality Experience in Multi-Scalable Video Multicast

6.1 Introduction

The resource allocation optimization introduced in Chapter 5 can be tailored to fit to different optimization scenarios. In this chapter, we aim to maximize the perceptual quality experience among clients. To adapt the proposed optimization to this particular application, we are mainly concerned with mapping video viewing quality to user utility by means of setting the proper utility parameters $\alpha_{m,j}$. Scalable video coders can generate video streams which embed different forms of scalability, mainly spatial, temporal and fidelity scalability. In this chapter we also investigate whether embedding multiple types of scalability in the generated video bitstream can benefit us in delivering the highest perceptual quality experience to clients.

6.2 Mapping Utility Parameters to The Perceptual Quality

Difficulties in modeling the perceptual quality of a media (particularly video) have been discussed in Chapter 2. Although, PSNR has been widely used as a measure of video quality in cross-layer optimizations, low correlation between PSNR and perceptual video quality ratings provided by human viewers, known as mean opinion score (MOS), has been reported [29]. Shortcomings of PSNR is even more prominent when comparing video playbacks of different spatial and temporal resolutions. Two encoded videos with different spatial-temporal resolutions may lead to the same PSNR, when they are compared against the uncoded video sequences in their respective resolution. However, perceived quality as a result of watching these two videos can be significantly different. In fact, it is impossible to calculate the PSNR between two videos of different spatial-temporal resolutions, however perceptual quality for videos of different resolutions can be quantified and compared.

Unlike MOS, objective quality measures do not need human viewing and can be fashioned as estimates of subjective MOS. Xue et'al [40] proposed the normalized MOS (NMOS_s). This function models the change of MOS among videos of various spatial resolution.

$$\text{NMOS}_s(s) = \frac{1 - e^{-b_s \frac{s}{s_{max}}}}{1 - e^{-b_s}} \tag{6.1}$$

Here, s is the number of pixels and s_{max} is the highest spatial resolution. b_s is the model parameter and function of the video content. In another work of the same team [41], it is demonstrated that the change of MOS with respect to frame rate f

and distortion level PSNR can be accurately modeled using

$$\mathrm{NMOS}_c(\mathrm{PSNR}, f) = \left(1 - \frac{1}{1 + e^{p(\mathrm{PSNR} - b_p)}}\right) \times \frac{1 - e^{-b_f \frac{f}{f_{max}}}}{1 - e^{-b_f}}, \tag{6.2}$$

where, f_{max} is the maximum frame-rate, b_f and b_p are model parameters and $p = 0.34$. PSNR measures the fidelity of the video against its respective uncoded video with the same spatial-temporal resolution. Combining (6.1) and (6.2), we obtain our desired normalized metric that takes into account distortion level, spatial and temporal resolution of the video, and the video content when quantifying the video quality.

$$\begin{aligned} \mathrm{NMOS} &= \mathrm{NMOS}_s(s) \times \mathrm{NMOS}_c(\mathrm{PSNR}, f) \\ &= \frac{1 - e^{-b_s \frac{s}{s_{max}}}}{1 - e^{-b_s}} \times \frac{1 - e^{-b_f \frac{f}{f_{max}}}}{1 - e^{-b_f}} \times \left(1 - \frac{1}{1 + e^{p(\mathrm{PSNR} - b_p)}}\right) \end{aligned} \tag{6.3}$$

According to (6.3), the perceptual quality of a video with the highest frame rate ($f = f_{max}$) and spatial resolution ($s = s_{max}$) and no compression (PSNR $= \infty$) corresponds to NMOS $= 1$. Other combinations of spatial, temporal and PSNR resolutions result in a lower NMOS score[1].

The highest layer of the media, successfully recovered by a terminal, might be of a spatial resolution lower than the terminal capability. For instance, let's consider a HD terminal receiving a SD video. The terminal may display the SD video as received in the middle of the HD display or adapt the video to the display by proper upsampling, if capable. Upsampling increases spatial resolution parameter s in (6.3) and potentially can increase the perceived quality. However, the PSNR of an upsampled video is

[1]In general, different anchor points (here different video adaptations) can be used to normalize the quality. Choosing different anchor points can result in different parameters in optimization and hence different solutions.

lower than PSNR in the lower spatial resolution. Therefore, the overall quality may in fact decrease.

We assume that clients adapt the received video stream such that the highest perceptual quality is attained. Using this model in receiver they can actually decide what would be the optimal adaptation that maximizes the the video quality for any possible set of recovered media layers. Thus, we define NMOS_j^m to represent the perceptual quality corresponding to the best possible adaptation —within the capabilities of the the class m terminals— that can be performed on the media up to layer $j \leq h_m$. Here h_m is the highest layer if the video stream that class m terminals can potentially decode. For example, video decoder implemented in a smart-phones may not be able to decode HD layers of the video while the embedded decoder in a laptop can possibly decode these layers.

Since being able to recover higher layers of a media can not reduce the quality experience, NMOS_j^m is non-decreasing with respect to layer index j. The incremental utility coefficients $\alpha_{m,j} \geq 0$ for $j \leq h_m$ capture the possible increase in utility for a class m terminal, when layer j is additionally recovered compared to layer $j-1$,

$$\alpha_{m,j} = W(m,j)\text{NMOS}_j^m - W(m,j-1)\text{NMOS}_{j-1}^m. \tag{6.4}$$

Here, $0 \leq W(m,j) \leq 1$, and non-decreasing with respect to j, models the dissatisfaction of the class m clients for only being able to decode up to layer j, relative to their terminal capability h_m. If class m users are unwilling to settle for the media quality of any layer $j < h_m$, then $W(m,j) = 0 \ \ \forall j < h_m$.

6.3 Numerical Results

We investigate the efficiency of multi-scalable video multicast using the proposed framework to maximize the perceptual quality experience of heterogeneous clients. The investigation is based on precoded bitstreams. i.e., video encoding is performed without knowledge of the actual clients. Using H.264/SVC, one GoP (16 frames) of the *Crew* video sequence is used to construct several scalable bitstreams for QCIF (176×144) and/or CIF (352×288) resolutions, according to Table 6.1. NMOS model parameters for this particular video are $b_f = 7.23, b_p = 29.68$ and $b_s = 8.84$. To avoid excessive complications, temporal scalability is not investigated in these scenarios ($f = 30$ for different adaptations). Each source symbol represents 5 bytes. Normative up-sampling and non-normative down-sampling is used to adapt the decoded video of one spatial resolution to the other [1].

Two hypothetical PSNR-scalable bitstream (with QCIF resolution) and spatial-scalable bitstreams are constructed using rate-distortion data from the combined-scalable bitstream. To enable fair comparison, we want layers of single-scalable bitstreams to supply the same rate-distortion values as their corresponding layers in the combined-scalable stream. There is always a rate distortion penalty when enforcing additional scalable layers in the encoded video. This penalty is due to the fact that the optimal coding strategy may conflict with the constraint of having additional layer. Therefore, to achieve the new scalable layer, the coding efficiency may be compromised[2]. Consequently, the bit rate needed to achieve a particular

[2]Any additional scalability layers is translated into an additional constraint in the source coding problem. Eventually, the solution of a more constraint problem yields an equal or higher maximum (lower minimum) compared to the solution of the problem with similar objective function and only a subset of those constraints.

Table 6.1: Specification of H.264/SVC video streams.

Stream	Layer ID	Resolution	FPS	Bit Rate (kbps)	S_j	Y-PSNR (dB)	
						QCIF	CIF
Combined Scalable	Q1	QCIF	30	52.57	701	29.8	27.0
	Q2	QCIF	30	82.58	400	31.1	27.8
	C1	CIF	30	299.83	2897	32.4	30.3
	C2	CIF	30	624.81	4333	38.8	35.3
PSNR Scalable (QCIF)	Q1	QCIF	30	52.57	701	29.8	27.0
	Q2	QCIF	30	82.58	400	31.1	27.8
PSNR Scalable (CIF)	C1	CIF	30	299.83	3998	32.4	30.3
	C2	CIF	30	624.81	4333	38.8	35.3
Spatial Scalable	Q2	QCIF	30	82.58	1101	31.1	27.8
	C2	CIF	30	624.81	7230	38.8	35.3

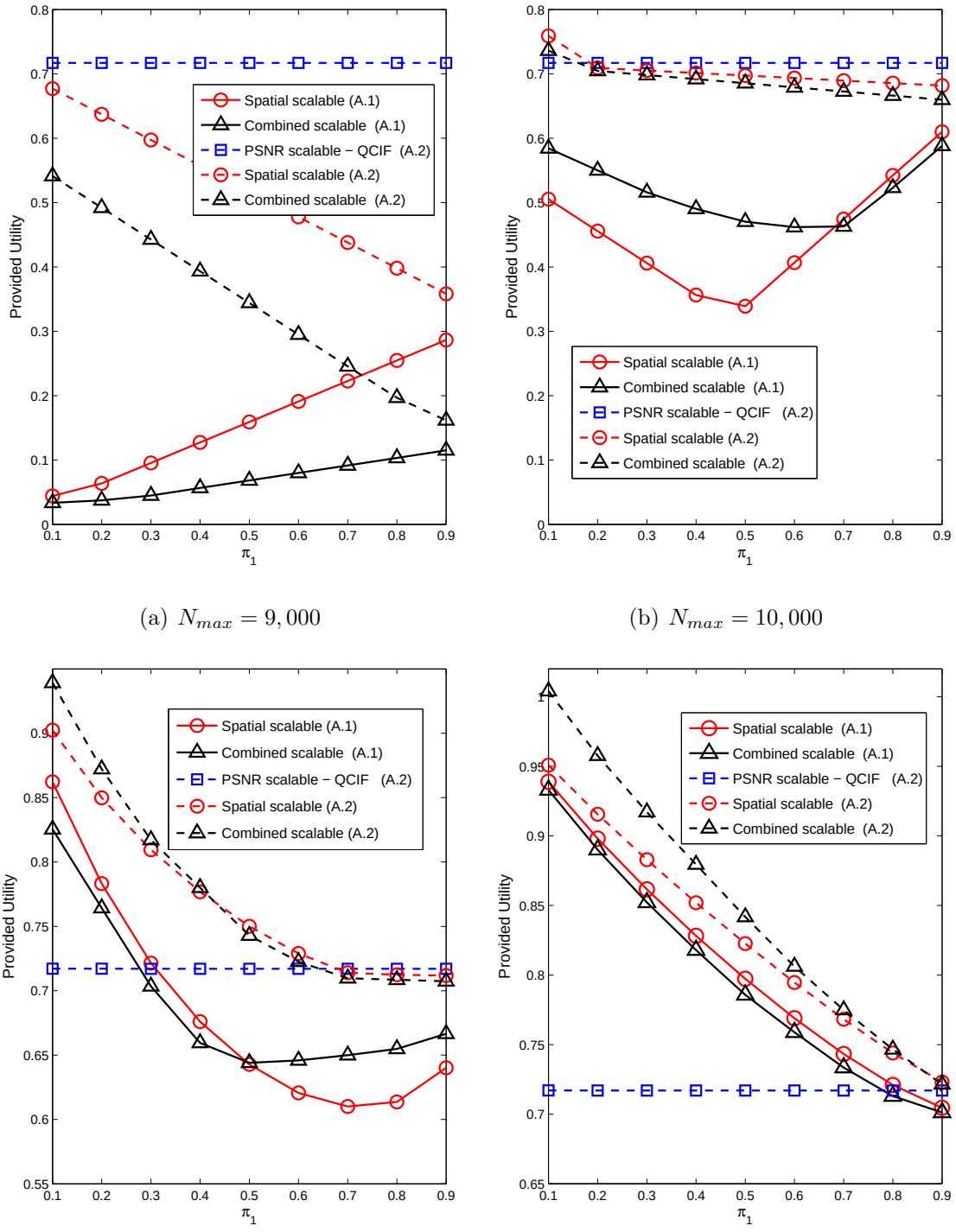
PSNR (in the desired spatial-temporal resolution) is lower with PSNR-scalable coding than multi-scalable coding. On the other hand, the small difference in bit rates may not be significant when the bit rates are high. This can be justified by studying the rate-distortion curves of the new video coders such as H.264/SVC [1]. Therefore, while acknowledging this fact, we neglect it in the proposed single-scalable bitstreams. To have a fair comparison, the outage probability constraints for the single-scalable streams are set to $P_{out}^{j} = \{10^{-4}, 10^{-3}\}$ vs. $P_{out}^{j} = \{10^{-4}, 3 \times 10^{-4}, 6 \times 10^{-4}, 10^{-3}\}$ for the combined-scalable stream. Client RCs of two different classes are modeled by $\Delta_1 \sim f(\delta; 0, 1, 0.65, 0.2)$ and $\Delta_2 \sim f(\delta; 0, 1, 0.85, 0.075)$. Here, $f(\delta; t_1, t_2, \mu, \sigma), t_2 > t_1$ represents a truncated Gaussian distribution with mean μ and variance σ^2 which is defined over the interval $\delta \in [t_1 \;\; t_2]$. Class 1 and class 2 terminals incorporate QCIF (176×144) and CIF (352×288) displays respectively. By setting $W(m, j) = 1, \forall m, j$,

we assume user utility to be given, only by NMOS-measured media quality.

6.3.1 Scenario A.1 & A.2

In scenario A.1, class 1 terminals are considered to be only capable of decoding QCIF layers while class 2 terminals can potentially decode all layers of the media. On the other hand, we assumed that terminals can not perform any extra adaptation on the decoded video, e.g., upsampling or downsampling. Clients who fail to decode any layer that matches their device screen resolution may not display anything on their screens. This scenario aims to evaluate the performance of the multi-scalable video encoding when potential limitations on decoding capabilities of different devices are considered. Because of these restrictions, the spatial-scalable and the combined-scalable bitstream are investigated in this scenario. Other bitstreams can not serve both classes according to the assumed limitations.

Scenario A.2 is similar to A.1. The only difference is that a terminal in class 2 that manages to decode the transmitted media up to QCIF resolution is able to display the QCIF video in the middle of its CIF screen. Similar to A.1, no up-sampling is permitted in this scenario. Performance of PSNR-scalable (QCIF), spatial-scalable and combined-scalable bitstreams are investigated, since only these formats are beneficial for clients in both classes. The results are illustrated in Fig. 6.1. The utility obtained in A.2 is higher than the utility in A.1 for those bitstreams involved in both scenarios, a direct result of the extra flexibility assumed for CIF terminals in A.2. When transmission deadline $N_{max} = 10,000$, the PSNR-scalable QCIF bitstream yields the highest utility in A.2 for $\pi_1 > 0.2$. The reason is that although the utility provided to CIF terminals from this bitstream is relatively low, the source rate is also low compared to



(a) $N_{max} = 9,000$

(b) $N_{max} = 10,000$

(c) $N_{max} = 11,000$

(d) $N_{max} = 12,000$

Figure 6.1: Provided utility for different bitstreams and prior settings in scenario A.1 & A.2. $\pi_2 = 1 - \pi_1$

other bitstreams with embedded CIF layers. According to Table 6.1, the size of the QCIF stream is 1101 source symbols while the size of the combined-scalable streams is 8331. By considering $N_{max} = 10000$, we realize that the small size of the QCIF stream enables the system to provide the utility associated with the highest PSNR to terminals with RC δ of roughly more than $\frac{1101}{10000} = 0.11$. According to the proposed distributions, approximately more than 95% of clients in class 1 and nearly all clients in class 2 have a RC above $\delta = \mu_1 - 2 \times \sigma_1 = 0.65 - 2 \times 0.2 = 0.25$. In Fig. 6.1(a)-(d), utility of PSNR-scalable QCIF bitstream is not increased by increasing N_{max} because nearly all terminals have already taken full advantage of the content utility of this stream. Conversely, the provided utilities of spatial- and combined-scalable streams are increased by N_{max}. Almost for every prior settings when $N_{max} \geq 12000$, the QCIF stream yields the lowest utility. In both scenarios ad when N_{max} is not very small $N_{max} \geq 10000$, the utility provided by combined-scalable bitstream is only slightly lower than the spatial-scalable stream in the worst cases (less than 0.02 on a scale of 0 to 1). This small inefficiency is mainly due to the lower efficiency of the rateless codes for the combined-scalable stream. Although codes with same degree distribution are considered in the simulations, the rateless codes for the combined-scalable stream suffer from higher overhead due to their smaller length S_j, according to (5.4).

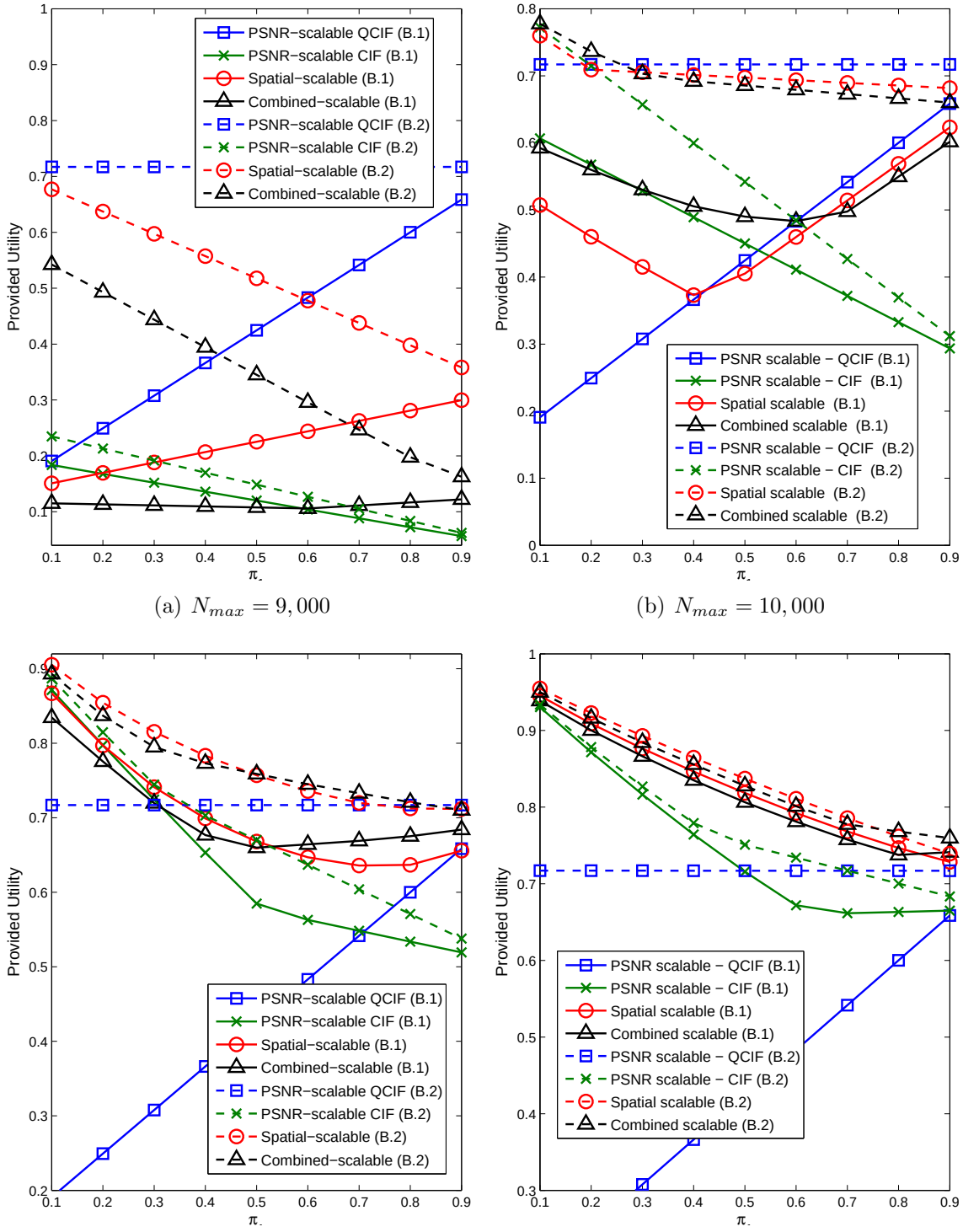
The combined-scalable bitstream provides significantly higher utility in various instances. For example in A.1 with $N_{max} = 10,000$ and $\pi_1 = 0.5$, the combined scalable stream results in 38.7% higher utility compared to the spatial-scalable stream. The reason is that the PSNR-scalable layers embedded in the combined-scalable stream enable the server to better tune the transmission rates of layers in the combined-scalable video to the distribution of clients in each class. For example, the system provides the CIF layer with the highest PSNR to the CIF terminals with high RCs

while some of the remaining terminals are provided the CIF layer with lower PSNR (or utilities associated with QCIF layers in scenario A.2). The optimal threshold $\delta_{min,4}$ differentiates CIF terminals with high RC ($\delta \geq \delta_{min,4}$) from the others.

6.3.2 Scenario B.1 & B.2

In these scenarios, terminals in both classes can potentially decode to the highest layer of the transmitted video bitstream ($h_m = J,\ \forall m$ where $J = 2$ in single-scalable streams and $J = 4$ in combined-scalable stream). In scenario B.1, terminals in each class adapt the video with respect to their displays. To ensure this adaptation, up-sampling (down-sampling) a decoded video of lower (higher) spatial resolution may be required in this scenario. In scenario B.2, CIF terminals are also permitted to display a QCIF video on their CIF screens, provided that higher perceptual quality is experienced in QCIF format. Unlike A.1 & A.2, these scenarios aim to explore the efficiency of multi-scalable video encoding when terminals with different screen sizes are capable of decoding all the layers of transmitted video and perform extra adaptations. Therefore, we expect the provided utility in these scenarios exceed the respective ones in A.1 & A.2 for every bitstream utilized in both.

The results are depicted in Fig. 6.2. When $N_{max} = 10,000$, the utility associated to transmission of combined-scalable media is very close to the highest achievable utility under different prior settings. In fact for $0.3 < \pi_1 < 0.6$, Combined scalable bitstream provides the highest utility in scenario B.1. Since priors reflect the population of clients in each class, the combined-scalable stream can be efficiently adapted to the population of multicast crowd. i.e., high utility is delivered without the need to add/remove spatial or fidelity layers in encoding stage in order to account for changes



(a) $N_{max} = 9,000$

(b) $N_{max} = 10,000$

(c) $N_{max} = 11,000$

(d) $N_{max} = 12,000$

Figure 6.2: Provided utility for different priors and streams (Scenario B.1 & B.2)

in population of clients in different classes. Similar to A.1, lack of PSNR-scalable layers prevents the spatial-scalable stream to be optimally adapted to the clients channel quality distributions when $N_{max} = 10,000$. The utilities of different bitstreams in scenario B.1 are higher than their counterparts in scenario A.1. the reason is that the same clients in B.1 can perform extra adaptations in their terminals and possibly increase their perceived quality experience. However, this increase in utility is not very significant for spatial-scalable and combined-scalable bitstreams. Considering the NMOS model parameters for this particular video ($b_f = 7.23, b_p = 29.68$ and $b_s = 8.84$) along with the rate-distortion data in Table 6.1, we realize that adapting the decoded video of an embedded QCIF layer to CIF resolution yields a very small PSNR. Consequently, the PSNR related term in (6.3) falls into the lower saturation region and a small NMOS is obtained. Therefore, there is a possibility that those extra adaptations do not contribute to the quality experience and the clients chose not to perform any extra adaptations on the video. In addition to that, this phenomena is also the main reason behind the difference between the corresponding curves in B.1 and B.2 for $N_{max} = 10,000$. Since the transmission deadline is not very large, most of the clients in both classes are unable to recover the embedded CIF layers. In B.2 CIF clients can display the the video in QCIF format. However, since the perceptual quality of adapting the video to QCIF resolution is more than quality of showing the video in CIF formats for these terminals (due to video parameters for this video), CIF terminals in B.2 experience higher perceptual quality compared to B.1. PSNR-scalable bitstreams of QCIF and CIF resolutions become more optimal when the prior π_m of their respective target terminals is increased.

6.4 Conclusions

In this chapter, the efficiency of multi-scalable encoded video in providing higher perceptual quality is investigated. The numerical results verified the embedded potentials of multi-scalable video along with the proposed optimization within the rateless multimedia multicast framework in delivering the highest quality experience in different scenarios. Compared to spatial-scalable video coding, multi-scalable video coding increased the provided perceptual quality experience by as high as 38.7%, according to the simulations. Based on the simulations, we can conclude that having both PSNR- and spatial scalable layers in video encoding stage can benefit us to achieve an optimal or close-to-optimal delivery of perceptual quality experience in video multicast. Unlike PSNR-scalable streams, spatial-, and more importantly combined-scalable streams can be tuned more accurately to the clients channel quality distributions regardless of the number of clients in different classes.

Chapter 7

Conclusions and Future Works

In this thesis, we have studied few applications of fountain codes in multimedia multicast and contributed to this field. In Chapter 3 we formulated the outage probability and obtained a novel closed-form equation for the outage probability with less computational complexity compared to the original formulation. Furthermore, we have proposed a new synthetic model that approximates the outage probability of fountain codes in memoryless erasure channel. Simplicity and accuracy of this model make it a suitable candidate for cross-layer optimizations. In Chapter 4, we derived problem formulations for packetized rateless multimedia multicast (PRMM) under two min-max cost criteria. By relaxing the integer constraints, we reformulated those problems and obtained convex optimizations. Analytical solution for those problems are also derived. Considering heterogeneity of clients' channels and their terminal capabilities, we introduced a new resource allocation and client classification framework in rateless multimedia multicast in Chapter 5. In the proposed optimization, users are dynamically assigned to various quality layers of the media based on channel quality distribution of clients and their terminal capability, in order to maximize the delivered

utility to the entire multicast population. The complexity of the proposed formulation is independent of the number of clients. This is one of the advantages of this framework compared to prior works. Moreover, by some simplifying assumptions, we managed to obtain convex optimization problems at the cost of less efficiency of the solution. However, the resulting inefficiency was negligible according to the simulations. Finally in Chapter 6, the proposed resource allocation scheme in Chapter 5 was employed to investigate the possible advantageous of using multi-scalable video coding in delivering the maximum perceptual quality experience among heterogeneous clients, in different scenarios. Our studies confirmed the benefits under certain conditions. These scenarios served as good examples by demonstrating how easily the proposed formulation in Chapter 5 can be adapted into different applications by means of setting proper connections between the utility coefficients and the optimization goal.

Further research can be conducted based on the contributions of this thesis. The closed-form outage probability and the proposed approximation can potentially improve the accuracy and/or reduce the complexity of the prior works in cross-layer design. Moreover, optimizations studied in this thesis are entirely devoted to single server broadcast scenarios. Server may broadcast the same content to heterogeneous receivers using multiple antennas. Therefore, the suggested optimization techniques can be generalized to investigate possible diversity gain in application layer. Further research on obtaining jointly defined optimization tasks wherein some of clients are treated as primary users with rigid QoS- and utility constraints while secondary users are flexible with their target media quality is also promising. Also, proposed optimizations can be implemented in laboratory-scale WLAN to investigate their performance more accurately based on the real wireless channels.

Bibliography

[1] Julien Reichel, Heiko Schwarz, and Mathias Wien, "Joint scalable video model JSVM-4," Doc. JVT-Q202, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Nice, France, Oct. 2005.

[2] J. Li, "A progressive to lossless embedded audio coder (PLEAC) with reversible modulated lapped transform," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, Apr. 2003, vol. 5, pp. V – 413–16 vol.5.

[3] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[4] M. Mitzenmacher, "Digital fountains: a survey and look forward," in *IEEE Information Theory Workshop*, Oct. 2004, pp. 271 – 276.

[5] Digital Video Broadcasting (DVB), "Oen 302304 v1.1.1: Transmission system for handheld terminals (DVB-H)," Tech. Rep., ETSI, Nov. 2004.

[6] 3rd Generation Partnership Project, "Technical specification group services and system aspects; multimedia broadcast/multicast service (MBMS);protocols and codecs," Tech. Rep., 3GPP TS 26.346 V10.0.0, Apr. 2011.

[7] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 1737 –1744, Nov. 1996.

[8] A.E. Mohr, E.A. Riskin, and R.E. Ladner, "Unequal loss protection: graceful degradation of image quality over packet erasure channels through forward error correction," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 819 –828, June 2000.

[9] P.A. Chou and Zhourong Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 390 – 404, Apr. 2006.

[10] N. Rahnavard, B.N. Vellambi, and F. Fekri, "Rateless codes with unequal error protection property," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1521 –1532, Apr. 2007.

[11] D. Sejdinovic, D. Vukobratovic, A. Doufexi, V. Senk, and R. Piechocki, "Expanding window fountain codes for unequal error protection," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2510 –2516, Sept. 2009.

[12] D. Vukobratovic, V. Stankovic, D. Sejdinovic, L. Stankovic, and Zixiang Xiong, "Scalable video multicast using expanding window fountain codes," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1094 –1104, Oct. 2009.

[13] U.C. Kozat and S.A. Ramprashad, "Unequal error protection rateless codes for scalable information delivery in mobile networks," in *26th IEEE International Conference on Computer Communications (INFOCOM'07)*, May 2007, pp. 2316 –2320.

[14] W. Sheng, W.-Y. Chan, S. D. Blostein, and Y. Cao, "Asynchronous and reliable multimedia multicast with heterogeneous QoS constraints," in *IEEE International Conference on Communications (ICC '10)*, May 2010, pp. 1 –6.

[15] Y. Cao, S. Blostein, and W. Y. Chan, "Unequal error protection rateless coding design for multimedia multicasting," in *IEEE International Symposium on Information Theory Proceedings (ISIT '10)*, June 2010, pp. 2438 –2442.

[16] Y. Cao, S. Blostein, and W. Y. Chan, "Optimization of rateless-coded asynchronous multimedia multicast," in *International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC '11)*, Sept. 2011.

[17] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, Aug. 1991.

[18] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, pp. 379–423, July 1948.

[19] D.J.C. MacKay, "Fountain codes," *IEE Proceedings-Communications*, vol. 152, no. 6, pp. 1062 – 1068, Dec. 2005.

[20] M. Luby, "LT codes," in *IEEE Symposium on Foundations of Computer Science*, 2002, pp. 271 – 280.

[21] A. Shokrollahi, "Raptor codes," in *IEEE Information Theory Workshop on Information Theory for Wireless Networks*, July 2007, p. 1.

[22] R. Gallager, "Low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21 –28, Jan. 1962.

[23] M. Luby, M. Watson, T. Gasiba, T. Stockhammer, and Wen Xu, "Raptor codes for reliable download delivery in wireless broadcast systems," in *3rd IEEE Consumer Communications and Networking Conference (CCNC '06)*, Jan. 2006, vol. 1, pp. 192 – 197.

[24] C. Harrelson, L. Ip, and W. Wang, "Limited Randomness LT Codes," in *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, 2003.

[25] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600 –612, Apr. 2004.

[26] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312 – 322, Sept. 2004.

[27] A. Rossholm and B. Lovstroem, "A new low complex reference free video quality predictor," in *IEEE 10th Workshop on Multimedia Signal Processing*, Oct. 2008, pp. 765 –768.

[28] Telecommunication standardization sector of ITU, "Objective perceptual multimedia video quality measurement in the presence of a full reference," Tech. Rep., International Telecommunication Union, August 2008.

[29] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660 –668, Sept. 2008.

[30] M. Luby, T. Gasiba, T. Stockhammer, and M. Watson, "Reliable multimedia download delivery in cellular broadcast networks," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 235 –246, Mar. 2007.

[31] A. Papoulis and S. Phillai, *Probability, Random Variables, and Stochastic Processes, 4th Edition*, McGraw-Hill, 2011.

[32] M. Chiani, D. Dardari, and M.K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 840 – 845, July 2003.

[33] G.T.F. de Abreu, "Jensen-Cotes upper and lower bounds on the Gaussian Q-function and related functions," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3328 –3338, Nov. 2009.

[34] W. Sheng, W. Y. Chan, and S.D. Blostein, "Rateless code based multimedia multicasting with outage probability constraints," in *25th Biennial Symposium on Communications (QBSC '10)*, May 2010, pp. 134 –138.

[35] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," `http://cvxr.com/cvx`, Apr. 2011.

[36] S. McCanne, M. Vetterli, and V. Jacobson, "Low-complexity video coding for receiver-driven layered multicast," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 983 –1001, Aug. 1997.

[37] D.G. Sachs, R. Anand, and K. Ramchandran, "Wireless image transmission using multiple-description based concatenated codes," in *Data Compression Conference, Proceedings*, 2000, p. 569.

[38] W.-H. Kuo and W. Liao, "Utility-based radio resource allocation for QoS traffic in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2714 –2722, July 2008.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[40] Y. Xue, Y. F. Ou, Z. Ma, and Y. Wang, "Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts," in *18th International Packet Video Workshop (PVW '10)*, Dec. 2010, pp. 201 – 208.

[41] Y. F. Ou, Z. Ma, T Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 286 –298, Mar. 2011.