PRIORITY AND PATH DIVERSITY BASED VIDEO STREAMING OVER BEST EFFORT NETWORKS

by
Dana Farouq Hussein

A Thesis Presented to the Faculty of the
American University of Sharjah
College of Engineering
in Partial Fulfillment of
the Requirements
for the Degree

Master of Science in
Electrical Engineering

Sharjah, United Arab Emirates
July 2014

Approval Signatures

We, the undersigned, approve the Master's Thesis of Dana Hussein.

Thesis Title: Priority and Path Diversity Based Video Streaming over Best Effort Networks

Signature	**Date of Signature** (dd/mm/yyyy)
_____	_____
Dr. Mohamed Hassan Associate Professor, Department of Electrical Engineering Thesis Advisor	
_____	_____
Dr. Taha Landolsi Associate Professor, Department of Computer Science & Engineering Thesis Co-Advisor	
_____	_____
Dr. Mohamed El-Tarhuni Professor, Department of Electrical Engineering Thesis Committee Member	
_____	_____
Dr. Mohamed Saad Associate Professor, Department of Electrical and Computer Engineering University of Sharjah Thesis Committee Member	
_____	_____
Dr. Mohamed El-Tarhuni Head, Department of Electrical Engineering	
_____	_____
Dr. Hany El Kadi Associate Dean, College of Engineering	
_____	_____
Dr. Leland Blank Dean, College of Engineering	
_____	_____
Dr. Khaled Assaleh Director of Graduate Studies	

Acknowledgements

I would like to thank my thesis advisors, Dr. Mohamed Hassan and Dr. Taha Landolsi, for their support and guidance through the completion of this thesis. I would also like to thank Dr. Tamer Shanableh for his help and contribution in the video quality work. I am thankful to the Electrical Engineering department at the American University of Sharjah for providing me with the opportunity to be part of their graduate program. Last but not least, I would like to deeply thank my parents, family and fiance for the endless support and compassion.

To my family. . .

Abstract

Successful delivery of important parts of video sequences is expected to maintain continuous playback at an acceptable video quality. In this thesis, priority and path diversity-based schemes are proposed in order to enhance the viewer's experience by reducing the delay and loss rates of important video packets that might occur due to routing to a congested path. The introduced schemes classify video frames into three priority levels based on their importance. In addition, the proposed schemes maintain a priority queuing system at each router to reduce the delay experienced by video frames with high priority and hence to reduce their probability of dropping. The performance of the proposed schemes is evaluated through simulations in comparison to each other as well as to a best effort-based model. A priority-based queuing model is introduced in which each router maintains a priority queuing system without the use of path diversity. Moreover, we introduced a priority-based path diversity (PBPD) scheme that integrates the priority-based queuing model with path diversity. The performance of the three models is studied in terms of network related metrics such as frame loss rates and end-to-end delay. Moreover, the performance of the three models is compared from the end users' perspective in terms of the peak-signal-to-noise-ratio (PSNR). In addition, a buffer occupancy metric is used to study the continuity of the reconstructed video. The results show that video frames experienced the lowest frame loss rates when transmitted using the proposed PBPD scheme. In addition, the PBPD scheme outperforms the priority-based queuing model and the best effort-based model in terms of the end-to-end delay. These results are reflected in the reduced number of starvation instants achieved by the proposed PBPD scheme at relatively higher PSNR values.

Search Terms—Path diversity, video streaming, priority queuing, quality of service, best effort networks, video quality.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

APD	-	Avalanche Photo Detector
ASE	-	Amplified Spontaneous Emission
BER	-	Bit Error Ratio
BPSK	-	Binary Phase Shift Keying
DLI	-	Delay Line Interferometer
EDFA	-	Erbium-Doped Fiber Amplifier
FEC	-	Forward Error Correction
FSO	-	Free-Space Optical
FP	-	Fabry-Pérot
FWHM	-	Full-Width Half Maximum
GEO	-	Geostationary Earth Orbit
HAP	-	High Altitude Platform
IM-DD	-	Intensity Modulation-Direct Detection
LED	-	Light Emitting Diode
LEO	-	Low Earth Orbit
NRZ	-	Non-Return-to-Zero
OOK	-	On-Off Keying
PBS	-	Polarization Beam Splitter
PDM	-	Polarization Division Multiplexed
POLSK	-	Polarization Shift Keying
PPB	-	Photons Per Bit
PPM	-	Pulse Position Modulation
PQ-PPM	-	PDM-QPSK PPM
QPSK	-	Quadrature Phase Shift Keying

RZ	-	Return-to-Zero
SNR	-	Signal-to-Noise Ratio
SWaP	-	Size, Weight, and Power
UAV	-	Unmanned Aerial Vehicles

Chapter 1: Introduction

1.1 Video Streaming over Best Effort Networks: An Overview

Today's Internet is witnessing a widespread deployment of multimedia applications such as video on demand and videoconferencing. With the increasing popularity of multimedia applications, video streaming is gaining more interest. Video streaming has strict quality of service (QoS) requirements. Even with the advances in video compression techniques, video streaming applications still require high bandwidth. In addition, video streaming applications are highly sensitive to end-to-end delay and delay jitter (delay variation) especially in real-time video applications. Delivering media content with strict QoS requirements is challenging especially when it comes to lossy packet networks such as the Internet.

Early media streaming systems over the Internet, such as VivoActive introduced in the 1970 used Hyper Text Transfer Protocol (HTTP) web servers to store and deliver media content [1]. HTTP server-client employs the reliable transport control protocol (TCP) which introduces additional delays to ensure reliability. However, these additional delays might affect the performance of real time applications such as media streaming applications. Moreover, the TCP protocol uses congestion control mechanisms which results in fluctuations in the available data rate [2]. To compensate for these drawbacks, the early VivoActive program used a large pre-roll buffer to buffer a sufficient number of frames before the start of video playback [1].

In 1995, the first streaming media system, which is based on the multimedia requirements and the underlying Internet, was RealAudio 1.0. RealAudio system used both TCP and user datagram protocol (UDP) in delivering the media content, which is known as the progressive networks architecture (PNA). RealAudio systems used the UDP protocol to deliver media content while maintaining a TCP connection for session control [1]. Transmitting media content over the UDP protocol made the transmission of the media content more continuous compared to the TCP protocol. However, the

unreliability of the UDP protocol caused several problems such as out of order delivery of packets and packet loss. To accommodate these drawbacks, several techniques have been proposed including variable bit rate (VBR) video compression and error control mechanisms [3], [4]. Video content has a dynamic nature which depends on the scene change and the motion introduced in the video. Because of this dynamic nature, the distribution of bits between frames is going to be different when compressed at the same level of distortion. This has led to the introduction of VBR video compression in the design of RealVideo codec [1]. VBR uses the dynamic nature of the video in producing a constant quality encoded video. RealVideo has also introduced a variety of forward error correction (FEC) codes and error concealment mechanisms to protect the most important parts of the encoded video [1].

1.2 Limitations of Best Effort Networks

Best effort networks with no traffic engineering do not offer deterministic or statistical QoS guarantees. The current level of service provided by today's Internet might not be sufficient to support the stringent QoS requirements of video streaming applications such as video on demand and video conferencing. As a result, video streaming applications are hampered by the widely varying packet loss rates as well as variations in delay and throughput caused by congestion at intermediate routers.

An example of a very famous and fast growing video delivery website is YouTube. YouTube aims to provide the QoS requirements of video over best effort networks by employing an overlay of content delivery networks which are close enough to end users [5]. Moreover, YouTube relies on large bandwidth usage to provide an acceptable QoS level for video streaming. A study conducted in 2008 has shown that YouTube consumed the same amount of bandwidth as that which was consumed by the entire Internet in year 2000. It was also estimated that YouTube had to pay roughly 1 million dollar a day for the large bandwidth consumption of its servers [6]. With these efforts and with the high speed access provided in academic institutions or corporations, users have a relatively acceptable video streaming experience. However, residential users

might suffer from lower video quality in which users experience long buffering delays and interruptions in the video playback [7].

1.3 Contributions

The contribution of this thesis is as follows:

a. A priority-based path diversity (PBPD) scheme is proposed for simple network topologies in which multiple paths exist between a video server and an end-user. The proposed scheme aims to improve the QoS requirements of video streaming applications in terms of delay and losses by forwarding video packets to the lowest congested path in a priority basis. The proposed scheme relies on two main aspects. First, the Internet is highly redundant in terms of the number of available paths between two communicating hosts and this redundancy is not fully exploited by the Internet's routing protocols. Second, video packets have different importance levels as they do not evenly contribute to the reconstructed video quality. The proposed scheme exploits the Internet redundancy by jointly considering the importance level (priority level) of video packets and the congestion among the candidate paths when forwarding video packets.

b. The proposed PBPD scheme is generalized for larger network topologies. Unlike the simple PBPD scheme which is based on the direct feedback from the network routers to the ingress router (i.e. the router directly connected to the video source), the generalized scheme forwards video packets to the least congested path by exchanging congestion state information between the routers in the network. This allows routers to select the least congested path for a transmitted video packet.

1.4 Thesis Organization

The rest of this thesis is organized as follows: In Chapter 2, a background that covers the network layer functionalities is presented in addition to several technical concepts which are directly related to the presented work. This includes introduction to video compression, scalable video coding, its operating modes, and path diversity. Chapter 3

discusses the work done in the area of video streaming over best effort networks, including: application layer-QoS control, TCP friendly-congestion control and network layer-QoS support. Chapter 4 addresses a detailed network performance evaluation of the proposed PBPD scheme in comparison to two streaming approaches (priority-based queuing approach and best effort approach). Chapter 5 studies the impact of the different routing schemes on the delivered video quality. Finally, chapter 6 concludes the work presented, summarizes the main findings and discusses our future work.

Chapter 2: Background

In this chapter, an overview of the Internet is presented, focusing on the network layer functionalities. Later, several technical concepts that are directly related to the presented work such as video compression, scalable videos, and path diversity will be illustrated. At the end, a focus on the idea of path diversity, its issues and use in multimedia communication will be presented.

2.1 Overview of the Internet

The Internet evolved into a universal network which delivers new services and connects different users throughout the world. The network is composed of millions of connected computing devices, which were initially limited to desktops and large data storage devices called servers. Web pages and video sequences are examples of the kind of data which is stored in these server devices for transmission. Today however, the Internet is witnessing a large deployment of nontraditional connected devices such as cell phones, television and gaming consoles.

End computing devices which run applications and transmit/receive data are called edge nodes. Other than edge nodes, the Internet is further composed of core nodes responsible for data delivery which are identified as routers and switches. The nodes can be connected using different types of communication links such as radio channels, coaxial cables and fiber optics. Packets are the units to transfer information over the Internet. Packets sent from one end host to the other are delivered with the aid of packet switching techniques, which are efficient alternatives to avoid dedicating physical paths for connecting any two hosts in the network.

A set of protocols govern the communication process between the end hosts. The protocols control the data format and the actions taken by the network during transmission, reception and any other event. The Internet follows a layered model, where each layer is responsible for controlling specific tasks to enable the communication process between end hosts. The upper layers of the system deal with the applications

and represent the direct interface to end users. Lower layers however, transform the abstracted data of the upper layers into a more suitable form that can be transmitted over the physical links. Open Systems Interconnection (OSI), the technical description for the Internet, consists of seven layers namely: application layer, presentation layer, session layer, transport layer, network layer, data link layer and physical layer.

Network Layer

Once the end host initiates a data transmission process, the data (in the form of packets) is passed through the transport layer to the network layer which is responsible for delivering packets between end hosts.

The packets travel inside the network through devices called routers. Each router in the network is connected to multiple input and output links. The network layer controls the forwarding functionality of routers which is responsible for passing the incoming data packets from an input link to the appropriate output link based on the destination address. In the forwarding process, the router relies on an information set known as the forwarding table which maps destination addresses to the correct output link. The collective interaction of all the routers in the network to define a particular path linking the source and destination end hosts is known as routing; the routing process is performed using a set of commands known as the routing protocols.

Routers

As discussed earlier, the main task of routers is to forward packets from their input links to their appropriate output links. Shown in Figure 2.1 is a typical router architecture which consists of input ports, output ports and intermediate units responsible for the routing process. The intermediate units are namely the switch fabric and the routing processor. As packets arrive from the different input links, they are queued in the input ports until the intermediate routing units are ready for data transmission. In the transmission process, the routing processor runs the routing protocol and maintains the

forwarding table which is used by the switch fabric in forwarding the packets from the input ports to the corresponding output ports. Once a packet reaches the output port, it is stored to be sent later to the corresponding output links.

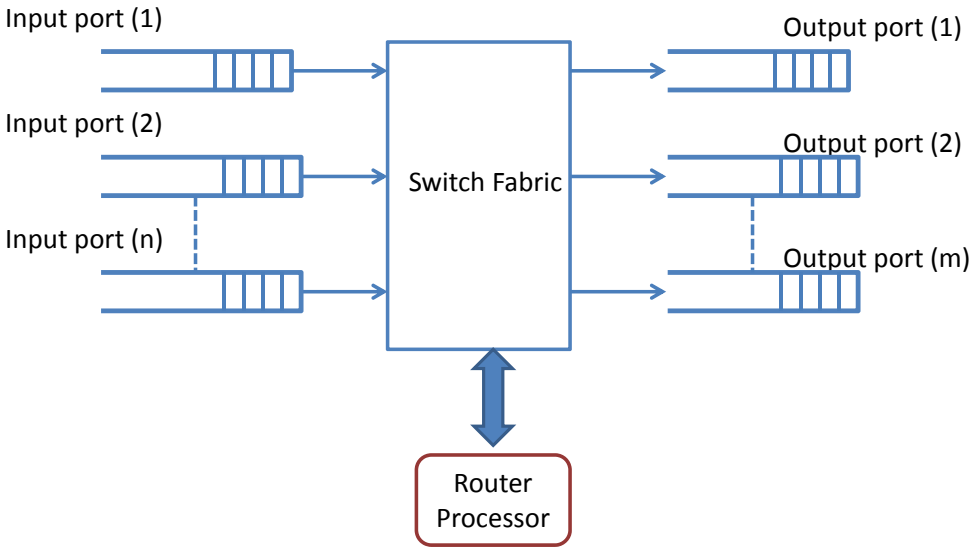


Figure 2.1: A simplified router architecture.

As the data is continuously being sent and received, the incoming and outgoing packets to the input/output ports of the routers will start forming queues. If the packets start accumulating in the queues, the storage capacity (also called the buffer space) might exceed resulting in packet loss due to dropping. Packet loss occurs when the incoming packet traffic rate is larger than the switch fabric speed, resulting in the packets having longer queuing delay and eventually getting dropped at the input queues as the buffer capacity is full. Similarly, if the incoming traffic rate to an output port is larger than the output link speed, packets will experience longer queuing delay and will more likely be dropped at the output port. Selection of a packet for transmission within the queue at the output port is done by a packet scheduler. The scheduler might be a simple First-Come-First-Serve (FCFS), in which packets are transmitted on the basis of arrival relative to other packets in the queue of the port. The scheduler can also be a Weighted Fair Queuing (WFQ) scheduler which is not concerned with the order of arrival of the

packets, but rather allows the packets to equally share the output link. In the literature, packet scheduling is proposed for future Internet to serve packets based on the priority that ensures better QoS guarantees [8], [9].

Furthermore, incoming packets to the input ports of a router have to be well managed with the buffer capacity condition. When an arriving packet overflows the buffer's capacity, a decision must be made by the router to either drop the arriving packet (also known as Drop-Tail decision) or to drop an already queued packet. Another method of managing the buffer space involves marking or dropping incoming packets before the buffer is full. This method is known as Active Queue Management (AQM), and it includes different policies which have been widely studied in the literature. One of the well-known AQM policies is Random Early Detection (RED), in which arriving packets are either accepted, dropped or marked at the queue based on the average queue length. When a packet arrives, the average queue length is compared with pre-defined threshold values called the min_{th} and max_{th}. When the average queue length is less than min_{th}, an arriving packet is accepted in the queue. Conversely, when the average queue length is larger than max_{th}, the packet is either dropped or marked for dropping. If the average queue length is between min_{th} and max_{th}, an arriving packet will be either accepted, marked or dropped with a probability that is a function of the threshold values.

Management of incoming and outgoing data packets at the input and output ports is different in concept. Scheduling schemes for the output ports are employed to serve or control the outflow of different packets. On the other hand queue management schemes at the input ports are designed for a router to decide on which packets to be dropped and when to drop them based on the condition of the buffer's storage capacity [10].

Routing Protocols

Routing refers to the process of finding the best path through which a host in a network sends packets to the least cost path. In a typical mesh network, such as the Internet,

there are multiple paths that can be used to communicate information between two destinations (hosts) amongst which a path is selected as the best path. The selection of the best path can be done based on the minimization of a cost function which considers the characteristics of the communication links or simply the number of communication links along the path. The characteristics of the communication links, that can be included in the cost function, are the expected latency, the available bandwidth and the reliability over the entire path. The rules that govern the choice of the best path are referred to as a routing protocol. Routing protocols used in today's Internet can be classified according to three aspects: how they govern the implementation of the routers forwarding tables, what information is used and shared between routers, as well as the location of the source and destination within a network.

Additionally, routing can be performed by utilizing a management center which finds the best path based on a certain protocol and downloads it to the routers in the network. The routers will then do the forwarding through the path selected by the management center. Alternatively, routing in today's Internet utilizes a distributed method. In distributed routing, each router acts as a centralized forwarding decision making entity within a larger network of other routers. The selection of the best path is done by a series of those centralized forwarding decisions along the path in the network. The decisions are done using a certain routing protocol. The routing process aims at interchanging network information between routers so that eventually each router knows at least the next hop correspondent to the best path to each network destination.

Routing protocols are classified as static or dynamic protocols depending on how the forwarding tables of the routers in the network are loaded [2]. In static routing protocols, the forwarding tables and any updates in the network (e.g.: when additional links are implemented or dismissed in the network) are managed and loaded directly from the administrator of the network to all routers in the network. On the other hand, in dynamic routing protocols, forwarding tables and updates are exchanged between neighboring routers. In more details, routers periodically receive forwarding tables

from directly connected routers, update their forwarding tables accordingly and send the updated forwarding table back to the directly connected routers.

Routing protocols can also be classified into link state routing and distance vector routing based on the information routers have and share. In link state routing, each router has full knowledge of the network topology and the cost of all the links in the network. In distance vector routing, each router only knows the cost to the directly connected routers. Moreover, in link state routing, routers share all the hops included in the best path. However, in distance vector routing, routers only know and share the next hop corresponding to the best path. Another way of classifying routing protocols can be based on the location of the source and destination within the global network and can be categorized into two main categories: inter-autonomous system and intra-autonomous system routing protocols [2]. An autonomous system (AS) is a network or a group of networks in which one administrator governs the network and all routers run the same routing protocol. The inter-autonomous system routing protocols are the protocols used in routing packets between AS; on the other hand, intra-autonomous routing protocols are the protocols used in routing within an AS. Currently, Routing Information Protocol (RIP) and Open Shortest Path First (OSPF) are the most commonly implemented protocols. Both of these protocols are intra-autonomous protocols; however, they differ in the information included in their forwarding table.

RIP is an example of distance vector routing protocols in which routers only know the next hop corresponding to the path with the lowest cost. Initially, each router has the cost of the directly connected links in their forwarding tables along with an initial estimate of the cost of the best path from the router to all other destinations in the network. Neighboring routers will then exchange their forwarding tables. This will allow each router to compare the cost of the paths in its forwarding table with that found in the received forwarding tables. Using this information, the forwarding tables are updated with a better estimate of the best path. This process is repeated every 30 seconds to ensure that the best path is periodically updated [2].

To further illustrate the distance vector concept of the RIP protocol, consider the network model shown in Figure 2.2. Suppose that router A initially estimates the best path to destination G to be through router B. When router A receives the forwarding table from its neighboring routers (B and C), it will compare the cost of the path entries of its forwarding table to that of the received forwarding tables. Suppose the cost of the path to destination G through router C is lower than that of the path through router B; then router A will update its forwarding table such that the path to destination G is best through router C. Since this is a distance vector protocol, router A will only know the next hop (router C) of the best path, but it has no knowledge of whether the best path after router C will be through router D or router F.

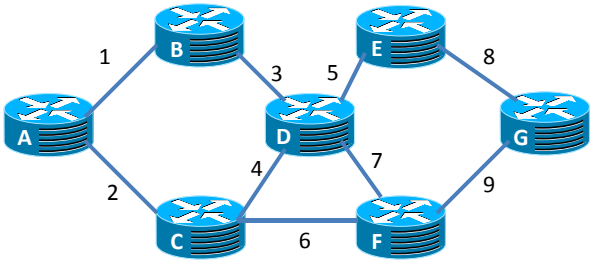


Figure 2.2: Network model (example).

RIP is a simple routing protocol which does not require sophisticated routing algorithms. However, it has some undesired characteristics such as the count to infinity problem. To illustrate the count to infinity problem, refer to the network shown in Figure 2.2. Suppose the two links 2 and 3 break, router A and router B are now isolated from the network. Suppose router A notices the link failures and updates its forwarding table such that the distance to all routers is infinity except the distance to router B. When router B receives the distance vector of router A, it will update its routing table and will set the cost to all other hops through router A as infinity. Router B will realize the failure and will understand that both routers A and itself are now isolated from the network. This information is stored in the forwarding table of router B which will then broad-cast to router A to update its tables and consequently, the algorithm converges.

Now, suppose router B -before detecting the link failure- the forwarding tables assumed that all paths through router A are possible. At the same time, router A had already detected the failure but was not able to send it to router B at the right instance. It rather received the forwarding tables of router B and updated its own tables with the wrong information that it received. Then, router A will mistakenly realize a path exists to all other routers through router B and sends its last updated tables to router B which will in turn mistakenly realize that a path exists to all other routers through router A. As a result, packets will keep bouncing between routers A and B creating congestion in these links. To solve this issue, RIP attaches a counter to each data packet that increments each time it passes through a router. By setting a maximum allowable number of hops each packet can pass through (usually it is set to 15), a data packet with a counter that reached this number will be dropped and its destination will be set as infinity (unreachable). This will reduce the congestion in the link and will update the tables of the routers accordingly and the algorithm will converge. Because of such undesired characteristics of RIP, it is mostly used in small networks. A more complicated link state algorithm such as OSPF is preferred for larger networks.

Unlike RIP, OSPF is an example of a link state routing protocol in which routers know the arrangement of the routers in the network (network topology) and the cost of all the links. This requires the network administrator to manually upload the map of the network topology and the cost of the links to each router. Also the updating of the routers forwarding tables is done by broadcasting the link state information at regular intervals as well as whenever a change on the link state is detected such as a link failure like the one described in the earlier example. By considering the same failure scenario in the network of Figure 2.2 when links 1 and 2 are broken, the difference between RIP and OSPF is evident. If RIP is used, there is a chance that the algorithm converges after the count to infinity is detected, but in the case of OSPF, convergence occurs without the need for detecting the count to infinity event. This is because, the OSPF protocol broadcasts the link state information immediately when a change occurs in the network. Furthermore, the OSPF protocol time stamps each broadcasted link state information

which will not allow older updates to be implemented in the routers forwarding tables. This will help to overcome the problem faced in the case of RIP when data packets keep bouncing between routers A and B when their forwarding tables were mistakenly updated.

2.2 Video Compression

A video is a sequence of still images (frames) displayed at a certain frame rate (frames per second) to perceive a continuous and natural object's motion by the human eye. The immense amount of data in a raw digital video results in video data rates that are high enough to overwhelm networks' links. Hence, practical video communication systems rely on video compression. Video compression standards such as MPEG and H.26x are based on two main principles: reducing spatial redundancy by intra-frame coding and reducing temporal redundancy by inter-frame coding.

In intra-frame coding, video frames are compressed in the same way as a still image. This is performed by dividing the still image (frame) into 8×8 pixel blocks and transforming them using the discrete cosine transform (DCT) to 8×8 coefficients. These 8×8 DCT coefficients are then quantized by an adjustable 8×8 quantization matrix. The quantization matrix is obtained by multiplying a base quantization matrix by a quantization scale which can be adjusted to control the source rate of a video encoder. A large quantization scale results in a coarser quantization and hence a smaller video size. However, this will come at the expense of downgrading the quality of the video. The quantized DCT coefficients are zigzag scanned then run-length coded.

MPEG inter-frame coding reduces the temporal redundancy by using motion estimation and motion compensation techniques between consecutive frames. Consecutive video frames are similar in the sense that an object or a scene that is presented in one frame is likely to be present in the following frame or in the following frames depending on the motion content of the video sequence. A low motion video (e.g. news broadcast) will have more similar video frames, therefore, higher temporal redundancy. To compress video frames in the temporal dimension, three types of frames

are introduced, namely intra-coded (I), predictive-coded (P) and bidirectional-coded (B) frames. These frames are structured in a periodic pattern called Group of Pictures (GoP). The sequence of frames between two consecutive I-frames corresponds to one GoP. The number of P-frames and B-frames present in the GoP defines the GoP structure as shown in Figure 2.3.

Intra-coded frames (I-frames) are independently coded using image coding techniques. In P-frames, a macroblock which consists of four blocks of 8×8 samples is inter-coded with a forward reference to a preceding I-frame or P-frame. Macroblocks of B-frames are inter-coded by bidirectional referencing to a backward I-frame or P-frame and a forward I-frame or P-frame. The inter-coding of macroblocks is done using motion estimation and motion compensation techniques. Given a macroblock, motion estimation identifies the best matching macroblock in the reference frame using motion vectors. The difference between the macroblock to be encoded and its best match in the reference frame is encoded using DCT and quantization as discussed earlier. This process is called motion compensation [11].

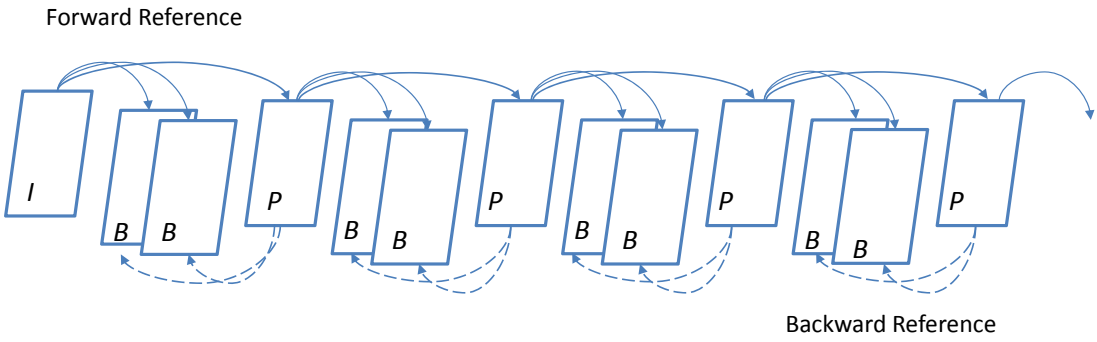


Figure 2.3: A typical MPEG group of pictures (GoP).

2.3 Scalable Video Coding (SVC)

Today's networks and applications are of a heterogeneous nature. On the one hand, networks provide variable bit rates to end users. In addition, links with different speeds

exist in the same delivery chain causing the heterogeneous network to be bottlenecked by the link with the lowest bandwidth. On the other hand, video display devices have diverse video formats requirements. This heterogeneity of application requirements and underlying networks is a challenge when it comes to streaming a single layer compressed video. A possible solution to this problem could be to encode the video into different versions each of a different encoding quality. However, a major drawback of this solution is the significant increase in the storage requirements. Moreover, in heterogeneous networks, different versions of the video stream need to be transmitted in adaptation of network bandwidth variations.

To overcome these drawbacks, scalable video coding was proposed. Scalable video coding provides different video formats at different video bit rates using a single encoded video [12]. A layered video encoder encodes the video into two or more layers which include a base layer and one or more enhancement layers. The base layer is independently encoded and provides a basic video quality. On the other hand, the enhancement layer is encoded in reference to the base layer and when received and decoded correctly, it enhances the quality of the reconstructed video. In what follows, we discuss common scalability modes namely, temporal scalability, spatial scalability and multiple description coding.

2.3.1 Temporal Scalability. Temporal scalability is one of the main scalability modes standardized by MPEG which provides an encoded video with adjustable frame rate (frames/sec) [12]. In this mode, the base layer provides a basic video quality at a lower frame rate where the addition of enhancement layer will increase the frame rate of the received video. This is done by interleaving the enhancement layer frames with the base layer frames.

One way to perform temporal scalability is shown in Figure 2.4. The base layer consists of I-frames and P-frames and the enhancement layer consists of B-frames. As discussed earlier, B-frames are encoded with reference to I-frames and P-frames. Therefore, the enhancement layer is encoded with reference to the base layer. On the

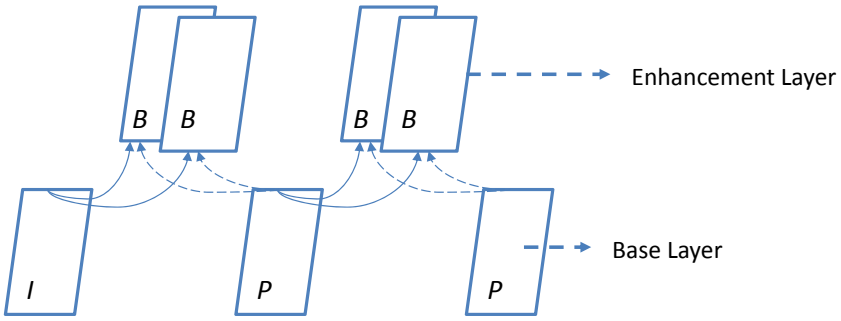


Figure 2.4: An example of temporal scalable encoding.

other hand, base layer is independently encoded. Base layer consists of independently coded (I-frames) and P-frames which are encoded with reference to other base layer frames (I-frames or P-frames) [12].

2.3.2 Spatial Scalability. Spatial scalability is another scalable coding mode standardized by MPEG which provides an encoded video with adjustable video format. In spatial scalable mode, the video is encoded into a base layer with a small spatial format (e.g.: Quarter Common Intermediate Format (QCIF)) and one or more enhancement layers with a large spatial format (e.g.: Common Intermediate Format (CIF)) [1].

To accomplish the above task, first the video is down sampled and encoded to represent the base layer. To obtain the enhancement layer, the base layer is locally decoded at the encoder and up-sampled. The difference between the up-sampled base layer and the original video is then encoded using the DCT based compression which will result in the enhancement layer. As shown in Figure 2.5, the base layer is independently encoded; base layer frames are referenced to other base layer frames. However, enhancement layer is encoded with reference to base layer frames which is called backward reference and with reference to other enhancement layer frames which is called forward reference.

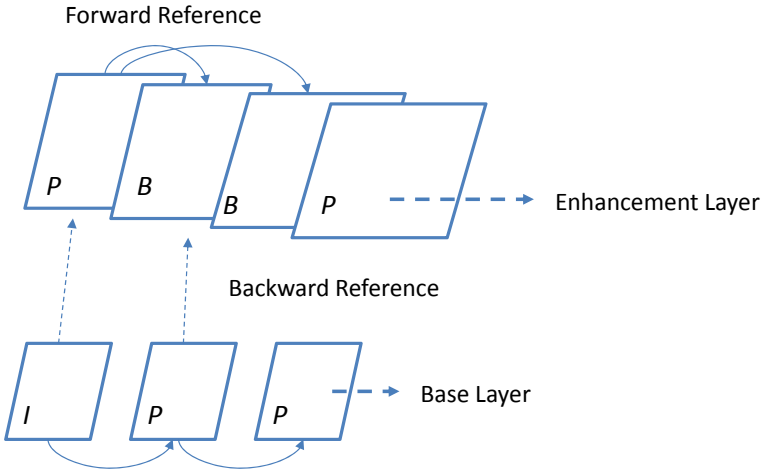


Figure 2.5: An example of spatial scalable encoding.

2.3.3 Multiple Description Coding (MDC) . Unlike SVC, multiple description coding encodes the video into multiple descriptions that are encoded independently of each other. The best received quality of the reconstructed video is obtained when all descriptions are correctly received and decoded. However, a lower video quality can still be achieved upon the correct reception of fewer descriptions with the least quality achieved if only one description is received.

One way to perform MDC in the temporal domain is shown in Figure 2.6. In this example, MDC is performed by splitting the video frames into odd and even frames. Each sequence of frames represents a description, which will be coded independently of the other sequence. However, the high redundancy in the descriptions will result in lower coding efficiency. The coding efficiency will be further lowered if we increase the number of descriptions. This is because as the number of descriptions increases the correlation between successive frames in each of the descriptions will decrease [13].

2.4 Path Diversity

In a typical Internet scenario, an end host drops packets to the network with a specific destination address. The underlying infrastructure of the Internet provides several paths

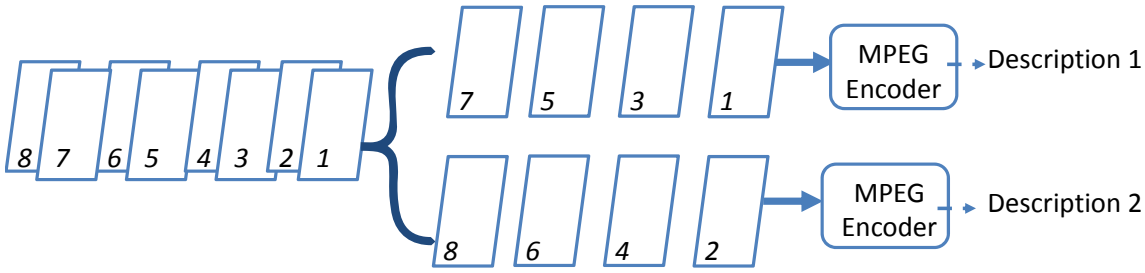


Figure 2.6: An example of multiple description encoding.

between communicating hosts. End hosts have no control on how packets are delivered. Instead, the underlying routing protocols control the packet delivery and the path selection. As discussed earlier, routing protocols select the lowest cost path based on a predefined cost metric. Packets of the same application session propagate along the selected lowest cost path. Even though multiple paths having the same lowest cost metric can be found, only one path is selected and advertised by the routing protocol. In packet switched networks, path diversity is a transmission technique that uses multiple paths to transmit packets from a source to a destination. More specifically, in a host to host communication, instead of sending packets through a single lowest cost path, packets are sent through multiple low cost paths.

In 1970's, dispersity routing was introduced by Maxemchuk in which messages are divided into sub-messages and sent to the receiver using multiple paths. Maxemchuk introduced dispersity routing by load balancing in the Advanced Research Projects Agency Network (ARPA-net) to provide more reliability and to overcome the problems of link failures [14]. In the early 1990's, the distribution of his work mainly focused on connection-oriented ATM networks [15]. Most of the research done on diversity techniques was related to wireless communication [16], [17]. Diversity techniques in wired networks did not gain interest until the early 2000's.

Although the infrastructure of today's Internet is highly redundant, routing protocols are known to be slow in reacting and recovering from network failures [18]. Studies show that 20% of path failures are not recovered within 10 minutes [19]. For

example, the fault recovery mechanism of the border gateway protocol (BGP) requires a long delay to converge [20]. Moreover, the underlying routing protocols lack the ability of adapting to network bottlenecked links. The underlying routing algorithms keep sending packets through the link as long as it is live. Other Internet measurements-based studies compared the quality of the default lowest cost path and other alternative paths in the Internet [21]. It was found that in 30-80% of the cases there exists an alternative path with a better quality, where the quality of the paths was measured in terms of loss rates, round-trip time and bandwidth. These constraints of the routing protocols have motivated the research of path diversity in wired networks.

Path diversity has both network related benefits and applications related benefits. From a network perspective, path diversity exploits the redundancy of the Internet infrastructure to better utilize network resources and to improve network performance [22]-[26]. Path diversity provides load balancing, which lowers the overall queuing delay of the network. Moreover, path diversity is proposed to reduce the delay of fault recovery mechanisms of the routing protocols.

From an application perspective, path diversity mainly has three advantages. It transforms burst losses into isolated losses, reduces the probability of an outage, and increases the available bandwidth of the application [27]. Burst losses results in a loss of multiple consecutive numbers of packets. This can be reduced by sending packets to disjoint paths; paths which do not have common links will de-correlate packet losses. Hence burst losses in one path transfer to isolated losses affecting only the packets transmitted to this path. Another type of packet losses, which is considered the worst type, happens when an outage occurs in which a total loss of communication remains for five to several seconds. Unlike sending to the default path, an outage occurs in a system that uses path diversity transmission technique when all paths between the sender and the receiver fail. For disjoint paths, the probability of failure of all the paths at a given period of time is smaller than the probability of failure of a single default path. Last, studies show that exploiting the redundancy of the Internet provides higher bandwidth to end users [28].

Issues in Path Diversity

To achieve an efficient path diversity-based transmission, several issues must be considered. This includes the implementation of a path diversity transmission technique, the selection of the paths, and scheduling packets transmission to the multiple paths. Commonly studied path diversity architectures in which a source can send packets through multiple paths to the receiver implements path diversity via source routing, path diversity via relays and path diversity via multihoming networks. Path diversity via source routing requires the source to explicitly assign the addresses of different nodes along the selected path. Although this method appears simple, it has several underlying problems. First of all, the source needs to have a good knowledge of the underlying network topology and network state in order to select the path. Even if it can be assumed that the network topology is manually inserted to the source, updating a large part of the network state is impractical. Because of its implementation difficulties, source routing was mainly proposed in intra-domain routing, where the network is relatively small [29].

Another common architecture is path diversity via relays, which is based on distributing an overlay of nodes over the network. The relay nodes are forward switching devices that are placed between the source and the destination. Unlike source routing, using relay nodes the source needs to only select the first hop of the path. Assuming that a source is given the addresses of two relays in the network, to achieve path diversity, the source will send part of the packets to the first relay and the other part to the second relay. The relays will use the destination addresses of the received packets to route them to their destination using the underlying routing protocols. One point to note here is that path diversity via relays does not guarantee that the multiple paths used for transmission are disjoint. This is because the relays route packets using underlying routing protocol; hence an overlap between the paths might occur [30].

Path diversity can be also achieved via multihoming networks [25]. Multihoming networks refer to networks with multiple connections to the Internet, for example a single network connected to multiple Internet service providers (ISP's). Path diversity

can be implemented by sending packets through multiple ISP connections. A source in a multihoming network can send the packets to the multiple nodes each in one ISP network. It is similar to the relay network in that the source needs to know the addresses of only the first hop of the selected path. However, path diversity via multihoming networks does not require any additional packet forwarding nodes (relays) to the network.

The second important issue is the selection of a path for a transmitted packet. The selection of a path can be done either in a simple round robin manner regardless of the difference in quality of the available paths or based on a best path selection optimization algorithm. Multiple paths have different qualities in terms of round trip time (RTT), packet loss rates and bandwidth. Thus, a better performance can be achieved when using an optimization selection algorithm. One way is to select the path based on the available bandwidth by using the average available bandwidth of the paths or by using the lower bounds on the available bandwidth variation range [30].

Instead of selecting the path based on one parameter (e.g. bandwidth), path selection can be also achieved over multiple parameters. This can be performed by using grey relational analysis (GRA) [31]. This method defines ideal path's parameters and finds the relation between this ideal path and each of the multiple paths. The relation is found by calculating the grey relational coefficient (GRC) in which the path with the optimal GRC is selected. Another way is to optimize the path parameters over a cost function which might be the expected distortion at the receiver. In other words, the path which is expected to deliver a lowest distorted video is chosen [32].

The third important issue is how to map different packets to multiple paths. This is especially important when it comes to transmitting packets of multimedia applications. Multimedia packets have different levels of importance because of the coding dependencies and time constraints. A multimedia packet is useful to the receiver if it is received before the decoding deadline and if reference frames needed for decoding are correctly received. In order to maximize the streaming quality, a scheduling algorithm that selects a packet to be transmitted at a given time instant to a specific path is necessary [33].

Chapter 3: Related Work

This section provides an overview of the contributions in the field of video streaming over the Internet focusing on the application layer, transport layer and the network layer. A general video streaming system is presented to clarify the relation among the different layers. Next, three sections are provided to discuss the different contributions, namely: application layer-QoS control, TCP friendly-congestion control and network layer-QoS support.

3.1 A General Video Streaming System

Extensive research has been conducted to address the challenges of video streaming over the Internet [34], [35]. In this chapter, three categories are discussed. Each category falls in one of the Internet layers, namely: the application layer-QoS control, TCP friendly-congestion control and network layer-QoS support. Figure 3.1, shows the different layers which collectively provides a video streaming system.

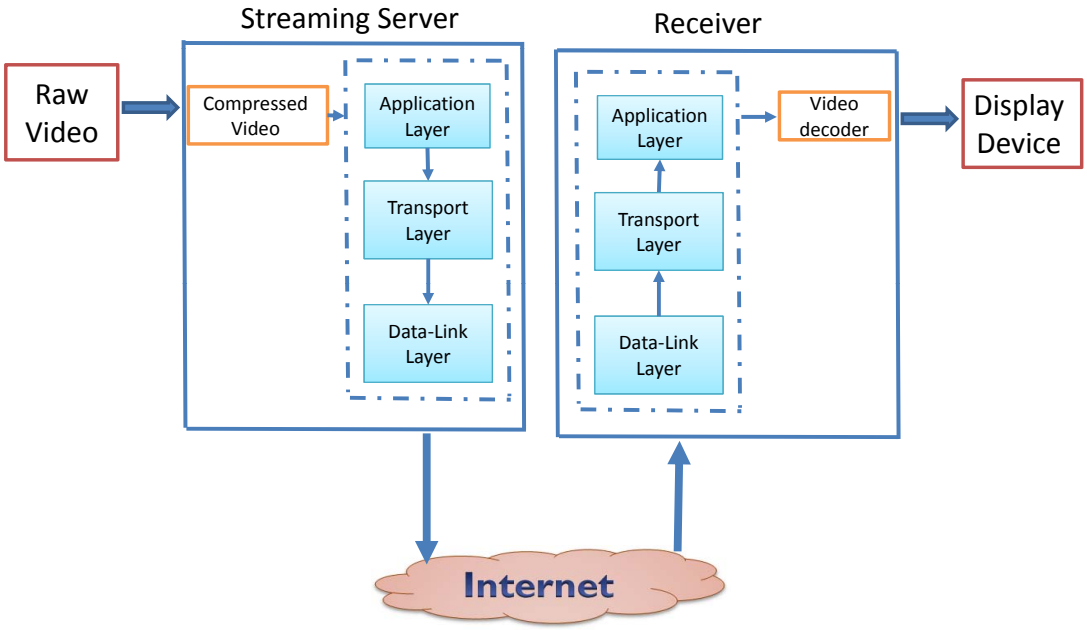


Figure 3.1: Architecture for video streaming showing in a layered structure.

For the case of pre-encoded video, the relation between the different layers is illustrated as follows. Raw video data is compressed by the compression algorithms (as described in the previous chapter) then saved in storage devices. Upon a client's request the compressed video data is retrieved and is handed to the application layer which controls the format of the transmitted data. After this, the transport layer adjusts the transmission rate of the packets. Finally, the network layer performs the routing algorithms in which the packets are delivered to their destination.

3.2 Application Layer-QoS Control

The objective of the application layer-QoS control is to avoid network congestion and provide the QoS requirements of video applications in the presence of packet loss. To achieve the aforementioned objective, application layer-QoS control employs adaptive techniques at the end systems which do not require any QoS support from the underlying network. These techniques can be used either by the application layer of the video source such as the source rate control and the rate-distortion optimization techniques or by the application layer of the video receiver such as adaptive playback.

Source rate control is a technique to adapt the sending rate of video traffic based on the estimated available bandwidth and QoS requirements of the application [36]-[38]. This is done at the application layer of the video source in an attempt to avoid network congestion by matching the transmission rate to the available network bandwidth. Typically, source rate control relies on network feedback to enable the source to adapt to bandwidth fluctuations. Moreover, source rate control requires a variable bit rate video stream which might be a challenge for a pre-encoded video. However, this is possible with the use of scalable video coding which provides a single encoded video stream with different bit rates.

Source rate control can also be used in adaptation of the playback buffer occupancy as the work presented in [39]. This work aims to maintain a continuous video playback by monitoring the buffer occupancy of the receiver and adapting the trans-

mission rate of the video source accordingly such that higher transmission rates are required when the buffer occupancy is below a certain threshold.

Another variation of the above mechanism, is the use of rate-distortion optimization framework in which the video source varies the transmission rate according to the expected distortion at the receiver [40], [41]. This is performed by receiving feedback signals indicating the distortion level of the reconstructed video.

Finally, an example of application layer-QoS control, which is done by the receiver is the adaptive video playback [42]. The objective of the adaptive playback is to provide a continuous video playback by adapting the playback rate in accordance to the occupancy of the playback buffer. This is done such that lower playback rate is used, if the playback buffer occupancy falls below a certain threshold; this allows the buffer to build up and helps avoid the buffer to underflow.

3.3 TCP Friendly-Congestion Control

The Internet provides data transportation services in two transport protocols including reliable transmission protocol (TCP) and an unreliable transport protocol (UDP). TCP protocol is not well-suited for video streaming applications because of several reasons. First, the feature of reliability it offers is obtained at the expense of large round trip delays. This is because the TCP protocol relies on automatic repeat request in which the video receiver requests the video source to repeat (retransmit) a lost packet. Second, the TCP employs sophisticated congestion control mechanisms which results in fluctuations in the available bandwidth [37].

Congestion occurs when the total arrival rate to a router's input ports exceeds the departure rate of the router's output ports. This will have severe effects on the network performance including: high delays in packet delivery, packet loss which results in wasted resources and, in worst case, congestion collapse might occur in which throughput drops to zero and end-to-end delay of packet delivery goes to infinity. This becomes especially critical to end users running real time applications such as video streaming applications. To prevent congestion, the TCP protocol employs congestion

control mechanisms, which complement the end-system protocols by avoiding and controlling congestion as to limit packet drop and delay.

The TCP congestion control mechanism is performed by having each sender control its transmission rate in response to the perceived network congestion. The sender perceives that there is congestion on the path between itself and the destination by receiving three duplicate ACKs from the receiver or by detecting a timeout event. The sender controls its transmission rate by defining a variable known as the congestion window. The congestion window sets a maximum bound on the amount of unacknowledged data. By limiting the amount of unacknowledged data, the mechanism indirectly limits the sending rate of the source.

The working mechanism of the TCP congestion control scheme is based on three components: the additive-increase multiplicative-decrease (AIMD), slow start and reaction to a loss or a timeout event. The AIMD algorithm increments the transmission rate of the source by increasing the congestion window linearly by one maximum segment size (MSS) after each round trip time (RTT). Whenever a loss of event is detected, the multiplicative decrease approach cuts the window size to half its initial value (before detecting the loss event). The TCP congestion control uses the AIMD algorithm in the congestion avoidance phase.

The sender starts its transmission by the slow start phase in which the sender exponentially increases its transmission rate. The TCP congestion control mechanism sets a threshold value to determine the point at which the slow start phase ends and the congestion avoidance state begins. The threshold value is initially set to a high value allowing the transmission rate of the sender to increase exponentially. Once the congestion window reaches the threshold value, the sender enters the congestion avoidance state and the congestion window grows linearly. If a triple duplicate ACKs loss event is detected, the threshold value and the congestion window are set to one half of the last congestion window value (when the loss event is detected). Finally, if a timeout event occurs the threshold value is set to one half the current congestion window and the congestion window is set to 1 MSS.

The above congestion control scheme is known as TCP Reno. It is noticed that in the occurrence of a loss event the mechanism reacts differently than the occurrence of a timeout event occur. The congestion window is reduced to 1 MSS when a timeout event occurs, whereas the congestion window is set to half its value in the case of detecting a loss event. An early version of TCP, known as TCP Tahoe, cuts the congestion window to 1 MSS and enters the slow start phase when either type of losses occurs. However, the newer version, which is TCP Reno, cancels the slow start phase after a triple duplicate of ACKs event and enters the congestion avoidance phase. This is because, although a packet is lost, the reception of three ACK packets shows that the network is able to deliver at least some packets. The cancellation of the slow start phase after receiving three ACKs is known as fast recovery [2].

The congestion control schemes have a fast response to network congestion; however, they lead to large variations in the delivered transmission rate. These variations in the transmission rate will degrade the performance of video streaming applications significantly. This is because the delivered video will have a significant quality variation over time which is undesirable for the viewer's experience [43], [44]. Thus, traditional congestion control mechanisms need to smooth out their transmission rate variation to allow a smooth playback quality for the end user. Moreover, video applications usually have a minimum bandwidth requirement which is not considered by the congestion control scheme. Thus, a TCP-friendly congestion control scheme is needed to meet the quality smoothness requirement and the bandwidth requirement of video applications. This should be achieved in a TCP friendly manner such that video applications share the bandwidth fairly with other TCP-flows.

Several efforts have been made to design a smooth TCP congestion control mechanism [45]. The equation-based rate control has been proposed as a solution for multimedia applications [34]. It has the desirable characteristic of achieving a smooth rate variation. However, it has a slow response to network congestion. Another, work is presented in [46], which proposes a linear increase and graded multiplicative decrease.

This scheme provides an additional degree of freedom in which the smoothness and the response time can be adjusted.

3.4 Network Layer-QoS Support

Currently, routers, switches and the routing protocols of the Internet are oblivious to the content of the packets; hence, packets are treated equally in the network. Although this design of the Internet has enabled it to reach its present size, it has impacted the multimedia flows which have divers QoS requirements as a result of the employed compression algorithms. Thus, there is a mismatch between the compression schemes and the Internet mechanisms that do not prioritize packets. QoS requirements of the video applications can be provided by the network layer in terms of the streaming schemes employed and the design of the routers and switches.

Several studies proposed the use of path diversity for providing QoS requirements for multimedia applications over best effort networks. Streaming approaches that consider the joint combination of path diversity and multiple description coding (MDC) were proposed in [17], [32], [47]-[50]. As previously discussed, MDC encodes the video into multiple independently coded streams. When path diversity is used, it is highly recommended to route descriptions along disjoint paths so that the probability of losing all the descriptions decreases.

Data partitioning and path diversity were combined in the scheme proposed in [51] where the video information is divided into different partitions with different contribution to the quality of the decoded video sequence. After packetization, the scheme duplicates packets containing partitions with higher importance data to multiple paths, while packets containing partitions with less importance are sent only once on any of the paths. Path diversity can also be used in providing dynamic QoS routing for different traffic classes as proposed in [52]. This can be done by coding the video into a base and an enhancement layer. Since the base layer is more important than the enhancement layer, a path diversity scheme can be used to provide best QoS requirements for the base layer.

The priority and path diversity-based schemes differ from the above mentioned work in several aspects. First, the proposed schemes exploit the differences in importance between video frames within a single layer video without the use of data partitioning or multiple description coding [17], [32], [47]-[51]. Second, unlike previous work, which focuses on the selection of multiple disjoint paths for transmitting video packets, the proposed path diversity scheme selects a best path based on the importance level of the transmitted video packet and the feedback congestion information from the network. Third, instead of transmitting multiple copies of the same video packet to the available paths [51], the proposed scheme selects a single best path for a transmitted video packet, as will be discussed in the next chapter.

Chapter 4: Performance Evaluation of the Priority and Path Diversity Schemes

This chapter proposes a priority-based path diversity (PBPD) scheme for video streaming over best effort networks. In the first section, the PBPD scheme is introduced using a simple network topology in which multiple paths exist between a video source and an end-user. This section also describes the priority queuing-based component and the proposed streaming algorithm of the PBPD scheme. In the second section, the PBPD scheme is generalized for more complicated network topologies.

After that, the performance of proposed PBPD scheme is investigated through simulations in terms of network related metrics such as end-to-end delay and frame loss rate. Two streaming approaches are also investigated for the sake of performance comparison with the proposed streaming approach. Namely, the performance of a best effort streaming model as well as a priority queuing-based streaming approach are compared with that of the proposed PBPD scheme.

4.1 Overview of the Proposed Scheme

The proposed PBPD scheme is designed to enhance the viewer's experience by reducing the delay and loss rates of important video packets that might happen if they are routed to a congested path. This is done by jointly considering the importance level of a video packet and the congestion level along the candidate paths when selecting a path to transmit a video packet. This in turn requires the network to continuously share its congestion information with the ingress router which is the router directly connected to the video source.

The PBPD scheme is proposed in the network layer with minimal addition to the underlying routing protocols in which a congestion state metric is added to the path selection criteria. As discussed in Chapter 2, routers running protocols such as OSPF continuously broadcast their link state information; hence, the PBPD scheme requires

the routers to include a congestion state information along with the broad-casted link state information to allow the selection of the minimum congested path. In addition, the PBPD employs a priority queuing system in the input ports of the network routers which maintains separate priority queues and serves video packets in a priority basis. Prior to their transmission, video packets are assigned a priority level based on the type of frame they contain, which will indicate the level of service to be provided by the network. Packets containing I-frames are assigned the highest priority level. Packets containing P-frames are assigned the second level of priority. Finally, packets containing B-frames are assigned the lowest priority level.

Assigning priority levels to the video packets can be easily implemented in today's Internet by using a special field in the packet header to assign the priority level. For example, the packet header format of the IPv4 reserves 8-bits field called the type of service field [2]. This field is included to allow the different types of packets to be distinguished from each other in which the specific level of service provided to each type is determined by the router's administrator. Furthermore, the IPv6 reserves a 20-bits field called the flow label [2]. The flow label field labels packets belonging to flows which request special services (non-default services). A flow might be a video transmission or it might be a traffic originating from a special user (i.e.: someone who is paying more). Moreover, the IPv6 includes 8-bits traffic class field which is similar to the type of service field in the IPv4 for which the service differentiation can be done within the same flow. Although these fields are not widely used, having them in both the IPv4 and IPv6 indicates that the designers foresee the eventual need for service differentiation among different type of packets in the network [2].

In what follow, we describe the priority-based component of the proposed PBPD scheme and the path diversity transmission scheme.

4.1.1 Priority Queuing System. The priority-based component of the proposed PBPD transmission scheme namely, the priority queuing system, maintains four priority queues. The highest priority queue serves packets of the most important frame

types (i.e.: I-frames). Similarly, the second priority queue serves packets of frame types with the second importance level (i.e.: P-frames). The third priority queue serves packets belonging to the least important video frames (i.e.: B-frames). Finally, the fourth queue is used by packets of best effort traffic. The queue dedicated for best effort traffic is maintained to demonstrate the possibility of integrating the proposed model in a real network that could be serving other types of traffic rather than video.

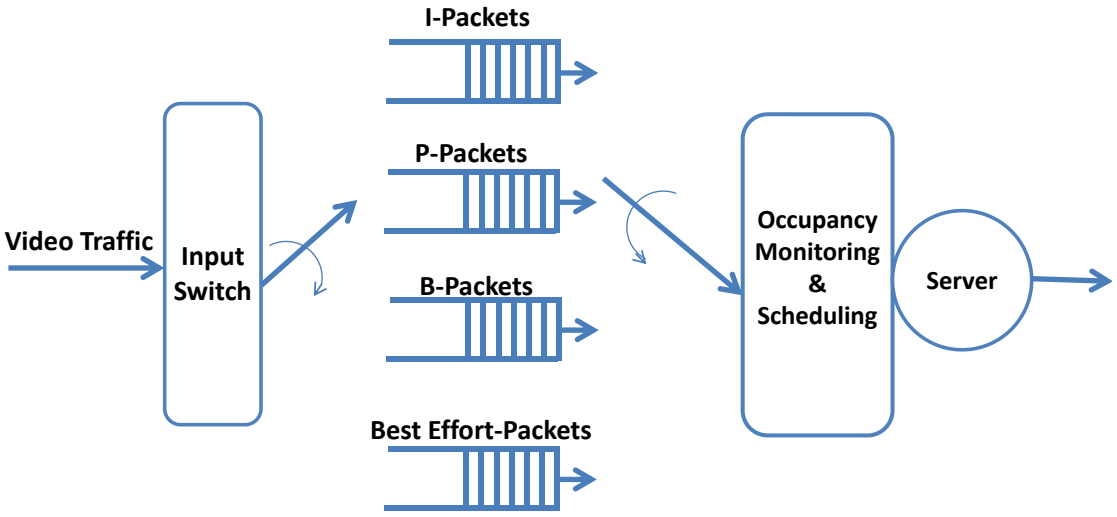


Figure 4.1: The priority queuing system employed in the PBPD scheme.

Figure 4.1 depicts a simple block diagram of the priority queuing system. Upon their arrival, packets of I-, P- and B-frames will be directed to the I-queue, P-queue and B-queue, respectively. The occupancy monitoring and scheduling unit continuously monitors the occupancy of each queue. Based on the priority level along with the instantaneous occupancy of the different queues, the scheduling unit decides which queue to serve. The scheduling unit serves the I-queue as long as its occupancy is not zero. When the I-queue is empty, the scheduling unit serves the P-queue if it is not empty. The scheduling unit only serves the B-queue when both higher priority queues (i.e., I- and P-queues) are empty. Similarly, the best effort queue is served only when the other three queues are empty. From the above discussion, it can be seen that the priority queuing system falls under the Diffserv architecture which aims to enable service differentiation within the Internet. However, in this work, the priority queuing system is

proposed specifically for video applications by considering the difference in importance between video frames.

4.1.2 PBPD Scheme for Simple Network Topology. Figure 4.2 shows a possible scenario in which the proposed PBPD scheme can be employed. The network consists of multiple paths between a video source and an end user. Prior to their transmission, video packets are given a priority tag depending on the frame type they contain which will be used in queuing and forwarding video packets in the network. At each involved router, the monitoring unit (see Figure 4.1) shares the queue length of each of the employed priority queues with the ingress router. Thus, the ingress router has a global overview of the status of the network allowing it to select the minimum congested path when transmitting video packets. The ingress router will select the path with the minimum cumulative occupancy of the I-queue buffers for packets belonging to I-frames. Similarly, the ingress router will select the path with the minimum cumulative occupancy of the I-queue and P-queue for the packets that belong to the P-frames. For B-frames packets, the video source selects the path with the minimum cumulative occupancy of all the queues but not the best effort queue.

It should be noted that although losses and queuing delays might occur at the input ports and/or the output ports of routers, this work is mainly concerned with reducing losses that take place at the input ports of routers due to congestion. For this reason, the problem is simplified by assuming that packets experience a relatively low delays at the output ports compared to the input ports and zero losses occur at the output ports of the network routers. However, the feedback information proposed in the PBPD scheme (cumulative queue lengths) can be modified such that it accounts for the queue lengths at both the input and output ports. In this case, multiple paths through a specific router will incur different costs (cumulative queue lengths) depending on the output ports they are assigned to take.

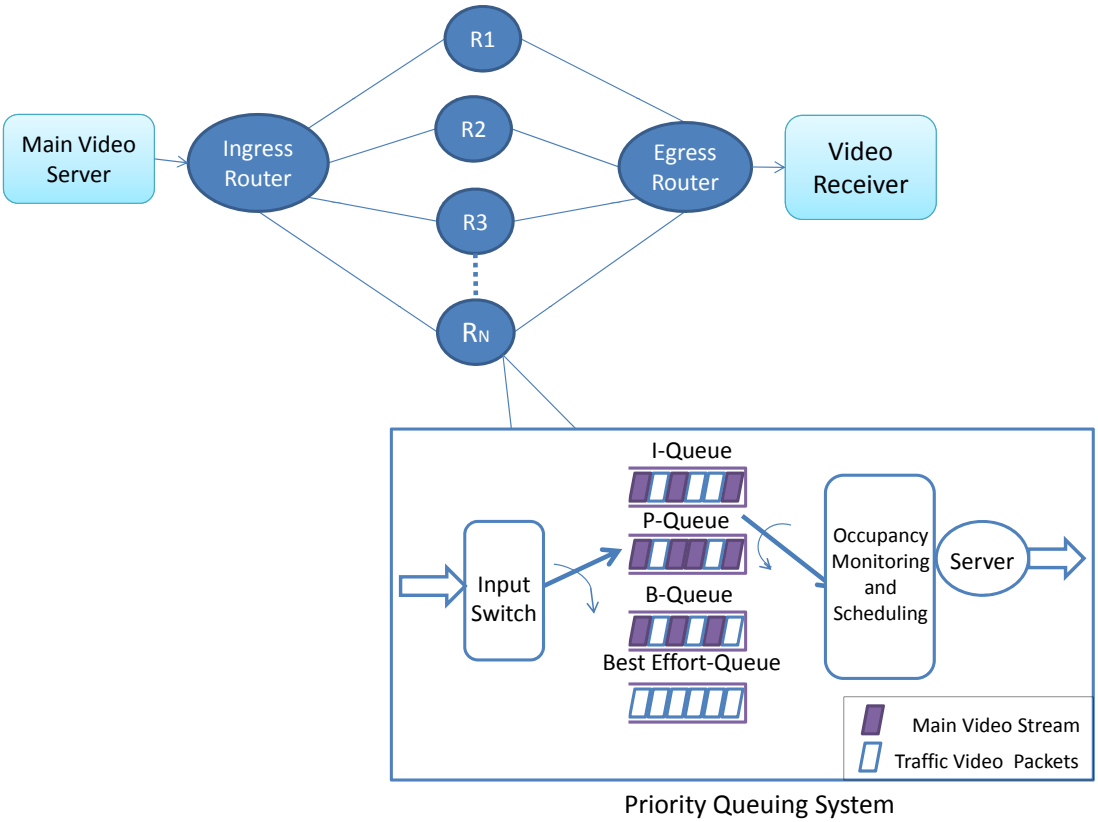


Figure 4.2: System model.

4.1.3 PBPD Scheme for General Network Topology. Figure 4.3, shows a more complicated network architecture in which the proposed PBPD scheme needs to be employed. The proposed PBPD scheme can be extended to larger networks such that packets are routed to the lowest congested path by means of routing in which routers exchange congestion information and decides the minimum congested path for a transmitted packet. In more details, routers are responsible for finding and updating the lowest congested path in their forwarding tables without the interference of the video source. As the video source transmits video packets to the network with a specific destination address, routers direct video packets to their destination through the minimum congested path using the generalized PBPD scheme.

The generalized PBPD scheme requires routers to monitor the average queue lengths of their priority queues and to broadcast routers' state information whenever a

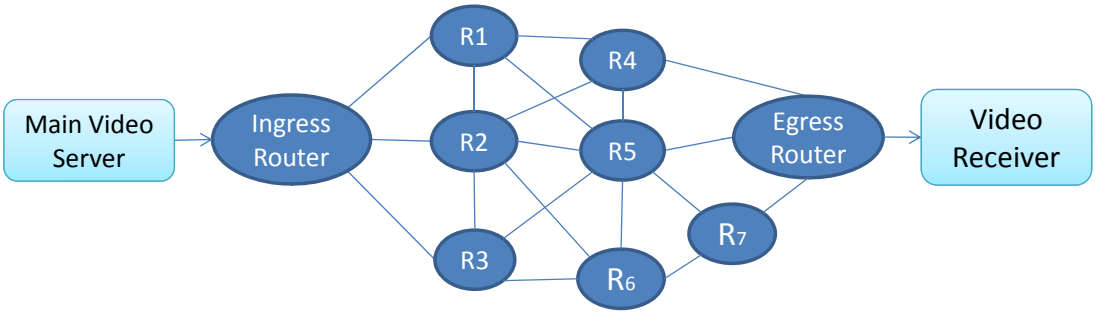


Figure 4.3: General network topology.

change in routers' state is detected. The routers are said to have one of three different states according to their average queue lengths. As shown in Figure 4.4, three congestion states are defined, namely: normal state, congestion avoidance state and congested state. The average queue length is calculated using the moving average equation [10], $\bar{q}(t) = \frac{\bar{q}(t-\Delta)+q(t)}{2}$, where $\bar{q}(t)$ is the average queue length at time t, $q(t)$ is the queue length at time t, and $\bar{q}(t-\Delta)$ is the average queue length at time $t-\Delta$. The value of Δ depends on the arrival and departure events of the queue such that the average queue length is calculated after an arrival or departure events. Based on the average queue lengths, a router is said to be in a normal state if its average queue length is below a minimum threshold Q_{min}. A router is in a congestion avoidance state if its average queue length is between the minimum threshold Q_{min} and a maximum threshold Q_{max}. Finally, a router is in a congested state if its average queue length is above Q_{max}. The queue threshold values are defined to be fractions of the buffer capacity. Without loss of generality, the threshold values used in this work are as follows, Q_{min} is set to 50% of the total buffer capacity and Q_{max} equals to 80% of the total buffer capacity.

The PBPD scheme defines the best path to be the path with minimum congestion and with minimum number of hops. This is true, with the assumption that all network links are of the same speed (bytes/sec); hence, the available paths between the video source and the end-user differ by the routers' states and the number of communication links along the path. Although we have simplified the routing problem by having the network links to be of the same speed, the algorithm can be generalized to networks

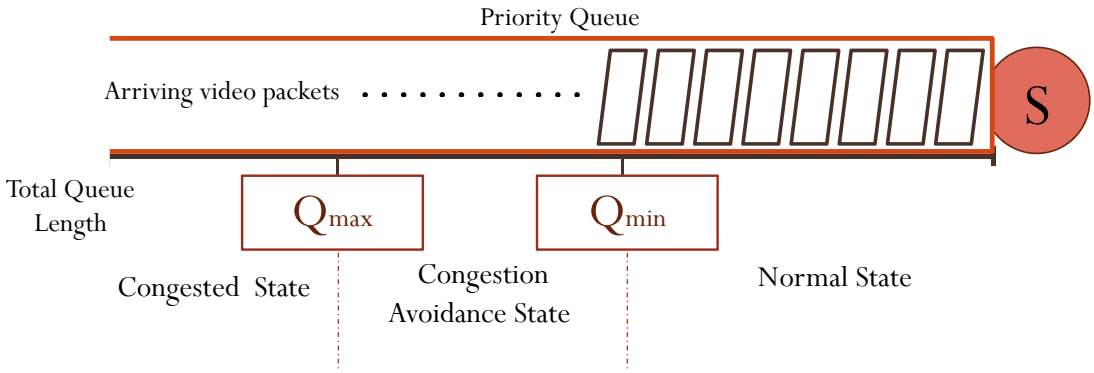


Figure 4.4: Congestion states of a priority queue.

with different links' speeds by an optimization framework for the path selection that considers the propagation delay and the queuing and serving delays of the routers along the path.

The selection of the best path in the proposed PBPD scheme is performed in the following steps. First, routers find the paths with the least hop count (or communication links) to each destination. This is possible because routers are assumed to have a full knowledge of the network topology (link state routing algorithm). Second, routers will save the multiple least hop count paths to a specific destination. Third, by receiving congestion state information, routers will select the best path for each network destination to be the minimum congested path from the saved least hop count paths. One point to mention is if a router along the selected best path has its state changed during the transmission of the video packet, the video packet will still propagate along the best selected path. However, these changes in the routers' states will be broadcasted to other routers in the network which will then update their forwarding tables accordingly.

The algorithm can be further explained by referring to the network in Figure 4.3 in which a video source transmits video packets through the network to the end-user. The video source transmits video packets to the ingress router (directly connected router). Initially, the ingress router will save the least hop count paths which are the paths with two hops. In other words, the least number of routers that packets might pass through from the ingress router to the egress router are two routers. Thus, the

ingress router will save a total of five paths which are two paths through router 1, two paths through router 2 and one path through router 3. Based on the updated congestion state information of the routers included in the least hop count paths (routers (1-5)), the ingress router will select the best path (minimum congested path) for each transmitted video packet.

As in the simplified PBPD scheme the selection of the best path is done by jointly considering the priority level of the transmitted video packet and the congestion level among the paths. Each router will broadcast four congestion state information which are the average queue lengths of its employed queuing system. In more details, each router consists of four priority queues each serving packets with specific priority level. Hence, a router's state includes four different congestion states depending on the average queue length of the I-queue, P-queue, B-queue and best effort-queue. When a packet with the highest priority level (containing I-frame) is to be transmitted, the router selects the path with the minimum cumulative congestion state of the I-queue. In the same manner, if a packet with the second priority level is to be transmitted, the router selects the path with the minimum cumulative congestion state of the I-queue and the P-queue. Finally, when a video packet containing a B-frame is to be transmitted, the path with the minimum cumulative I-queue, P-queue and B-queue is selected. Based on the above discussion, it should be noted that different types of video packets (video packets with different priorities) might be transmitted along different best paths.

The implementation of the proposed schemes in today's Internet requires modifying the queuing system of the existing routers such that they consist of four priority queues. Moreover, the existing routing algorithms can be modified such that it consider the congestion in selecting the best path. For example, the open shortest path first (OSPF) algorithm used in todays Internet to find the minimum hop count can be adjusted such that it finds the lowest queue lengths. This will require adding a congestion information in the broadcasted link state packets.

4.2 Simulation Results

In this section, the performance of the proposed PBPD streaming scheme is evaluated in a simulated multi-path scenario where an encoded video is transmitted using the proposed scheme. For the purpose of performance comparison, three models are investigated namely: a best effort-based model, a priority-based queuing model and the proposed path diversity-based model. While the best effort-based model reflects the performance of today's Internet, the priority-based queuing model and the path diversity-based model investigate the different components of the proposed PBPD scheme. In particular, the priority-queuing based model is used to study the performance of the priority-queuing system component of the proposed PBPD scheme; the path diversity-based model complements the priority-based queuing model by adding the proposed path diversity streaming scheme.

The three models differ in the employed queuing and streaming approaches. The best effort-based model employs the queuing and forwarding functionalities of today's Internet. In more details, video packets are treated equally in the network regardless of the type of video frame they contain. Moreover, routers in the best effort-based model maintain a First-In First-out (FIFO) queue which is served by a fixed service rate reflecting the speed of the outgoing link at the output of the routers. Therefore, an arbitrary packet will experience a queuing delay that depends on the number of packets ahead of it in the FIFO queue and a service delay that depends on its size. Unlike the best effort-based model, routers involved in the priority-based queuing model maintain the priority queuing system which consists of four different queues served according to their priority as previously discussed. Finally, the path diversity-based model presents the proposed PBPD streaming approach in which routers employ the priority queuing system and they guarantee that video packets belonging to more important frames will be always routed to the least congested paths.

To mimic real-life transmission scenarios with variable levels of congestion, traffic video sources are introduced in the network. These video sources generate video

traffic with Gamma distributed frame sizes [53], [54], which is the distribution generally used to model the sizes of video frames. The generation rate of the video sources follows a Poisson random process. The level of congestion in the network is controlled by varying the generation rate of these sources. This is done by varying the inter-generation time of the video traffic using $E[T_g] = 1/(f_e\, C_L)$ for $C_L \in [C_{\text{Lmin}}, C_{\text{Lmax}}]$, where C_L denotes the congestion level, f_e denotes the nominal encoding rate of the video source and T_g denotes the generation period of video traffic. In what follows, the video source generates frames at a constant rate of $f_e = 30$ frames/sec and the congestion level is an integer chosen in the interval [0, 6]. A value of $C_L = 0$ indicates that there is no traffic in the network other than that generated by the primary video source. When $C_L = 1$, the inter-generation time of video traffic sources in the network is exponential with a mean that is equal to $1/f_e$ sec while $C_L = 2$ corresponds to exponential inter-generation time of $1/2f_e$ sec and so on.

Beside the traffic video sources, a primary video source is used to transmit the tested video sequence to the end user. It is important to note that, in studying the performance of the different streaming models, video packets generated by the primary video source and that generated by traffic video sources are treated similarly and according to the employed streaming model.

In what follows, the performance of the PBPD scheme is studied in terms of network related metrics such as the end-to-end delay and the frame loss rates. The results are obtained for the simple PBPD scheme and for the generalized scheme in which each is compared to a best effort-based model and a priority-based queuing model.

4.2.1 End-to-End Delay. Video applications are highly time sensitive due to the tight delivery deadlines associated with video frames. More specifically, for a video frame to be useful at the receiver, it must arrive not only before its play-out instant but also before a decoding deadline. This can be thoroughly explained by considering the application level functionalities of a video streaming process. Figure 4.5, shows a time line of the transmitted and received video frames in which f_t is the transmission

frame rate that varies based on the frame sizes and the available bandwidth and f_d is the display frame rate controlled by the application layer of the video receiver.

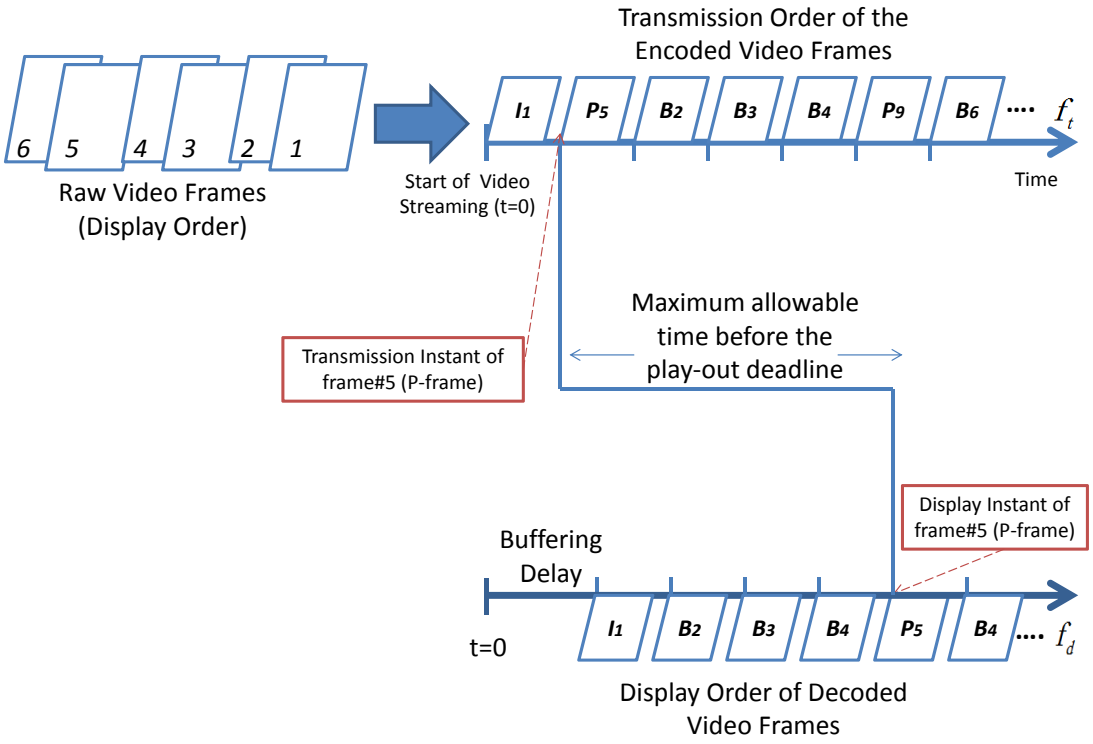


Figure 4.5: Time line diagram of a video streaming process.

The first point to consider is that the encoder outputs the video stream in the so-called encoding or transmission order which results from the interdependencies introduced by the compression algorithm. The encoding order can be illustrated through the following example. Suppose the first frame in the raw video stream is intra-frame coded (independently coded) into an I-frame. As part of the encoding process, the encoder locally decodes the I-frame and stores it in a local buffer after which it is used as a reference frame to encode other frames in the GoP. The encoder locally decodes reference frames so that the same reverse operation is performed at the receiver side. More specifically, the receiver will receive an encoded version of the I-frame; hence, it will decode the dependent frames with reference to the decoded I-frame. Noting that video compression standards are lossy compression schemes in which a certain error

(noise) is introduced, resulting in the decoded I-frame being different than the original I-frame (first frame). Still however, the correct decoding process is preserved by having the encoder locally decoding the reference frames, buffering them, and using them to encode dependent frames.

In the same manner, the fifth frame in the raw video stream is then encoded with reference to the decoded I-frame; thus, it is a predictive coded frame (P-frame). The P-frame is also decoded and buffered locally in the encoder. Lastly, the second, third and fourth frames in the raw video stream are encoded with reference to bi-directional video frames which are the decoded I-frame (forward reference) and P-frame (backward reference).

To successfully decode dependent frames, reference frames should be delivered and decoded before the decoding instants of their dependent frames. Moreover, if a reference frame has failed to arrive before the decoding deadline of its dependent frames, it still needs to be delivered before its own decoding deadline so that it can be displayed at the correct displaying instant. This can be better explained by referring to the time line presented in figure 4.5 which shows the maximum allowable time that frame number 5 (P-frame) can take from its transmission instant to its display instant. In this example, P-frames and I-frames (other than the first I-frame in the video stream) are transmitted in advance proceeding 3-frames (B-frames) from their display order. Once a reference frame -for example frame number five (P-frame)- is received, the decoder will decode it and buffer it in its local buffer to be used for decoding dependent frames, which are frames number two, three, four (B-frames). After that, at the display instant of frame number five, the receiver will display the decoded P-frame. However, a copy of it is still needed for decoding frames number six, seven, eight (B-frames) and nine (P-frame). If the P-frame is received after the decoding deadline of its dependent frames, the receiver will use error concealment schemes to decode dependent frames which will result in the degradation in the quality of the reconstructed video.

Finally, the transmission order of B-frames delays their transmission by one frame (reference frame) from their display order. However, they should arrive before their decoding deadline so that they are not discarded at the receiver.

From the above discussion, two main points can be concluded. First, unlike non-real time traffic, video packets have stringent time delivery deadlines after which video packets are not useful to the receiver. Second, in the same video stream, reference video frames (I-and P-frames) have more important deadlines as their unsuccessful arrival will result in not only missing the display instant of the reference frames themselves but also of the incorrect decoding of dependent frames which will degrade the quality of the reconstructed video.

In this section, the three schemes are compared in terms of the end-to-end delay, which is defined to be the time taken by a video packet from its transmission instant until it reaches the destination including the total propagation delay in the network and the queuing and serving delays in the intermediate routers. The propagation delay depends on the speed of the network links which are set to be the same in all simulated models. On the other hand, the queuing and serving delays vary in the three simulated models based on the employed queuing system and transmission scheme. As mentioned earlier, the best effort-based model employs a simple first-in-first-out (FIFO) queue; thus, the queuing delay of an arbitrary packet depends on the number of packets ahead of it in the queue and the service delay depends on the packet's size. Unlike the best effort-based model, the path diversity-based model and the priority-based queuing model use a priority queuing system in their routers which serves arriving packets based on their priority. Serving packets in the network in a priority basis by the use of the priority queuing system and the PBPD transmission scheme is expected to reduce the end-to-end delays experienced by reference frames (I- and P- frames).

End-to-End Delay: Results

In this section, the performance of the PBPD scheme is investigated in terms of the average end-to-end delay when employed in the network topology of Figure 4.2 and the

network topology of Figure 4.3. The average end-to-end delay is obtained for packets with different priorities (containing different frame types) when transmitted over the three simulated models.

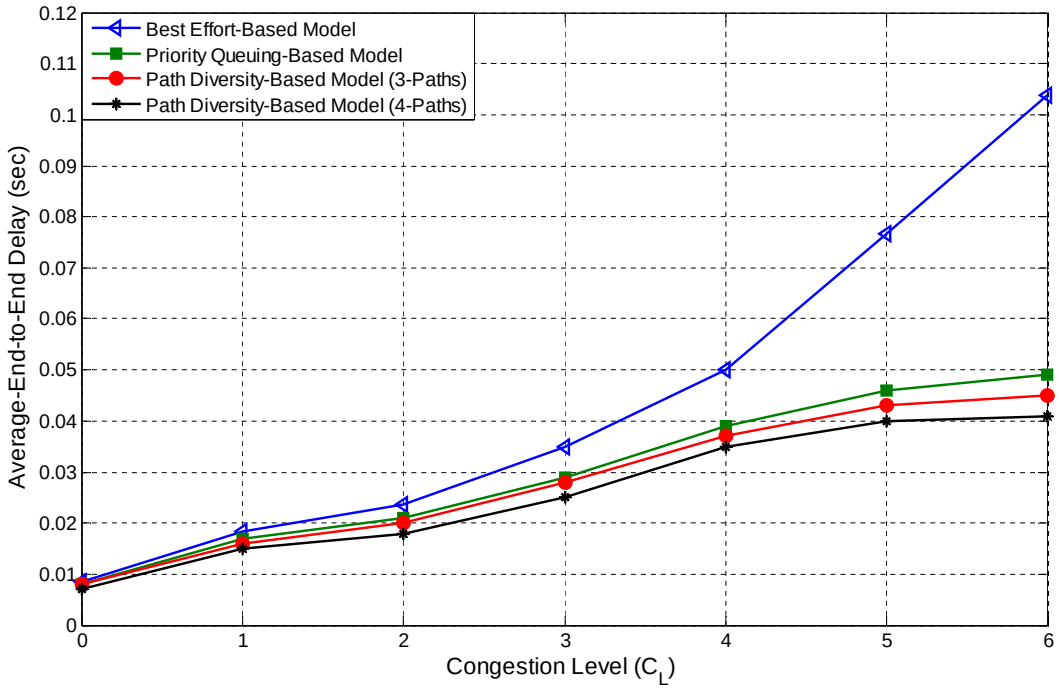


Figure 4.6: Average end-to-end delay of I-frames (the simple network topology).

Figure 4.6 shows the average end-to-end delay of I-frames versus C_L when transmitted over the simple network topology. At small C_L values, I-frames experienced almost the same end-to-end delay for the three models. Intuitively, this is explained by the low congestion levels over the network. As C_L increases, I-frames experience the worst end-to-end delay for the best effort-based model which is typically not acceptable by streaming applications. The figure shows that the path diversity-based scheme outperforms the priority-based queuing model for all congestion levels. It also illustrates that a better performance in terms of the end-to-end delay is achieved as the number of available paths is increased.

A similar end-to-end delay performance is achieved for I-frames transmitted over the general network topology which is shown in Figure 4.7. The best performance

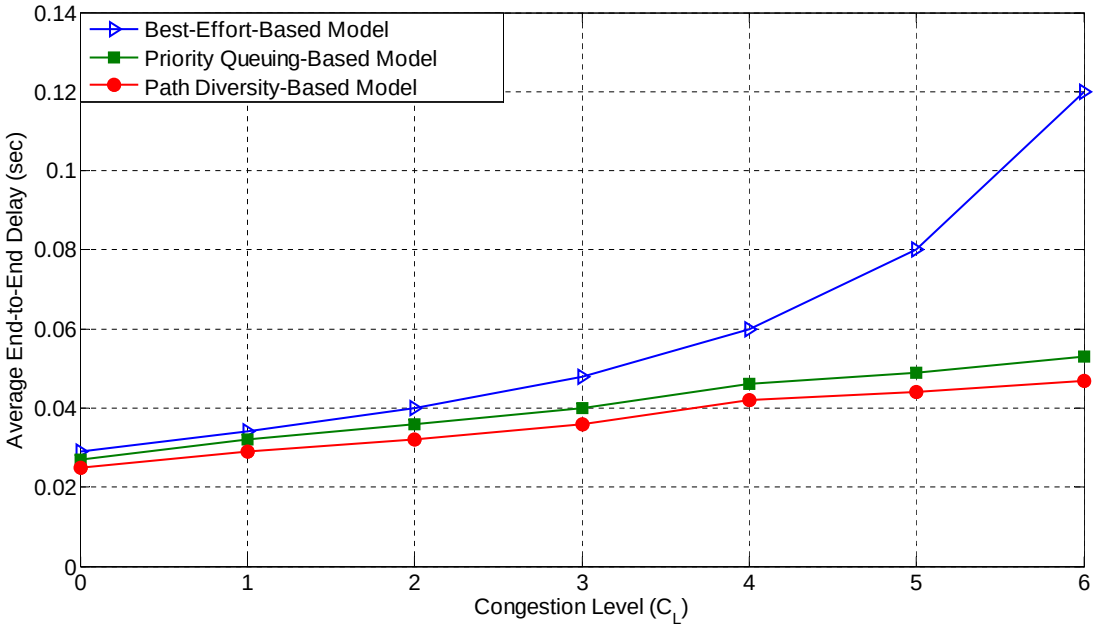


Figure 4.7: Average end-to-end delay of I-frames (the general network topology).

is achieved in the path diversity-based model which is slightly better than the priority-based queuing model but with noticeable enhancement compared to the best effort-based model.

Figures 4.8 and 4.9 depict the average end-to-end delays of P-frames versus C_L for the simple network topology and the general network topology, respectively. The average end-to-end delay of P-frames shows almost similar results to the case of I-frames where the path diversity-based model outperforms the best effort-based model and the priority-based model. These results confirm the initial assumption that by considering the priority level of video packets, reference frames (more important frames) experienced lower end-to-end delays compared to best effort networks where packets are treated regardless of their QoS requirements. Moreover, the results indicate that the PBPD scheme outperforms the priority-based queuing model and the best effort-based in both network topologies.

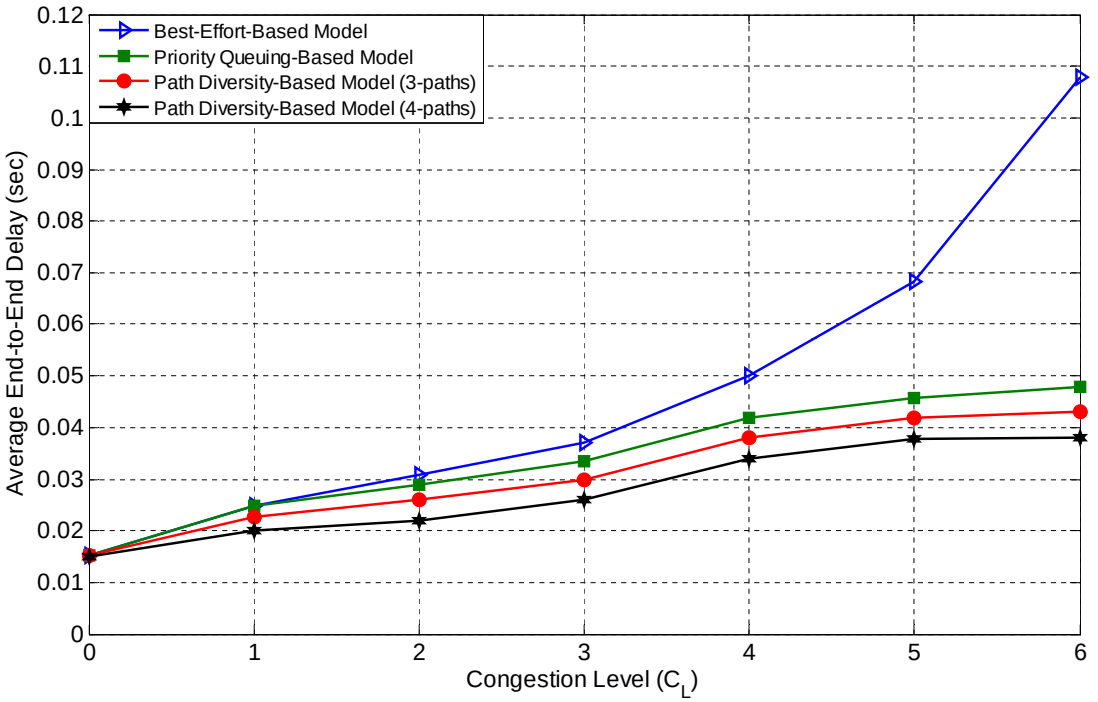


Figure 4.8: Average end-to-end delay of P-frames (the simple network topology).

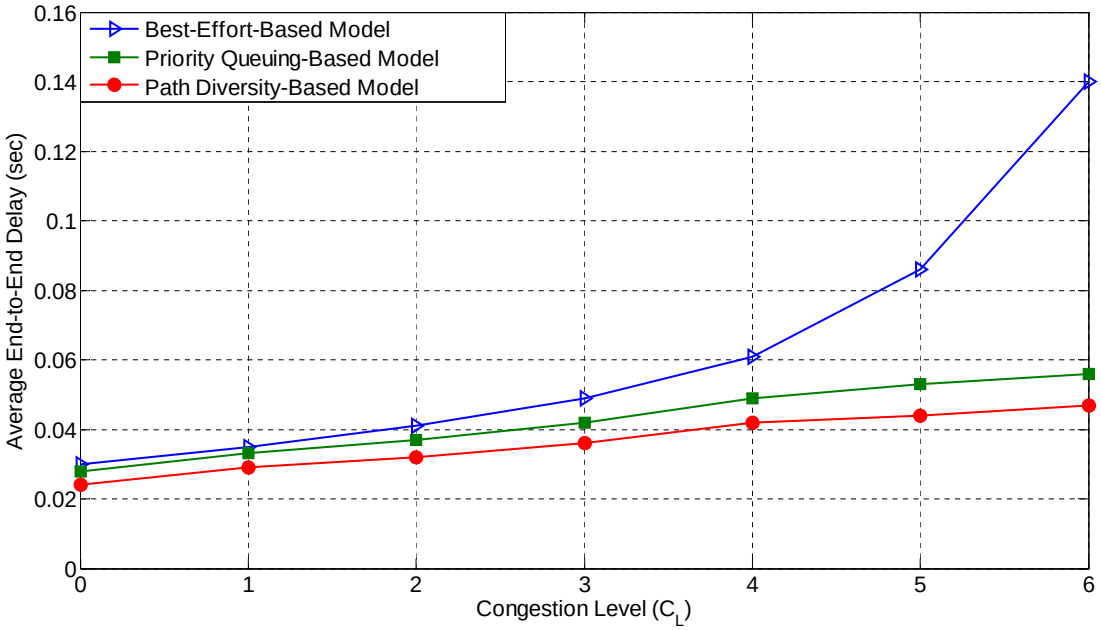


Figure 4.9: Average end-to-end delay of P-frames (the general network topology).

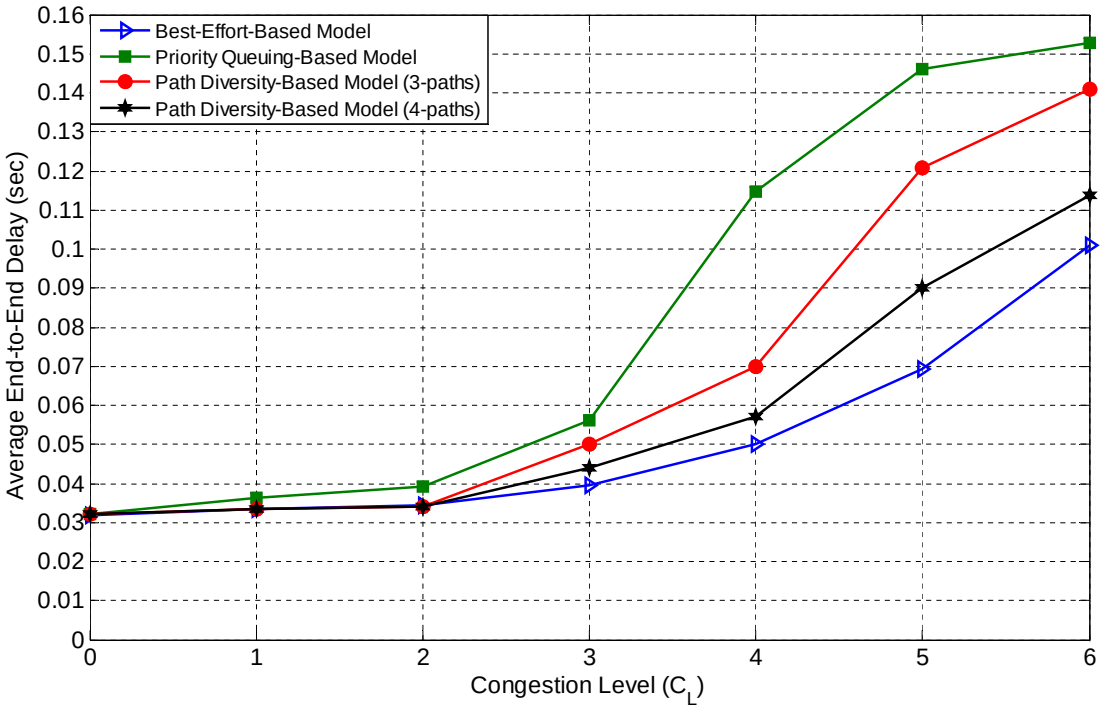


Figure 4.10: Average end-to-end delay of B-frames (the simple network topology).

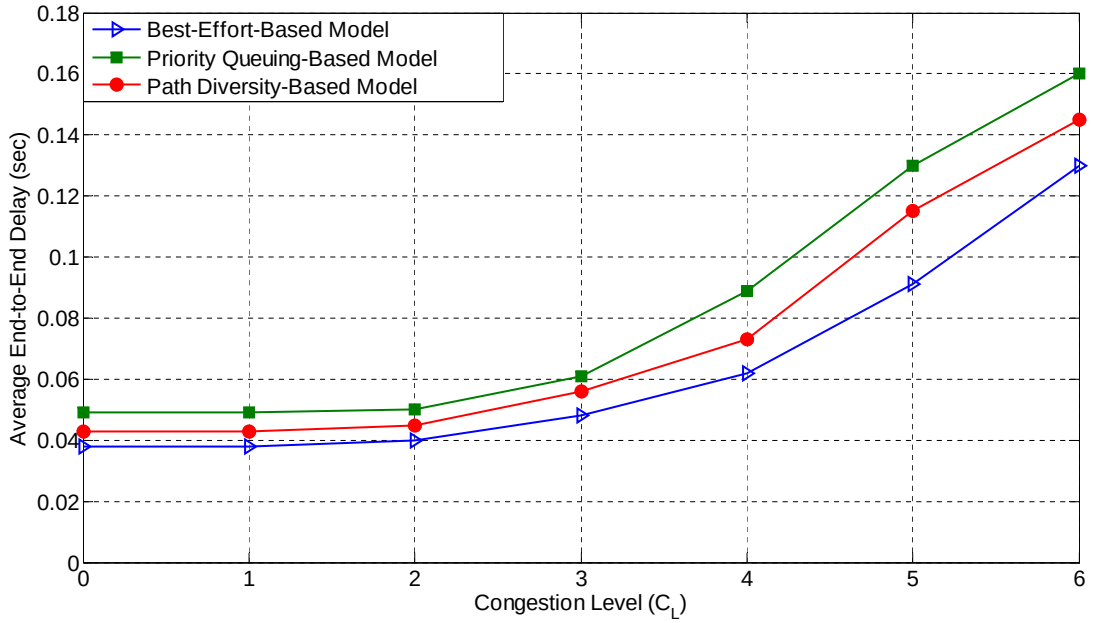


Figure 4.11: Average end-to-end delay of B-frames (the general network topology).

Figures 4.10 and 4.11, show the average end-to-end delay of B-frames versus C_L for the two network topologies. Unlike P- and I-frames, B-frames experienced the lowest end-to-end delay for the best effort-based model. This is explained by the fact that B-frames in the best effort-based model are treated in the same manner as P- and I-frames on a first-come first-serve basis and hence they experience smaller queuing delays.

As C_L increases, the differences in the average end-to-end delay B-frames experience using the best effort-based model becomes more visible in comparison to the other two streaming models. This is expected since as the congestion level increases the priority-based and the path diversity-based models take into considerations the importance of transmitted frames which is not the case in the best effort-based model. The average end-to-end delay figures of the B-frames also assert that by employing the priority queuing system and the PBPD scheme in the path diversity-based model, the performance is improved in comparison to the use of only the priority queuing system in the priority-based queuing model. They also indicate that the average end-to-end delay reduces as the number of available paths increases.

Figure 4.12, shows the histogram of the achieved end-to-end delay for the three models for the worst congestion level considered in the simulations (i.e $C_L = 6$). It is once again clear from the figure that the best effort-based model is the worst in the achieved end-to-end delay. This can be derived from the relatively high number of I- and P-frames experiencing average delays higher than 0.16 seconds which results in incorrect decoding of all dependent frames which in turns causes error propagation and possible severe interruptions in the playback process. In studying the proposed schemes, no deadline was set for video frames. However, in real scenarios, video applications require the delivery of video frames before a certain deadline.

Figures 4.12-(b), 4.12-(c) and 4.12-(d) indicate that when the importance of frames are taken into account, the achieved average end-to-end delay improves when compared to the case of the best effort-based model. Figures 4.12-(c) and 4.12-(d) also indicate that when congestion along the path is taken into consideration jointly with the

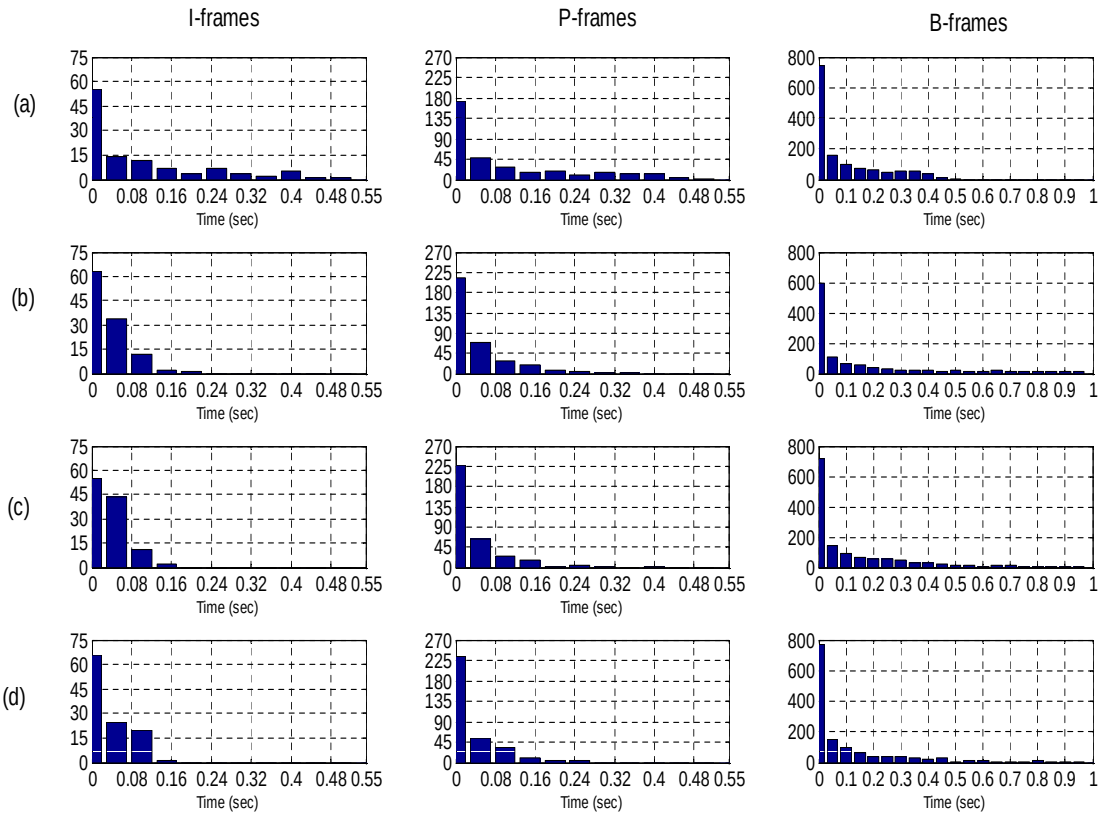


Figure 4.12: Histogram of the average end-to-end delay: (a) Best effort-based model, (b) Priority-based queuing model, (c) Path diversity-based model (3 paths), (d) Path diversity-based model (4 paths).

importance of frames, the average end-to-end delay further improves. It is also clear from Figure 4.12 that as the number of available paths increases, important frames (I, P) experience lower delays. Similarly, B-frames in the best effort-based model experience the lowest end to-end delay when compared to the other two streaming models. Again, this is because B-frames in the best effort-based model are treated in the same manner as P- and I-frames on a first-come first-serve basis and hence experience smaller queuing delays.

4.2.2 Frame Loss Rate. As the primary video source transmits video packets to the end-user, a number of video packets will be lost in the network because of congestion at the intermediate routers. More specifically, each router in the network

has a finite capacity; hence, if the arrival rate to a specific router's input port exceeds the departure rate at some time intervals, packets will accumulate in the router's queue causing any additional arriving packet to be dropped.

In this section, the performance of the three simulated models is compared in terms of the frame loss rate for the different types of video frames (I-frame, P-frame and B-frame). While the first part studies the performance of the proposed PBPD scheme implemented in the network of Figure 4.2, the second part studies the generalized PBPD scheme implemented in the network of Figure 4.3. Moreover, the results are obtained for two GoP lengths to study the effect of the GoP length, if any, on the frame loss rates achieved by the streaming approaches. Namely, GoP length 16 and GoP length 32 are studied in which the structure of the two GoPs has the following sequence $I_0B_1B_2B_3P_4B_5B_6B_7P_8B_9B_{10}B_{11}P_{12}....$

Furthermore, to have a more accurate representation of the frame loss rate results, the confidence interval is found for the frame loss rates for the different models. The confidence interval is a way to represent the information by an average value and an assessment of the error that might be considered when estimating the real average value. The confidence interval provides a more meaningful representation because it tells how close the estimated average value from the real value. It is calculated using the following formula:

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \tag{4.1}$$

where $\overline{x}$ is the estimated average value that follows a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, n is the number of runs, $1-\alpha$ is called the degree of confidence and $z_{\alpha/2}$ is the value in the standard normal distribution with probability $\alpha/2$. This equation requires the sample space (n) over which the sample mean is calculated to be large enough that is at least 30 so that the sample mean is normally distributed. If the sample space is less than 30, the sample mean is modeled by a t-distribution and the

following equation should be used:

$$\overline{x} - t_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}. \tag{4.2}$$

Using any of the confidence interval equations, it can be said that the real average value falls in the confidence interval with a probability of $(1 - \alpha)\%$. Commonly, the values used for α are 0.95 and 0.99 in which a 95% and a 99% confidence intervals are obtained.

In this work, the number of repeated simulation runs represents the sample space (n) in which the frame loss rate is found for each run. The average frame loss rate of these runs is the sample mean ($\overline{x}$). The simulation is repeated until the confidence intervals are narrow. This happened when the simulation is repeated 20 times (n=20), where 300 frames are transmitted at each simulation run. Thus, Equation 4.2 is used to calculate the confidence intervals.

Simple Network Topology

In this section, the simple PBPD scheme is studied in comparison to the priority-based queuing model and the best effort model in terms of the achieved frame loss rate for the highest C_L value considered in this simulation.

Table 4.1: Frame loss rate of GoP 16 for the simple network topology.

Frame Type	Best Effort	Priority Queuing	Path Diversity
I-frames	0.35%	0.05%	0%
P-frames	1.5%	0.1%	0.02%
B-frames	6.2%	8.8%	6.5%

Table 4.2: Frame loss rate of GoP 32 for the simple network topology.

Frame Type	Best Effort	Priority Queuing	Path Diversity
I-frames	0.2%	0%	0%
P-frames	1.7%	0.2%	0.02%
B-frames	6.1%	8.7%	6.5%

Tables 4.3 and 4.4, indicate the frame loss rates for the different frame types of GoP length 16 and GoP length 32, respectively. The tables clearly show that for the I-frames and P-frames, the frame loss rates have significantly improved for the priority-based queuing model and the path diversity-based model. These improvements in the frame loss rates of I- and P-frames are expected to provide a better video quality at the end-user as will be demonstrated in the next chapter.

Unlike I-frames and P-frames, B-frames have the lowest frame loss rate when transmitted using the best effort-based model. This is explained by noting that B-frames are given the lowest priority level in both the priority-based queuing model and the path diversity-based model. Thus, in terms of queuing and routing, B-frames in the priority-based queuing model and the path diversity-based model are served with the least importance where better services are provided to reference frames (I-frames and P-frames). This has resulted in lower frame loss rate for B-frames when served in a fairly manner which is the case in the best effort-based model.

Another point to notice is although the path diversity-based model assigns the lowest priority level to B-frames, the frame loss rate of B-frames is lower in the path diversity-based model than that of the priority-based queuing model. This is because, unlike the priority-based queuing model, the path diversity-based model uses the PBPD scheme in which B-frames are routed to the least congested path. In more details, in the priority-based queuing model, B-frames are transmitted without considering the congestion along the different paths. On the other hand, the path diversity-based model selects the minimum congested path for a transmitted B-frame; thus, lower frame loss rates are achieved in the path diversity-based model.

Another point to study is the comparison between the I-, P- and B-frame loss rates in each simulated model. It is noticed that I-frames had the lowest frame loss rates compared to the frame loss rates of the P- and B-frames in each model. Moreover, P-frames had a lower frame loss rates compared to B-frames in each simulated model. In particular, this might be mistakenly contradicted with the fact that packets are treated equally in best effort networks. Nevertheless, this is due to the lower number of transmitted I-frames compared to the number of P-frames which is in turn lower than the number of B-frames. For example, in the GoP of length 16 frames, in every 16 frames there is only one I-frame, whereas 3 frames are P-frames and 12 frames are B-frames. Thus, lower fraction of the total number of transmitted video packets will be lost from the transmitted I-frames. This is true for the best effort-based model in which packets are equally probable to be lost in the network. On the other hand, the priority and path diversity-based models are designed to reduce the losses in reference frames; thus, the frame loss rate of reference frames have been reduced.

Finally, the last point to consider is the difference in performance between the simulated models when a video is encoded with different GoP sizes. The tables indicate that the frame loss rates of I-frames and B-frames in the best effort-based model and the priority-based queuing model of GoP 32 are lower than the corresponding results of GoP 16. On the other hand, P-frames have larger frame loss rates in GoP 32 than that of GoP 16. This can be explained by highlighting the differences between the two GoPs. For the same video structure, when the video is encoded in different GoP lengths the number of I-frames and P-frames presented in each GoP length will be different. This can be better explained in Figure 4.13 which shows the frame types contained in a 33 frames of the two GoP lengths. As shown in the figure, lower number of I-frames are presented in GoP32 which will result in a lower frame loss rate of I-frames. Conversely, larger number of P-frames are presented in GoP 32 which will result in a larger frame loss rate of P-frames. Another way to explain this difference in the priority-based queuing model is that when the number of I-frames is reduced in the video stream of GoP 32, the arrival rate to the I-queue has been reduced as well which

will in turn reduce the losses in the I-frames. However, when the number of P-frames has increased in GoP 32, the arrival rate to the P-queue has increased which will result in more dropping of P-frames.

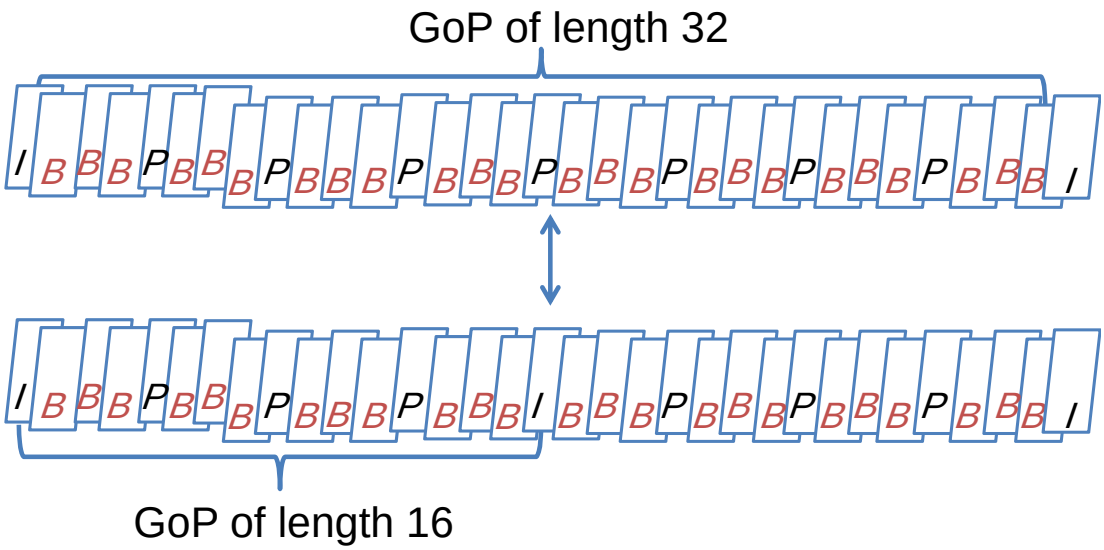


Figure 4.13: GoPs of length 16 and 32.

Another difference between video sequences with GoPs of sizes 16 and 32 is that the lower number of I-frames presented in GoP 32 will lower the average frame sizes of the video stream. The average frame sizes of the tested video sequence considered in this simulation is larger by 7.5% when encoded with GoP of length 16 than that encoded with GoP of length 32. This is true because I-frames are intra-frame coded (coded as a still-image) so they are of larger sizes compared to the P- and B-frames. Hence, if the number of I-frames when compressing a video stream is reduced (by increasing the length of the GoP), the average frame sizes will be reduced as well. In what follows, the queuing and serving delays of routers are discussed to study the effect of reducing the average frame sizes on the transmission process.

A router in the network can be modeled as a queuing system and a server. While the queuing system holds arriving packets until they are being served, the server represents the delay associated with forwarding the packets from the input ports to the appropriate output ports in which the service delay depends on the packets size. More

specifically, the service delay of a specific packet is: $T_S = \frac{L}{S}$, where L is the packet size and S is the speed of the server which is set to a specific fixed value for each router. Therefore, the effect of reducing the average frame sizes on the network performance will be increasing the average departure rate. This can conclude the following: having the arrival rate and the service speed of a router to be the same (for both GoP lengths), increasing the departure rate will lower the dropping in the system. This can be seen from the lower frame loss rate of B-frames and the lower overall losses percentage in GoP 32 compared to that of GoP 16. These differences are more significant in the general network topology as will be seen in the next section.

Finally, the path diversity-based model had the same frame loss rate for the two GoP lengths. This can be explained by noting that the arrival rate of the video stream to the router's input port in the path diversity-based model will vary based on the relative congestion state of the router to the other routers in the network. More specifically, the path diversity-based model employs the PBPD scheme in which video packets are transmitted to different paths in response to network congestion. Hence, the arrival rate to a specific router is not necessarily the same for the two GoP lengths. This was not the case in the best effort-based model and the priority-based queuing model in which video packets are transmitted through a single lowest cost path; thus, the arrival rate to the router on the selected best path will be the same for the two GoPs.

In addition to frame loss rates, Figure 4.14 plots the confidence intervals for the two GoP lengths. Moreover, the figure shows two intervals for each frame loss rates. Intuitively, when the degree of certainty increases, the confidence interval becomes wider; thus, it tells less about the estimated variable. As indicated in the frame loss rates, the figure shows a noticeable reduction in the average frame loss rate of the I-frames and P-frames when transmitted over the priority-based queuing model and the path diversity-based model. The priority-based queuing model outperforms the best effort-based model and achieves lower loss rates for the P- and I-frames. The reduction further increases in the path diversity-based model in which the two confidence limits of the loss rate of I-frames falls into zero. This indicates that there is a 99% chance for

the average frame loss rate of I-frames transmitted using the path diversity-based model to be zero at the highest C_L value considered in this study. In contrast, one can notice that the frame loss rate of B-frames has increased in the priority-based queuing model compared to the best effort-based model. As discussed previously, this is because of the low priority level assigned to the B-frames in the priority-based queuing and the path diversity-based models which resulted in lower frame loss rates of B-frames when served in fairly manner as in the best effort-based model.

Generalized PBPD Scheme

In this section, the frame loss rate results are obtained for the network topology of Figure 4.3 in which the PBPD scheme is employed in a general network topology.

The results show a similar behavior to the results obtained for the PBPD scheme in the simple network topology. Again, the frame loss rates of reference frames have significantly improved in the path diversity-based model and the priority-based queuing model. This indicates the effectiveness of using the PBPD scheme in general network topologies in reducing the frame loss rates of reference frames. Moreover, B-frames had the lowest frame loss rates in the best effort-based model. The results are further confirmed in the confidence intervals shown in Figure 4.15.

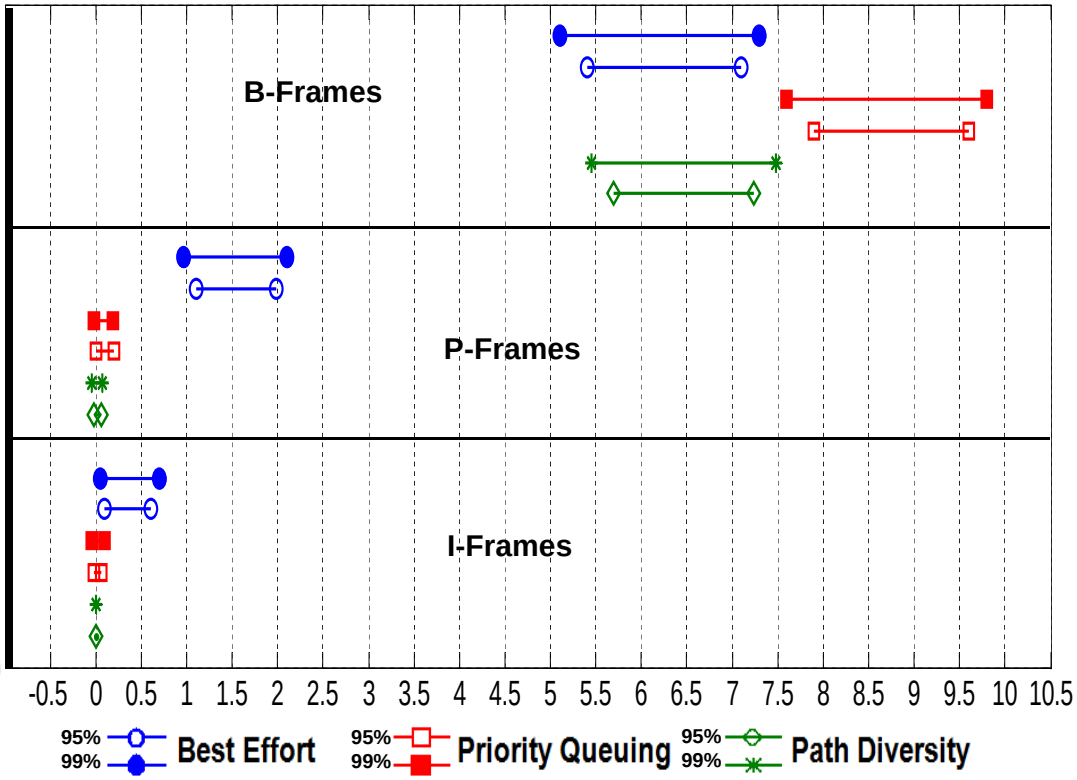
Table 4.3: Frame loss rate of GoP 16 for the general network topology.

Frame Type	Best Effort	Priority Queuing	Path Diversity
I-frames	1.4%	0.07%	0%
P-frames	1.6%	0.39%	0.1%
B-frames	9.3%	10.1%	8.8%

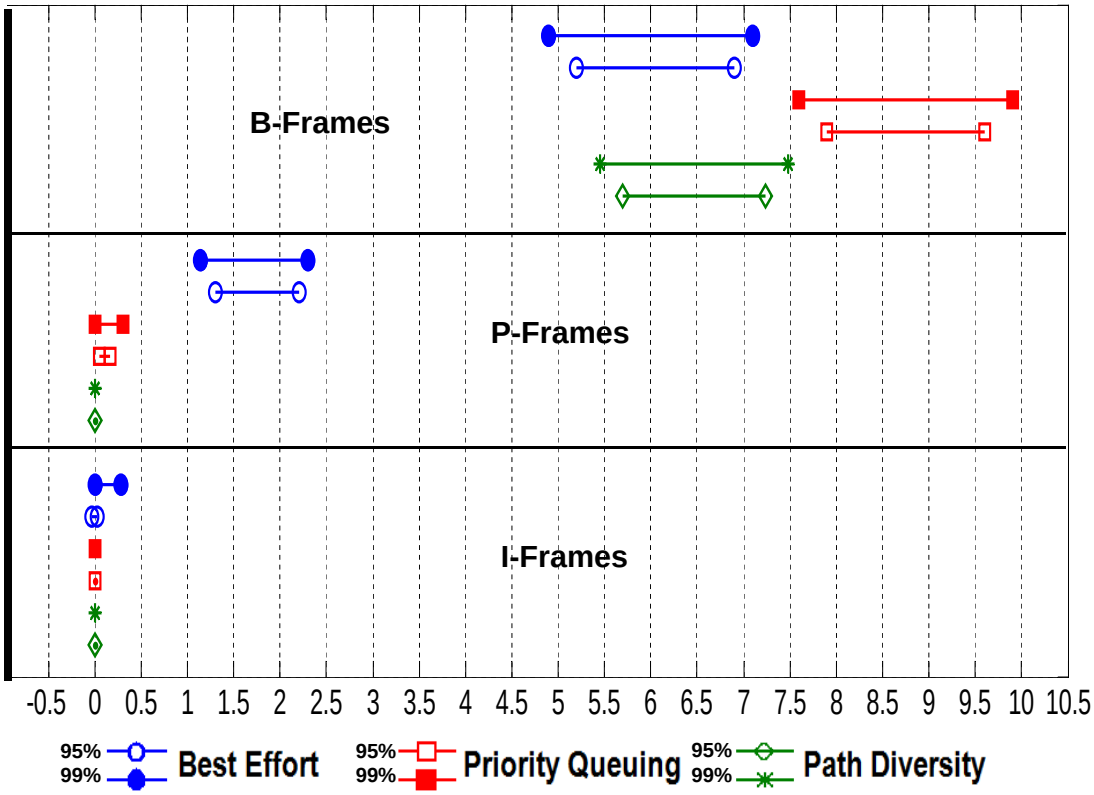
Table 4.4: Frame loss rate of GoP 32 for the general network topology.

Frame Type	Best Effort	Priority Queuing	Path Diversity
I-frames	0.4%	0.05%	0%
P-frames	2.1%	0.4%	0.1%
B-frames	8.4%	9.5%	8.8%

However, one difference is found in the relative frame loss rates of the priority-based queuing model when a video stream of GoP 16 is transmitted compared to that of GoP 32. In the simple network topology, it was found that in the priority-based queuing model the frame loss rate of I-frames have decreased in GoP 32 and the frame loss rate of P-frames have increased compared to the corresponding results in GoP 16. This difference in performance of the priority-based queuing model of the two GoP lengths has effectively been reduced in the general network topology. This can be explained by noting that, while the best path in the simple network topology consists of a single router, the best path (minimum number of hops) in the general network topology consists of two routers. Although the arrival rate of I-frames will decrease and the arrival rate of P-frames will increase in the first router of the general network topology when using GoP 32, this might not be the case for the second router in the network. More specifically, the relative departure rates of the I-frames and P-frames from the first router in the path when using GoP 16 and GoP 32 cannot be easily predicted due to the random characteristics of the traffic.

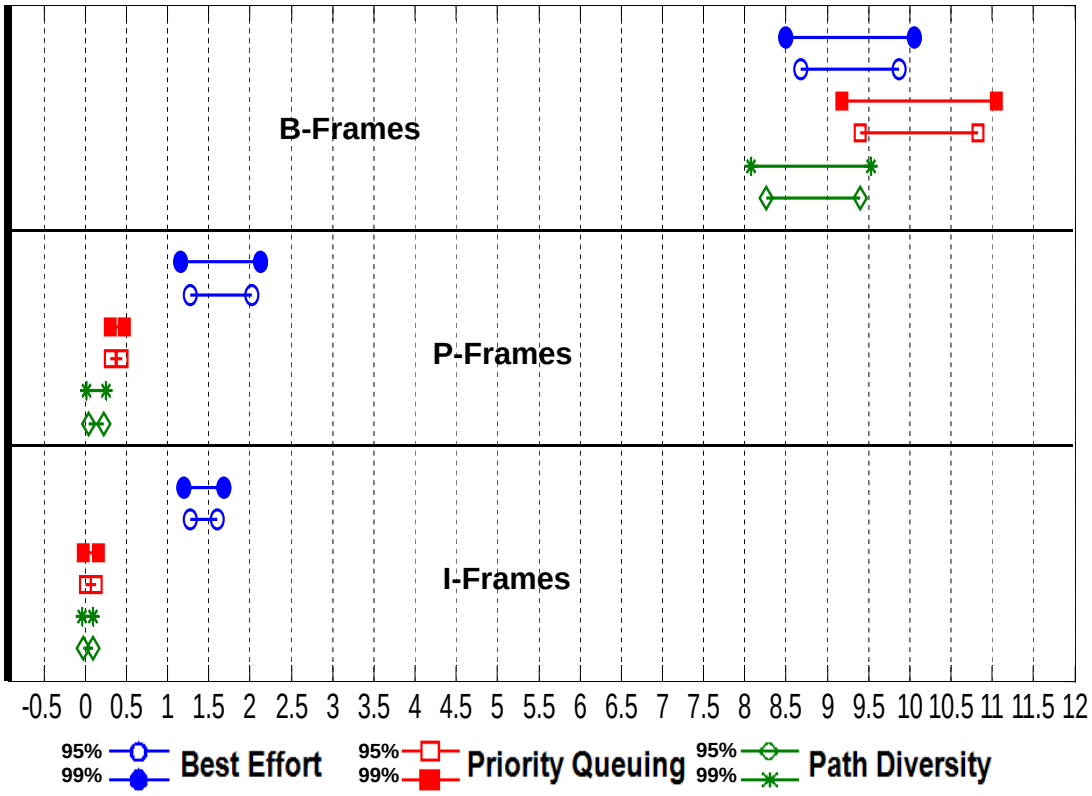


(a) GoP 16

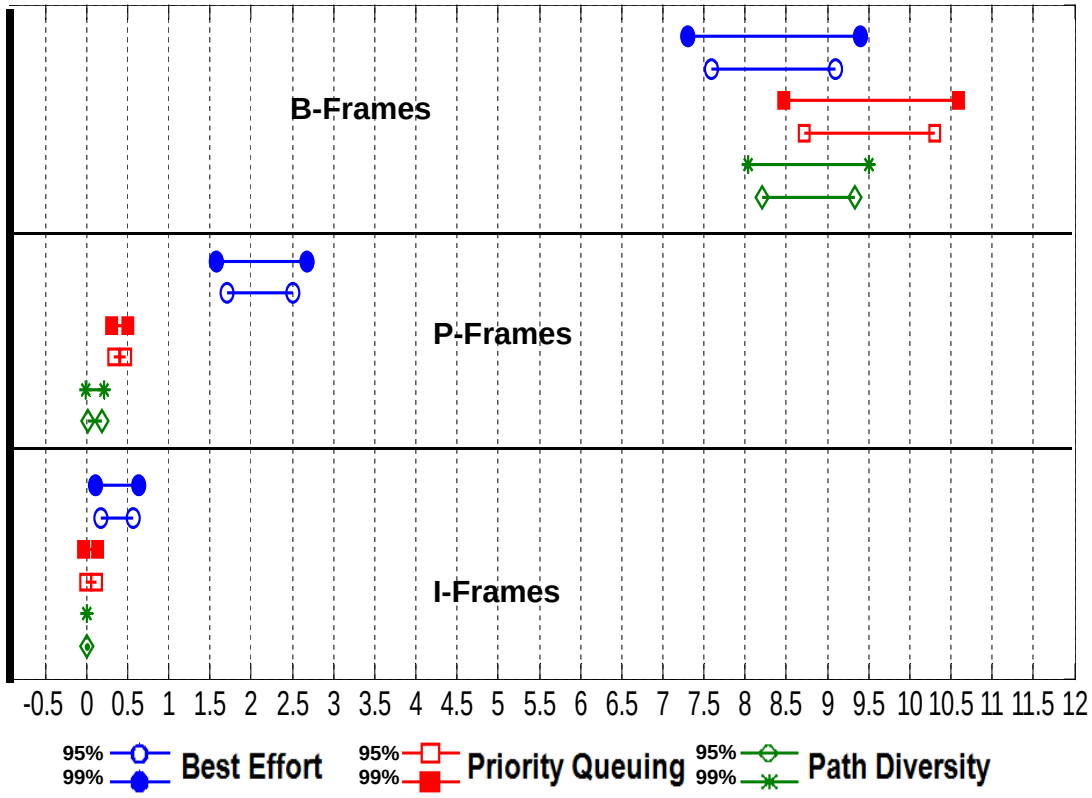


(b) GoP 32

Figure 4.14: Confidence intervals for the frame loss rates of the simple network topology.



(a) GoP 16



(b) GoP 32

Figure 4.15: Confidence intervals for the frame loss rates of the general network topology.

Chapter 5: Quality Assessment of the Reconstructed Video using the Proposed Approach

In this chapter, the performance of the priority based path diversity (PBPD) scheme is evaluated in terms of the achieved video quality. For this purpose, two performance metrics are studied namely, the peak-signal-to-noise ratio (PSNR) to measure the spatial video quality and the buffer occupancy to study the continuity of the reconstructed video. The performance of the PBPD scheme is studied in comparison to the priority-based queuing model and the best effort-based model. Moreover, an overview of the streaming process is presented at the start of the chapter to help demonstrate the indications of the used metrics.

5.1 Video Streaming System

In what follows we briefly discuss the video streaming process and the actions taken by the receiver on the occasions of lost/delayed video frames to be able to understand the indications of the obtained results.

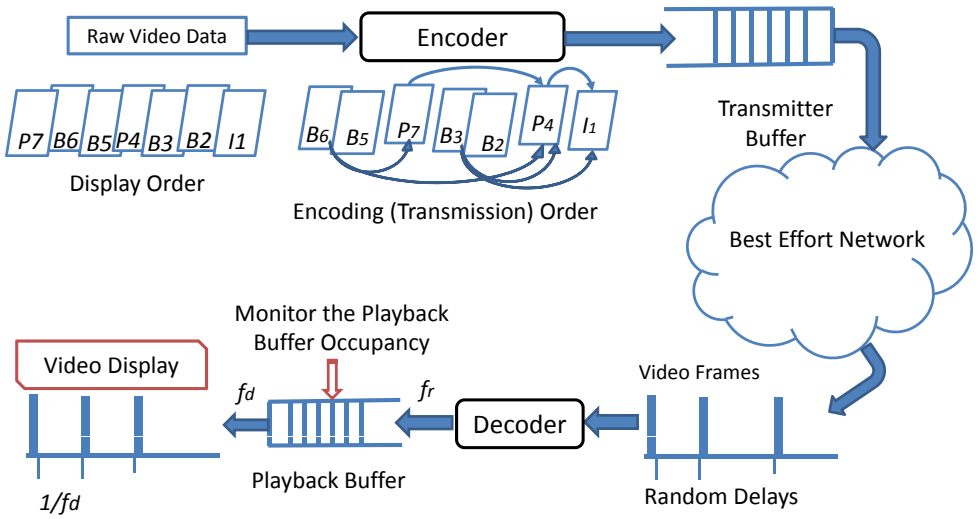


Figure 5.1: Block diagram of a video streaming system.

Figure 5.1 shows the basic building blocks of a video streaming system. The video encoder converts the high data rates of the raw video to lower data rates that are

more suitable for transmission. The encoded video stream is then divided into one or more packets depending on the size of the video frame and the underlying network. Video packets are stored in the transmitter buffer after which they will be transmitted at a transmission rate that depends on the underlying network conditions. When transmitted over the Internet, video packets experience varying end-to-end delays and packet loss rates which depend on the network congestion.

These losses and delays will affect the temporal and spatial quality of the reconstructed video. Typically, the PSNR metric is used to measure the spatial video quality. It is calculated using:

$$\text{PSNR} = 10\log_{10}\left(\frac{1}{XY}\sum_{i=0}^{X-1}\sum_{j=0}^{Y-1}\frac{255^2}{|f(i,j)-g(i,j)|^2}\right), \tag{5.1}$$

where $f(i,j)$ is the (i,j)th pixel value in the original video frame, $g(i,j)$ is the (i,j)th pixel value in the reconstructed video frame, 255 is the largest possible pixel value (peak) when the number of bits per pixel is 8, and X and Y are the horizontal and vertical dimensions of the video frame. On the other hand, the continuity can be examined by monitoring the playback buffer occupancy. In what follows the role of the playback buffer occupancy is studied to demonstrate how it can be used to reflect the smoothness of the displayed video.

In video on demand (VoD), at the start of the streaming process, the playback is delayed for a certain period (buffering delay) to allow for buffering sufficient video frames before displaying the video. This buffering delay can be for a fixed period of time (eg. 10 seconds) or a fixed number of decodable frames arrive (eg. one GoP). When the buffering delay ends, video frames will be displayed sequentially (in their display order) at a displaying rate in frames per second which is controlled by the application.

Although the displaying rate is set to a specific value by the application, it actually depends on the rate of successfully decoded frames. This can be demonstrated as follows, video frames are displayed for $1/f_d$ sec at specific time instants depending on

their displaying order. For example, if a 1 min video is to be played at 30 frames/sec, a total of 1800 frames are displayed each for $1/30$ sec of the full 1 min video so that a continuous motion is perceived by the viewer. On the occasions of frame loss, the receiver action depends on the employed error concealment mechanism. For example, it could pause the last displayed video frame or it might display the last received reference frame, at the displaying instant of the lost frame. Nevertheless, lost frames will result in discontinuity in the playback of the video. The situation will be even worse in the case of having consecutive number of frames experiencing long delays or losses; this will cause the buffer to under flow and probably starve. This will in turn lead to a disturbed video playback. From the above discussion, it can be seen that a key point to study the smoothness of the displayed video is by monitoring the playback buffer occupancy.

5.2 Simulation Results

In the previous chapter, it was shown that the proposed PBPD scheme outperforms the priority-based queuing model and the best effort-based model in terms of network related metrics such as the end-to-end delay and the frame loss rate. This section, however, evaluates the performance of the proposed PBPD scheme in terms of the improvement achieved in the streaming process from an end user's perspective. Specifically, the performance of the three simulated models, namely: path diversity-based model, priority-based queuing model and best effort-based model, are compared in terms of the PSNR and the buffer occupancy to assess the playback process for the three cases. For this purpose, the H.264/AVC Joint Model reference software [55] [56] is used to encode four video sequences: crew, mobile, foreman and walk. At the decoder, a copy of the last arrived reference frame is used for error concealment whenever a frame is lost in the network. In more details, if a video frame is lost in the network a copy of the last received reference frame is displayed in the displaying instant of the lost frame.

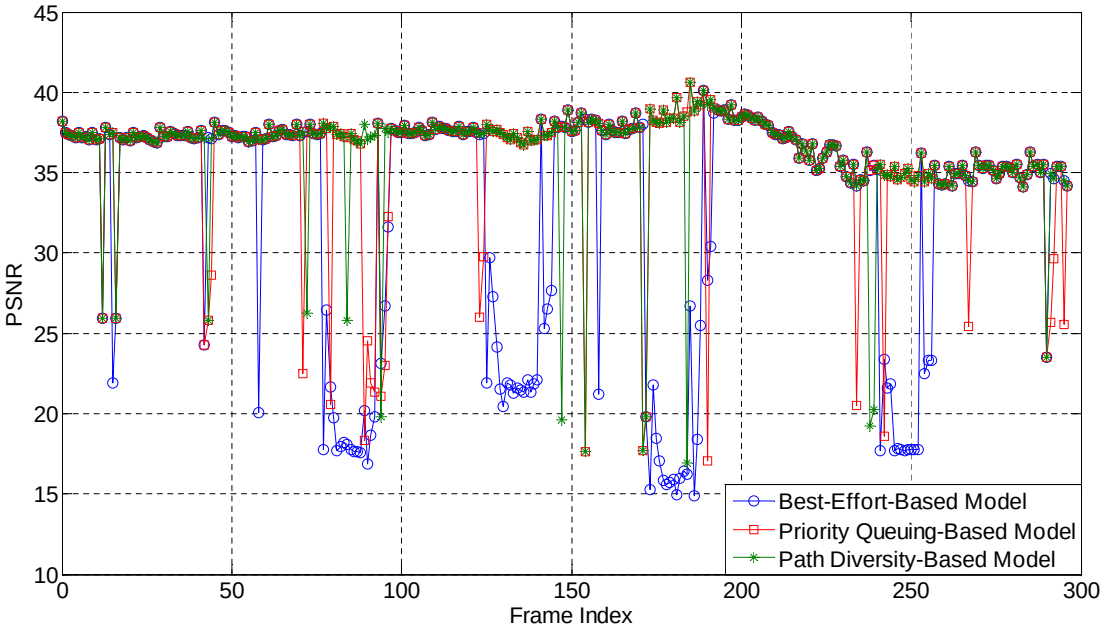
In this simulation, two GoP sizes ($M = 4,\ N = 16$ and $M = 4,\ N = 32$) are considered to study the impact of the GoP size, if any, on the streaming process. The GoP structure used for the two GoP sizes is $I_0B_1B_2B_3P_4B_5B_6B_7P_8B_9B_{10}B_{11}$,

where the total number of frames in the GoP is N and $(M-1)$ B-frames are presented between two reference frames. Each of the four video sequences used in this simulation is encoded in two versions: the first one is with GoP structure ($M = 4,\ N = 16$) and the second one with GoP structure ($M = 4,\ N = 32$).

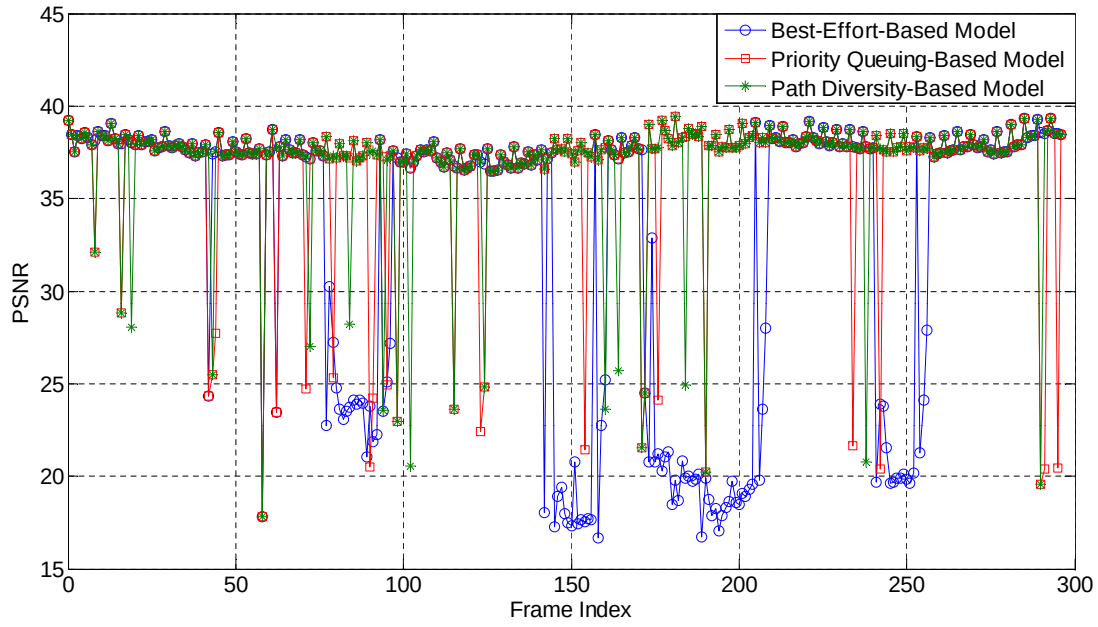
5.2.1 PSNR Results. The primary video source transmits the tested video sequences through the lossy network to the end user. Whenever a frame is lost in the network, the error concealment used replaces the lost frame with the last arrived reference frame. Consequently, if the lost frame is a reference frame (I-frame or P-frame), all dependent frames in the GoP will be decoded with reference to the last arrived reference frame. The PSNR is obtained for the video sequences when transmitted through the three simulated models.

Figures 5.2 and 5.3 show the PSNR of the four video sequences for the GoP of size 16. An important point to highlight when comparing the average PSNR values obtained by the three models is that the drops in the PSNR values of the path diversity-based model are located in single dips. In other words, a drop in the PSNR value of the path diversity-based model occurs in one frame at a time. However, the drop in the PSNR values of the best-effort-based model and the priority-based queuing model occur in one frame at a time and in consecutive number of frames. The same can be noticed in the PSNR values of sequences with GoP of size 32 shown in Figures 5.4 and 5.5.

This is explained by noting that the drop in the PSNR value occurs for one frame at a time only if a B-frame is lost. In this case, the error concealment used will replace the lost frame by the last received reference frame causing a drop in the PSNR value of the lost frame. However, when an I-frame or a P-frame is lost, dependent frames will be decoded with reference to the last arrived reference frame causing error propagation in the GoP. This will not only result in a drop in the PSNR value of the lost reference frame but also of all consecutive dependent frames.

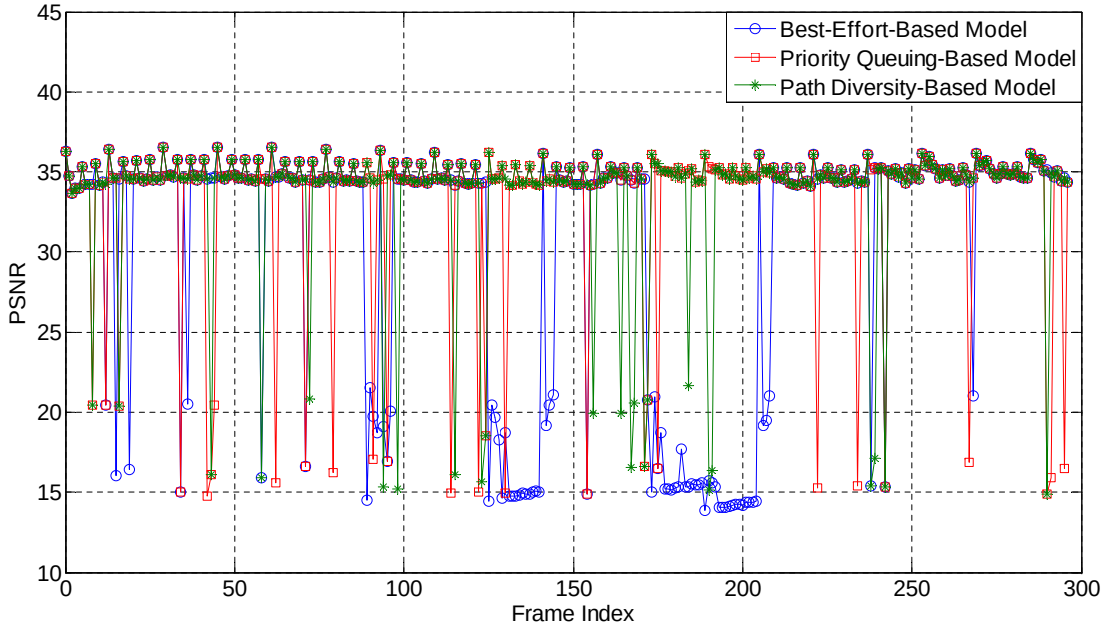


(a) Foreman Sequence

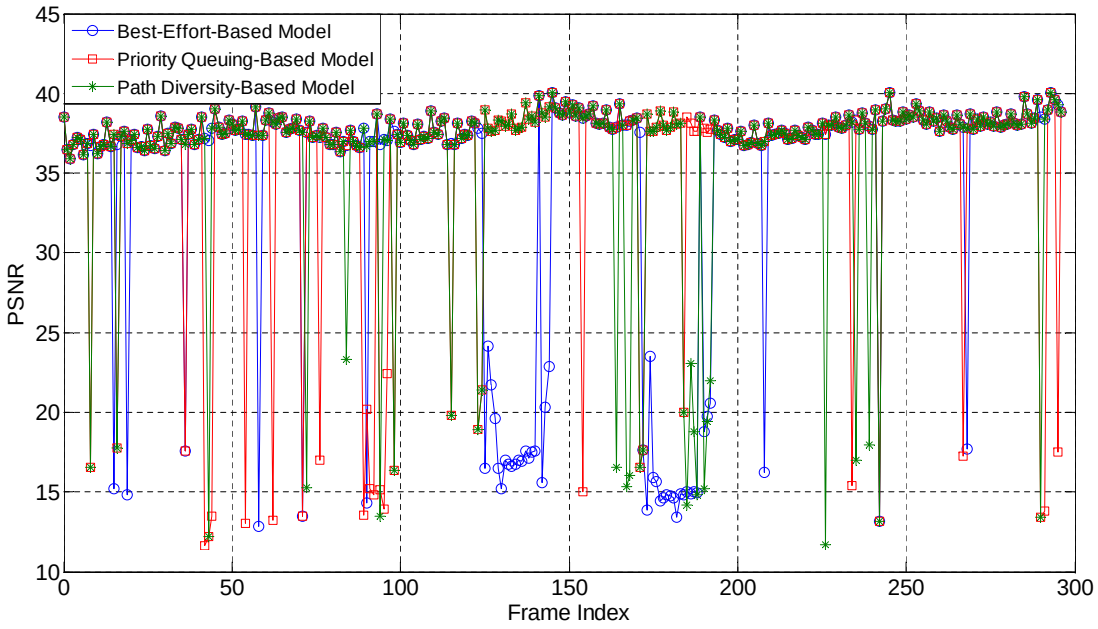


(b) Crew Sequence

Figure 5.2: PSNR of the foreman and crew video sequences for a GoP structure $M = 4$ and $N = 16$.

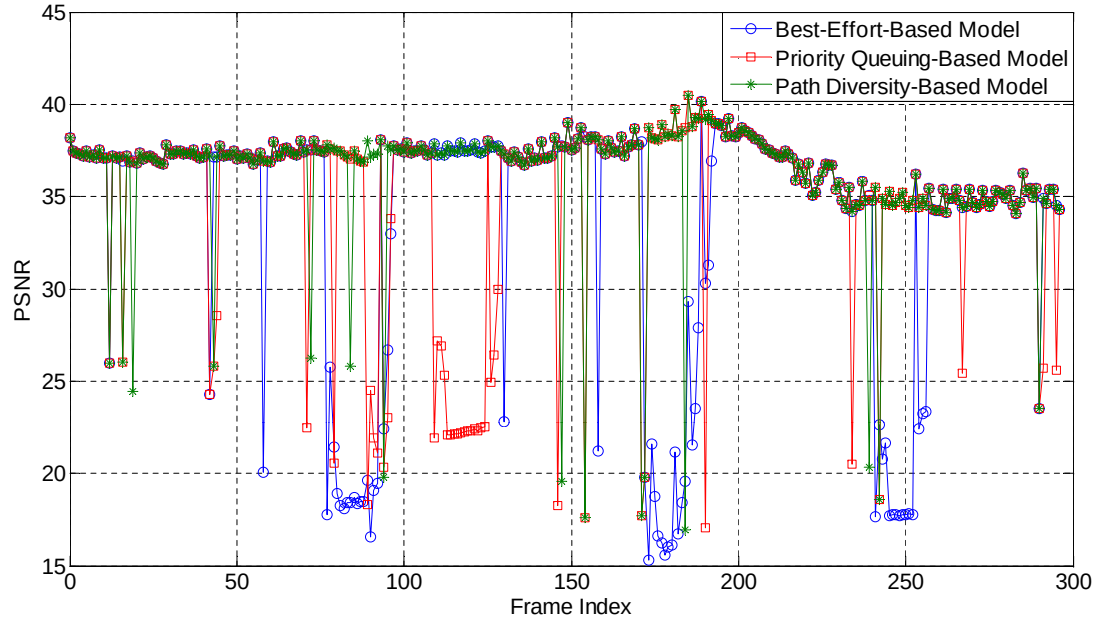


(a) Mobile Sequence

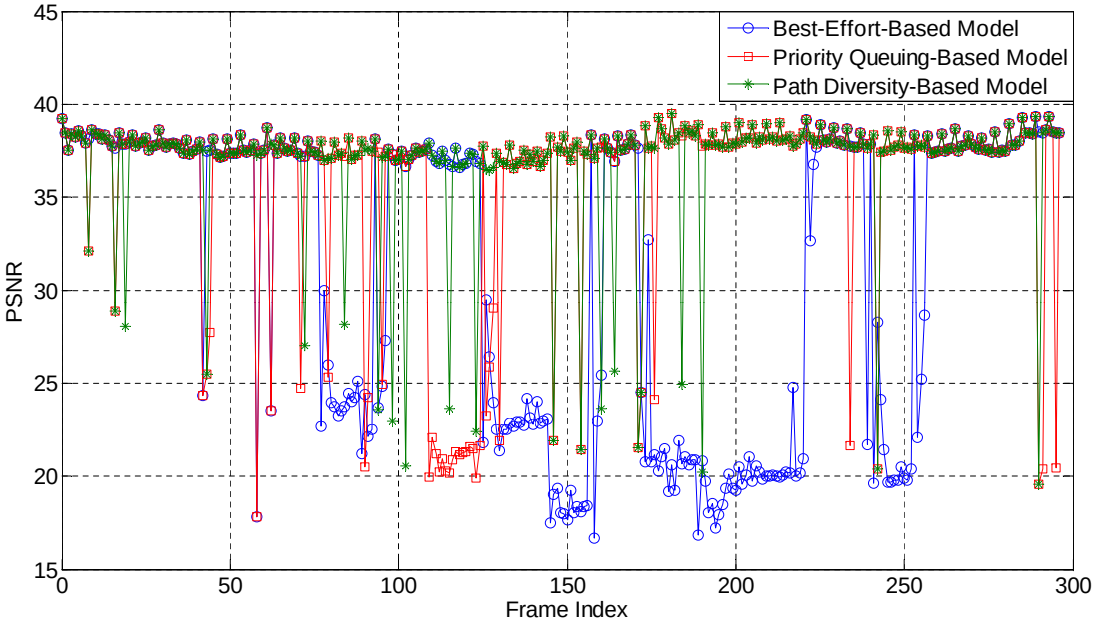


(b) Walk Sequence

Figure 5.3: PSNR of the mobile and walk video sequences for a GoP structure $M = 4$ and $N = 16$.

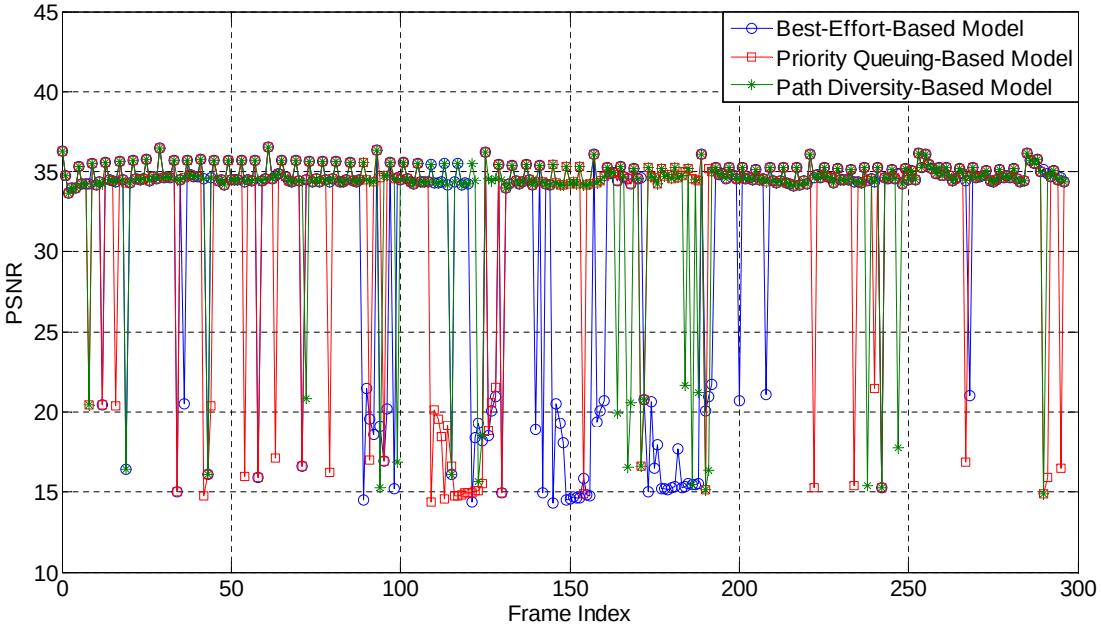


(a) Foreman Sequence

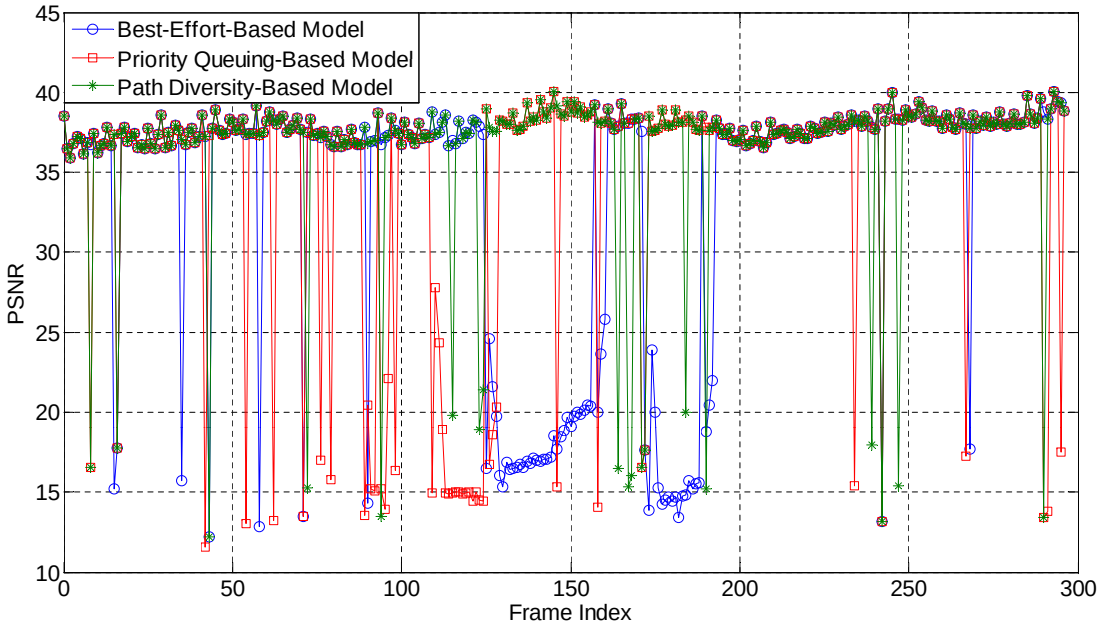


(b) Crew Sequence

Figure 5.4: PSNR of the foreman and crew video sequences for a GoP structure $M = 4$ and $N = 32$.



(a) Mobile Sequence



(b) Walk Sequence

Figure 5.5: PSNR of the mobile and walk video sequences for a GoP structure $M = 4$ and $N = 32$.

Figure 5.6 plots the PSNR of the walk sequence with GoP of size 32 when transmitted using the three different models. The figure focuses on a particular frame interval from frame number 109 to frame number 163, where the frames are plotted in their decoding order. The marked frames are the lost frames. As one can notice, whenever a B-frame is lost a drop in the PSNR value occurs only at the position of that B-frame. Nevertheless, on the occasion of the loss of an I-frame or P-frame, the PSNR of all frames in the GoP is degraded. In the considered frame interval, an I-frame was lost in the best effort-based model specifically, frame number 126. All following frames in the GoP where the I-frame is lost are decoded with reference to the last arrived P-frame in the previous GoP which is in this case frame number 122. As a result, a degradation in the PSNR remains until the reception of the next I-frame (frame number 158). The same degradation is noticed when a P-frame is lost in the priority-based queuing model. However, the degradation in the PSNR when a P-frame is lost is expected to remain for a shorter interval compared to the degradation when an I-frame is lost. This is due to the smaller number of dependent frames on the P-frame compared to that of the I-frame.

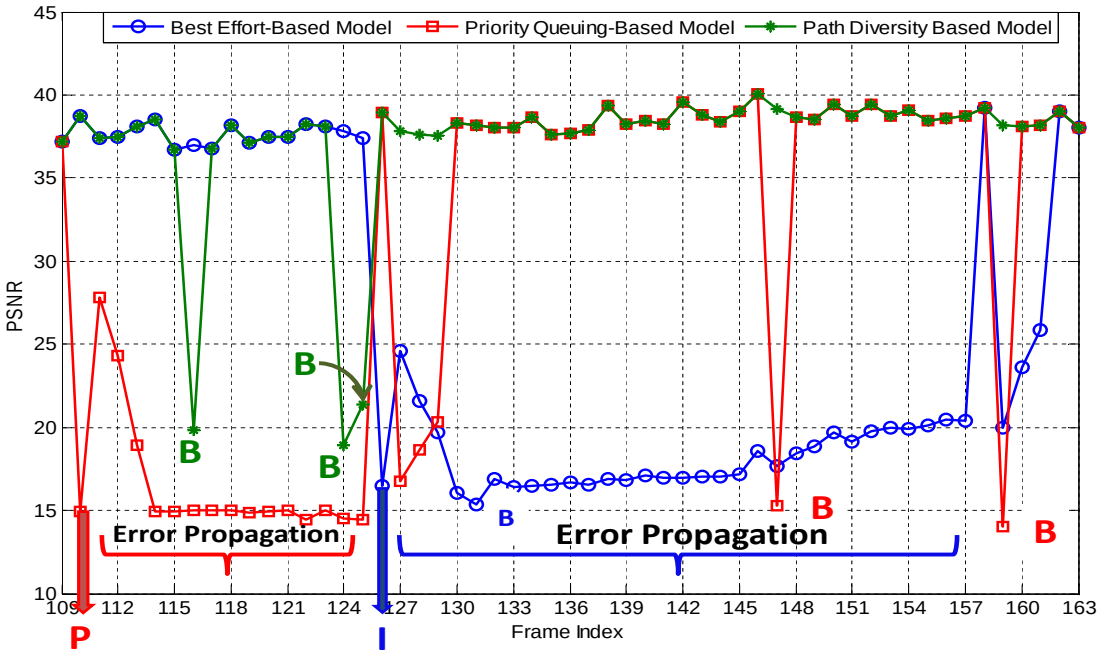


Figure 5.6: PSNR of the walk sequence (GoP $M = 4$ & $N = 32$), which marks the lost frames in the specified frame interval.

The visual artifacts caused by error propagation due to the loss of an I-frame are shown in Figure 5.7 in which a P-frame and a B-frame in the walk sequence are decoded with reference to the last received P-frame. The figure clearly shows the degradation in the perceptual quality caused by error propagation. This is especially feasible because the frames should capture the movement of the man in the video; hence, the incorrect decoding of the bi-directional frame (B-frame) has incorrectly shown an object (the moving hand) which is not there in the corresponding correctly decoded frame of the path diversity-based model.

(a) Best Effort-Based Model

(b) Path Diversity-Based Model

Figure 5.7: Error propagation effect in the walk sequence when a P-frame (frame number 181) and a B-frame (frame number 182) in the best effort-based model are decoded using the last received reference frame.

Another point to notice is the decrease and increase of the PSNR values whenever a reference frame is lost. Considering the loss of the I-frame in the best effort-based model of the walk sequence shown in Figure 5.6, the error concealment used will replace the lost I-frame with the last received P-frame which will result in having the PSNR value of the I-frame to be low (around 17 dB). As previously discussed, the loss of the I-frame will result in unsuccessful decoding of the dependent frames in the GoP. However, the PSNR values of the following dependent frames do not all remain in the 17 dB level. Instead the PSNR of the following frame jumps to a better PSNR after which it gradually decreases to the 17 dB level in two frames. To explain this behavior, the coding dependencies between the frames should be considered. The three frames following the lost I-frame are B-frames that depend on two bidirectional reference frames. While one of the reference frames is lost, which is the I-frame in this example, the other one is received and decoded correctly (P-frame). In their display order, the position of the first B-frame following the lost I-frame is closer to the received reference frame. Therefore, it is decoded with a better PSNR compared to the other two B-frames. As the position of the B-frame becomes closer to the lost I-frame the value of the PSNR decreases.

A similar behavior occurs upon the reception of an I-frame after a previous loss in a reference frame in which a decrease and increase in PSNR values occur. Referring to the same example, when the I-frame of the next GoP is received, the PSNR increases to a high value around 39 dB. However, the PSNR of the following frame drops back to a low value and gradually increases in the next two frames after which it goes back to a high value. Similarly, this is explained by noting that those three frames are B-frames which depend on a correctly received I-frame and an incorrectly decoded P-frame. Although the P-frame might have been received correctly, it was not correctly decoded because of the loss or the incorrect decoding of its reference frame. Once more, the first B-frame which is the closest to the unsuccessfully decoded P-frame had the lowest PSNR compared to the following two B-frames; as the position of the B-frames becomes nearest to the correctly received I-frame, the PSNR increases.

Table 5.1: Average PSNR of the walk sequence in the considered frame interval.

Number of Lost Frames	Best Effort	Priority Queuing	Path Diversity
I-frames	1	0	0
P-frames	0	1	0
B-frames	1	2	3
Average PSNR	24.6 dB	26.5 dB	35.7 dB

To have more insight on the overall effect of the loss of the different frame types on the obtained PSNR, Table 5.1 states the number of lost frames for each type and the average PSNR for the three models in the specific interval of Figure 5.6. As shown in the table, the lowest number of lost frames is obtained in the best effort-based model including an I-frame and a B-frame. However, the best effort-based model had the worst average PSNR. Moreover, the average PSNR of the priority-based queuing model is better by around 2 dBs than the best effort-based model even though the priority-based queuing model have also lost one reference frame (P-frame) and two B-frames. Based on our last demonstration of Figure 5.6 this is due to the longer error propagation resulted from the lost I-frame compared to that of the lost P-frame. Last, the path diversity-based model has the best PSNR as the loss occur only for B-frames which are not referenced by any other frames.

Table 5.2 states the average PSNR found for the video sequences using the three simulated models. The average PSNR results show that the highest average PSNR is obtained for video sequences transmitted over the path diversity-based model. On average, a gain of 3.2 dB was achieved for video sequences of GoP with length 16 when transmitted over the path diversity-based model and an average gain of 2.8 dB when transmitted over the priority-based queuing model. Again comparing the average PSNR values of the GoP length 32, an average of 3.4 dB was achieved using the path diversity-based model and an average of 2 dB was achieved for the priority-based queuing model.

Note that the average PSNR achieved by a specific model will differ from one video sequence to the other. This is due to the different motion content of the video

Table 5.2: Average PSNR.

(a) **GoP Structure:** $M = 4$ **and** $N = 16$

Network Model	Best Effort	Priority Queuing	Path Diversity
Foreman Sequence	32.68 dB	35.81 dB	36.30 dB
Crew Sequence	32.79 dB	36.57 dB	36.92 dB
Mobile Sequence	30.36 dB	33.25 dB	33.55dB
Walk Sequence	34.43 dB	35.62 dB	35.93 dB
Average PSNR	32.56 dB	35.3 dB	35.68 dB

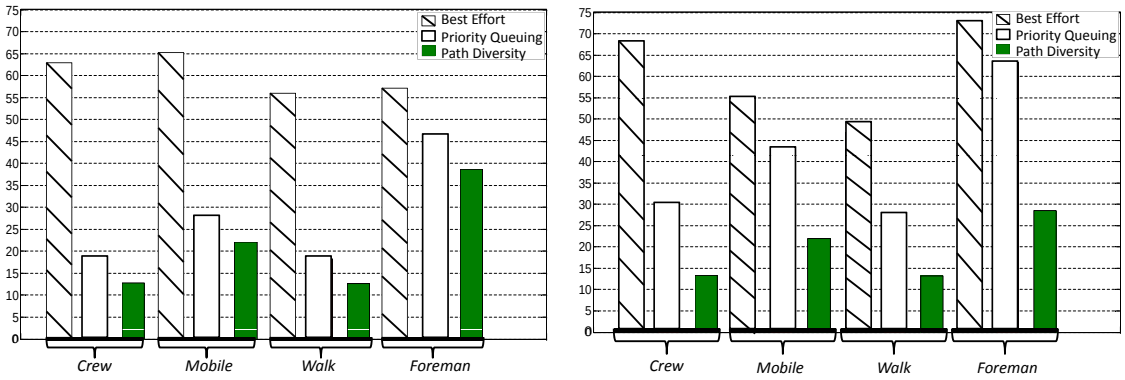
(b) **GoP Structure:** $M = 4$ **and** $N = 32$

Network Model	Best Effort	Priority Queuing	Path Diversity
Foreman Sequence	33.66 dB	34.87 dB	36.19 db
Crew Sequence	31.12 dB	35.72 dB	36.85 dB
Mobile Sequence	31.00 dB	32.03 dB	33.45 dB
Walk Sequence	33.38 dB	34.60 dB	36.50 dB
Average PSNR	32.29 dB	34.31 dB	35.75 dB

sequences, which will result in different average PSNR when using the error concealment discussed previously. In more details, suppose that an I-frame is lost in two video sequences, then the error concealment used will replace the lost I-frame with the last received reference frame; the average degradation in the PSNR will depend on how similar is the lost I-frame from the last received reference frame. This in turn depends on the content of the video. While low motion video will have more similarity between video frames resulting in lower average PSNR degradation, high motion video will have less similarity between adjacent video frames resulting in larger average PSNR degradation.

Another probabilistic measure that helps to compare the average PSNR obtained using the three models is the variance which measures how close the PSNR values are from the average PSNR value. A small variance indicates that the PSNR values are close to the mean value, while a large variance indicates that the PNSR values are spread away from the average PSNR. Figure 5.8, shows the variance of the obtained PSNR values for the different video sequences. The figure shows that the path diversity-based model has the lowest variance. This means that the PSNR values of the individual frames are close to the average PSNR value which indicates a better perceived video quality.

The difference in the obtained variance of each video sequence can be explained by the nature of the video (the different motion content of the video sequences). However, one can notice that the variance of the priority-based queuing model of GoP32 is always larger than that of GoP16, for all video sequences. This highlights the importance of considering the impact of the GoP size on the streaming process.



(a) GoP $M = 4$, $N = 16$ (b) GoP $M = 4$, $N = 32$

Figure 5.8: Variance of the PSNR values for the tested video sequences.

Two important points should be considered when comparing the average PSNR values of the two GoP lengths. First, GoPs with different structure contain different numbers of I-,P- and B-frames. These differences in the number of frames have affected the loss rates of each frame type as discussed in the previous chapter. In summary, the number of I-frames has decreased in GoP32 which resulted in a lower percentage loss of I-frames. This however was translated into larger number of P-frames in GoP32 which resulted in larger percentage drop in GoP32. This is true for the priority-based queuing model and the best effort-based model in which video packets are propagated along a single lowest cost path.

The second important point is, the effect of losing a reference frame in one GoP is related to the number of dependent frames which in turn depends on the length of the GoP and the location of the lost reference frame in the GoP (eg. P-frame). In more details, Figure 5.9 plots a scenario of losing the same I-frame in the two GoP sizes. This figure focuses on the specific part of the crew video sequence in which the error

propagation took place upon the loss of an I-frame. In the case of GoP 32 the error propagates among the GoP causing a degradation in the PSNR of 31 frames until the correct reception of the next I-frame. On the other hand, in the case of GoP 16 the error propagates among 15 frames causing a degradation in a lower number of frames compared to the case of GoP 32; hence, resulting in a better average PSNR. These two points should be considered when explaining the different percentage drops for the two GoPs with different length.

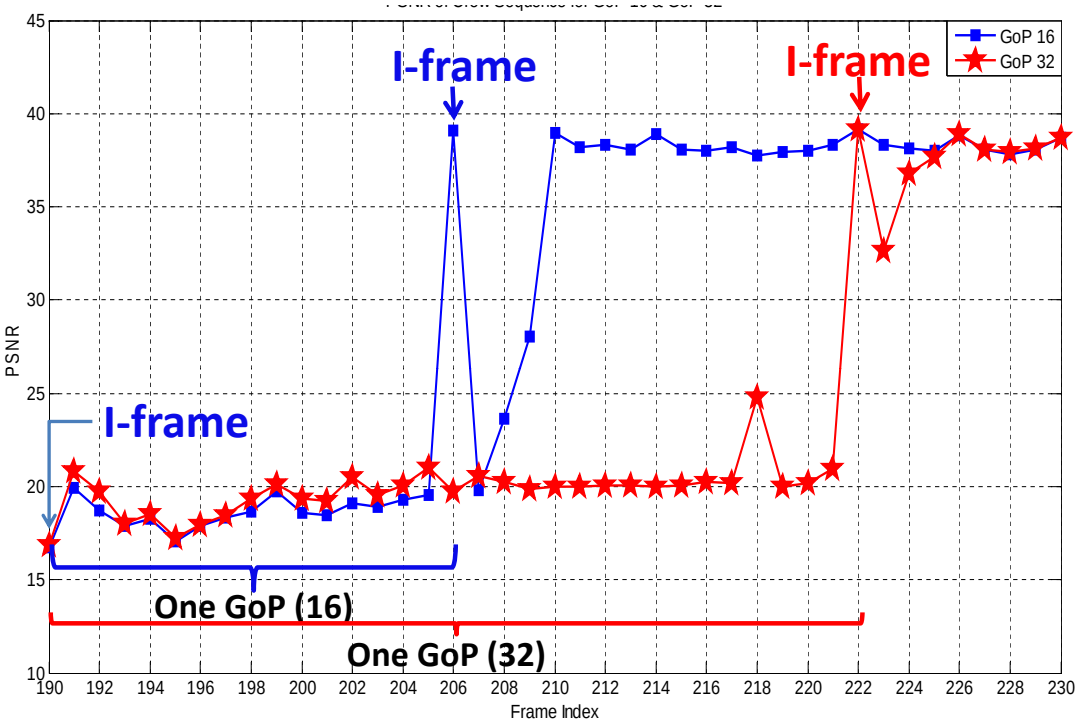


Figure 5.9: PSNR of the crew sequence (GoP $M = 4$ & $N = 16$) and (GoP $M = 4$ & $N = 32$).

To investigate the impact of GoP length on the performance of the three models, Table 5.3 shows the percentage drop in the average PSNR of the two GoP lengths when transmitted using the three simulated models. Again, comparing the results of the two GoP lengths for the priority-based queuing model, it is found that the PSNR percentage drop of GoP32 is larger than that of GoP16. This is explained by recalling that the loss rates of P-frames were found to be larger in GoP 32 compared to GoP16. Another

reason for this is the larger number of dependent frames in GoP32 compared to GoP16. The design of the priority-based queuing model aims to lower the loss in I-frames and P-frames; hence, a lower number of reference frames was lost in the priority-based queuing model compared to the best effort-based model.

Considering the best effort-based model, the loss rate of P-frames has increased in the case of GoP32 but the loss rate of I-frames has decreased by the same percentage as the P-frame. Thus, it is not clear how these differences in the frame loss rates of the I- and P-frames in the two GoPs will be reflected in the average PSNR of the video sequence. As indicated in Table 5.3, in the best effort-based model, the percentage drop in the average PSNR of the two GoPs differ from one video sequence to the other. Lastly, when comparing the results of GoP 16 and GoP 32 of the path diversity-based model the differences were not significant. This is explained by noting that the path diversity-based model has mainly lost B-frames (non reference frames); therefore, the small difference in percentage is due to the different motion content of the video sequences.

Table 5.3: Percentage drop in the PSNR.

Model	Video Sequence	GoP $M = 4$ & $N = 16$	GoP $M = 4$ & $N = 32$
Best Effort	Crew Sequence	13%	18%
	Mobile Sequence	13%	11%
	Foreman Sequence	12%	9%
	Walk Sequence	9%	12%
Priority Queuing	Crew Sequence	3%	5%
	Mobile Sequence	5%	8%
	Foreman Sequence	3%	6%
	Walk Sequence	6%	9%
Path Diversity	Crew Sequence	2%	3%
	Mobile Sequence	4%	4%
	Foreman Sequence	1%	2%
	Walk Sequence	5%	4%

5.2.2 Buffer Occupancy. In what follows, the performance of the three streaming models is compared in terms of the occupancy of the playback buffer at the receiver

side. At the start of the video session, a number of correctly decoded frames (i.e. buffering delay) will be stored in the receiver's playback buffer before playback starts. When playback starts, frames will be displayed at a fixed display rate of $f_p = 30$ frames/sec. While frames are drained from the playback buffer at a constant rate, their arrival rate depends on the network conditions. Therefore, periods with high congestion levels will not only cause the occupancy of the buffer to underflow but will cause it to starve as well.

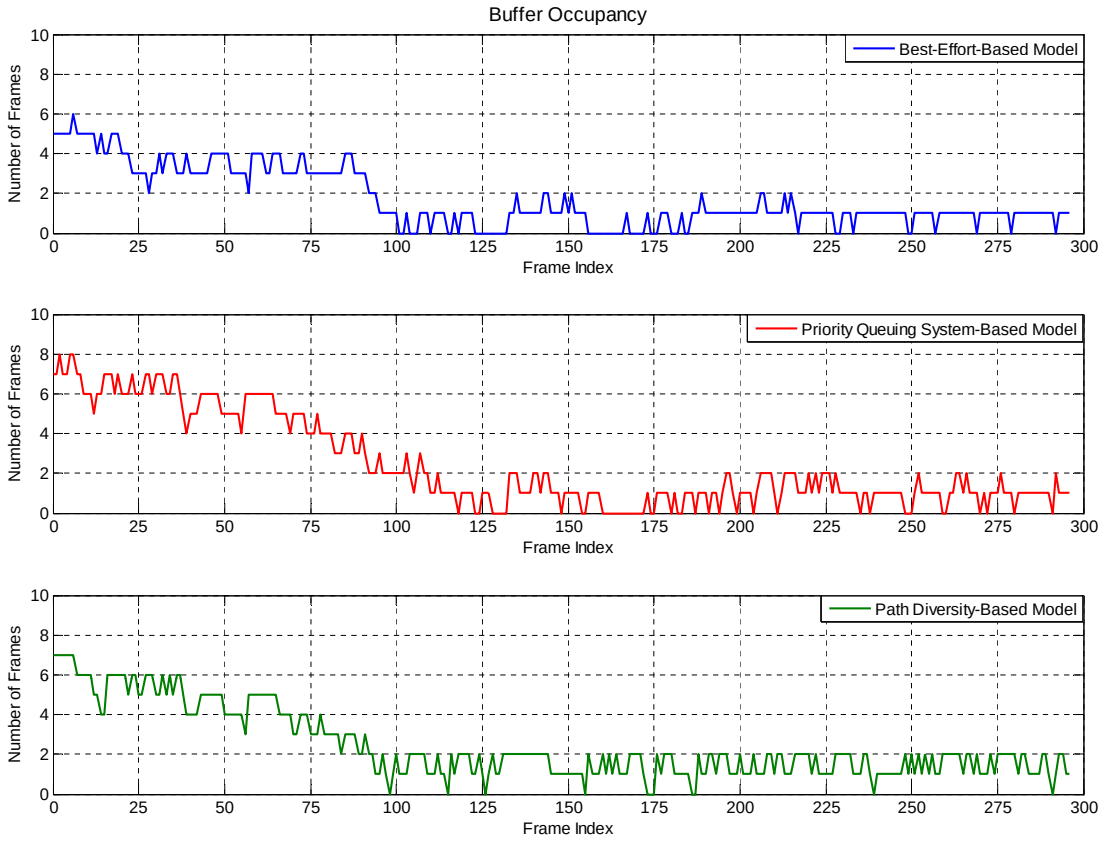


Figure 5.10: Buffer occupancy versus frame Index ($M = 16$ and $N = 4$).

Figures 5.10 and 5.11 plots the occupancy of the playback buffer at the receiver side for the three streaming models and for GoP16 and GoP 32, respectively. The buffer occupancy is found at the highest congestion level considered in this study. These figures clearly highlight the capability of the proposed PBPD streaming model in reducing the starvation instants at the playback buffer. Moreover, the figures also show that on

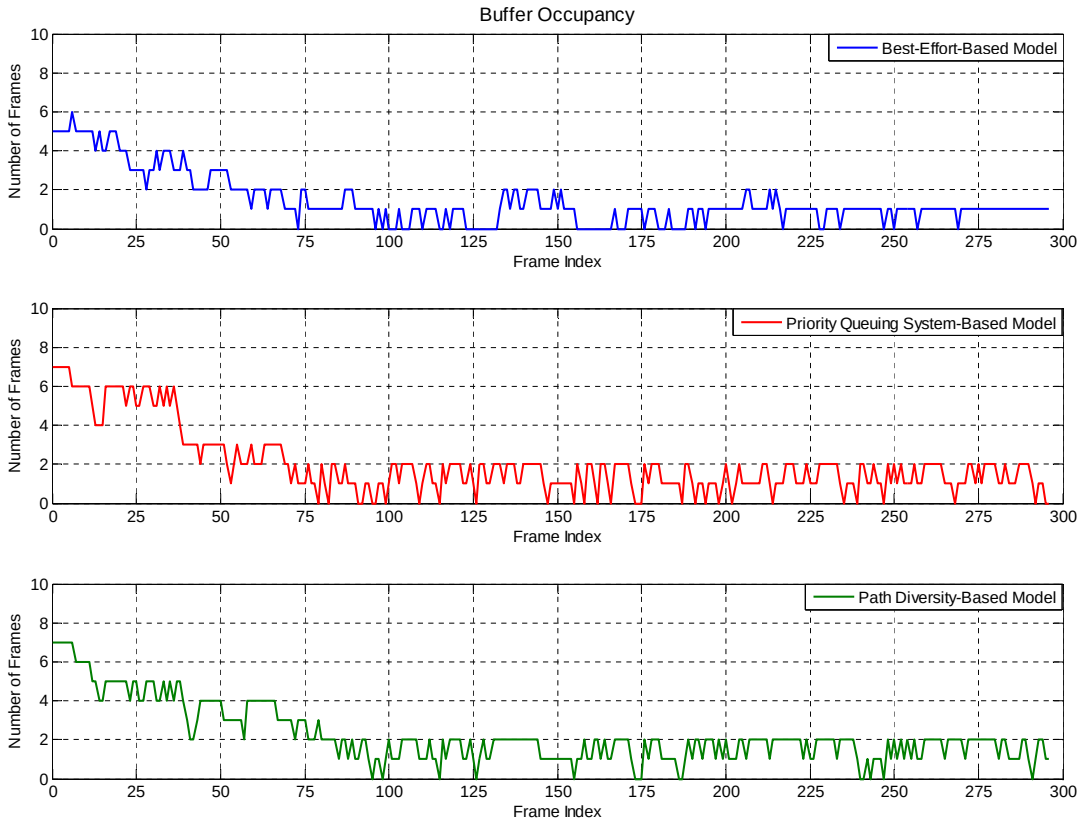


Figure 5.11: Buffer occupancy versus frame index ($M = 32$ and $N = 4$).

the occurrence of a starvation it will not last for consecutive frames. Such a separation between starvation instances is expected not only to improve the perceptual quality of the reconstructed video but also reduces interruptions in the playback process.

The skip length and inter-starvation distance metrics are also used to com pare the performance of the three simulated models [57]. The skip length denoted by $\overline{L}$ which, on the occurrence of any starvation instant, indicates how long (in frames) this starvation will last on average. The average skip length is computed using $\bar{L} = \sum_{k=1}^{L} L_k/L$, where L is the the total number of starvation intervals and L_k is the number of frames a client misses on the occurrence of the kth starvation interval. The inter-starvation distance denoted by $\overline{D}$ which, on the occurrence of any starvation instant, is the distance in frames that separates successive starvation instants. The inter-starvation distance is computed by $\bar{D} = \sum_{k=1}^{D} D_k/D$, where D is the number of uninterrupted

playback intervals, and D_k is the number of frames played between two starvation intervals. Table 5.4 shows the values of skip length $\overline{L}$ and inter-starvation distances $\overline{D}$ found when using the three simulated models.

Table 5.4: Skip length and inter-starvation distances (in frames) using the three models.

(a) GoP Structure: $M = 4$ and $N = 16$

Network Model	Best Effort	Priority Queuing	Path Diversity
Skip Length ($\overline{L}$)	2.8	2	1.4
Inter-Starvation Distance ($\overline{D}$)	12.2	10.5	31.8

(b) GoP Structure: $M = 4$ and $N = 32$

Network Model	Best Effort	Priority Queuing	Path Diversity
Skip Length ($\overline{L}$)	2.4	1.2	1.4
Inter-Starvation Distance ($\overline{D}$)	10	10.2	23.5

The skip length values of the path diversity-based model and the priority-based queuing model are lower than that of the best effort-based model. No much difference is found between in the skip length values of the priority-based queuing model and the best effort-based model of GoP 32. However, a remarkable difference is found in the inter-starvation distance values of the priority-based queuing model and the path diversity-based model. This indicates that when a starvation moment occur in the two models, it lasts for almost the same average number of frames but the separation between starvation moments in the path diversity-based model is much larger compared to the priority-based queuing model. The small values of skip length combined with the large values of inter-starvation instants achieved when using the path diversity-based model result in less playback interruptions and better quality of the reconstructed video. This enhancement in performance of the path diversity-based model is a result of the lower delays and losses experienced by video frames when using the priority queuing system and the PBPD scheme.

Chapter 6: Conclusions

In this thesis, priority and path diversity-based schemes are studied for video streaming over best effort networks such as the Internet. The novelty of the proposed approaches is in the use of path diversity by considering the difference in importance between video frames for the objective of successfully delivering the most important parts of the video.

The performance of the proposed PBPD scheme is evaluated and analyzed through simulations in comparison with a priority-based queuing model and a best effort-based model. The priority-based queuing model is designed such that each router maintains a priority queuing system without the use of a path diversity transmission approach. On the other hand, the best effort-based model is designed to reflect the performance of today's best effort networks, like the Internet. The performance of the three models is studied in two sections. The first section compares the performance of the three models in terms of network related metrics such as the frame loss rate and the end-to-end delay. The second section compares the performance of the three models from the end users' perspective which are the PSNR to measure the spatial video quality and a buffer occupancy metric is used to study the continuity of the reconstructed video.

The results show that video frames experienced the lowest frame loss rates when transmitted using the proposed PBPD scheme. In addition, the PBPD scheme outperforms the priority-based queuing model and the best effort-based model in terms of the end-to-end delay. These results are reflected in the delivered video spatial and temporal qualities in which larger average PSNR values are achieved and lower buffer starvation instants are experienced using the PBPD scheme.

The impact of the proposed schemes on best effort traffic is neglected as non-real time applications can overcome packet loss by employing retransmission schemes. However, having high packet loss rate in best effort traffic will require the retransmission of large number of packets which in turn might increase the congestion in the network. When employing the proposed priority queuing system in the network routers,

best effort traffic will only be served if all higher priority queues are empty. One way to reduce losses of the best effort traffic is to assign different weights to the priority queues such that large weights are assigned to high priority queues allowing them to be served more frequently compared to lower priority queues. At the same time, by assigning weights we ensure that the best effort traffic is being served especially when the incoming rate of video traffic is high.

Another point to study is the response of the queuing system when a priority queue experience buffer overflow. In this study, when an I-frame arrives to a full I-queue, the arriving I-frame will be dropped. However, a queue management scheme can be studied to decide to either drop the arriving I-frame or to drop a queued packet in a lower priority queue.

References

[1] J. Hwang, *Multimedia Networking: From Theory to Practice*. Cambridge University Press, 2009.

[2] J. F. Keith and W.Rossi, *Computer Networking A Top-Down Approach Featuring the Internet*. Pearson Education, 2003.

[3] L. Li, Q. Han, and X. Niu, "Enhanced Adaptive FEC Based Multiple Description Coding for Internet Video Streaming over Wireless Network," in *Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010*, pp. 478–481, Oct 2010.

[4] C. Diaz, J. Cabrera, F. Jaureguizar, and N. Garcia, "A Video-Aware FEC-Based Unequal Loss Protection System for Video Streaming over RTP," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 523–531, 2011.

[5] A. Rafetseder, F. Metzger, D. Stezenbach, and K. Tutschku, "Exploring YouTube's Content Distribution Network Through Distributed Application-layer Measurements: A First View," in *Proceedings of the 2011 International Workshop on Modeling, Analysis, and Control of Complex Networks*, Cnet '11, pp. 31–36, International Teletraffic Congress, 2011.

[6] X. Cheng, J. Liu, and C. Dale, "Understanding the Characteristics of Internet Short Video Sharing: A YouTube-Based Measurement Study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184–1194, 2013.

[7] P. Juluri, L. Plissonneau, Y. Zeng, and D. Medhi, "Viewing YouTube from a metropolitan area: What do users accessing from residential ISPs experience?," in *IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), 2013*, pp. 589–595, May 2013.

[8] T. Szymanski, "Max-Flow Min-Cost Routing in a Future-Internet with Improved QoS Guarantees," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1485–1497, 2013.

[9] T. Szymanski, "Future Internet Video Multicasting with Essentially Perfect Resource Utilization and QoS Guarantees," in *IEEE 19th International Workshop on Quality of Service (IWQoS), 2011*, pp. 1–3, 2011.

[10] R. Adams, "Active Queue Management: A Survey," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, pp. 1425–1476, 2013.

[11] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-Constrained Coder Control and Comparison of Video Coding Standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, 2003.

[12] P. Seeling, M. Reisslein, and B. Kulapala, "Network Performance Evaluation using Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial," *Communications Surveys Tutorials, IEEE*, vol. 6, no. 3, pp. 58–78, 2004.

[13] P. Correia, P. Assuncao, and V. Silva, "Multiple Description of Coded Video for Path Diversity Streaming Adaptation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 923–935, 2012.

[14] N. Maxemchuk, "Dispersity Routing: Past and Present," in *Military Communications Conference, 2007. MILCOM 2007. IEEE*, pp. 1–7, 2007.

[15] N. Maxemchuk, "Dispersity Routing on ATM Networks," in *INFOCOM '93. Proceedings.Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future, IEEE*, pp. 347–357 vol.1, 1993.

[16] J. Apostolopoulos and M. Trott, "Path diversity for Enhanced Media Streaming," *Communications Magazine, IEEE*, vol. 42, no. 8, pp. 80–87, 2004.

[17] S. Somasundaram, K. P. Subbalakshmi, and R. N. Uma, "MDC and Path Diversity in Video Streaming," in *International Conference on Image Processing, 2004. ICIP '04. 2004*, vol. 5, pp. 3153–3156 Vol. 5, 2004.

[18] J. Han, D. Watson, and F. Jahanian, "An Experimental Study of Internet Path Diversity," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 4, pp. 273–288, 2006.

[19] N. Feamster, D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "Measuring the Effects of Internet Path Faults on Reactive Routing," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, pp. 126–137, ACM, 2003.

[20] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, pp. 293–306, 2001.

[21] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The End-to-End Effects of Internet Path Selection," in *ACM SIGCOMM Computer Communication Review*, vol. 29, pp. 289–299, ACM, 1999.

[22] X. Misseri, J. Rougier, and D. Saucez, "Internet Routing Diversity for Stub Networks with a Map-and-Encap Scheme," in *IEEE International Conference on Communications (ICC), 2012*, pp. 2861–2866, 2012.

[23] C. Hu, K. Chen, Y. Chen, B. Liu, and A. Vasilakos, "A Measurement Study on Potential Inter-Domain Routing Diversity," *IEEE Transactions on Network and Service Management*, vol. 9, no. 3, pp. 268–278, 2012.

[24] V. Sharma, K. Kar, K. Ramakrishnan, and S. Kalyanaraman, "A Transport Protocol to Exploit Multipath Diversity in Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 1024–1039, 2012.

[25] H. Han, S. Shakkottai, C. Hollot, R. Srikant, and D. Towsley, "Multi-Path TCP: A Joint Congestion Control and Routing Scheme to Exploit Path Diversity in the Internet," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1260–1271, 2006.

[26] P. Arjona-Villicana, C. Constantinou, and A. Stepanenko, "The Internet's Unexploited Path Diversity," *IEEE Communications Letters*, vol. 14, no. 5, pp. 474–476, 2010.

[27] J. G. Apostolopoulos, "Reliable Video Communication over Lossy Packet Networks using Multiple State Encoding and Path Diversity," in *Visual Communications and Image Processing*, vol. 4310, pp. 392–409, 2001.

[28] N. Gogate, D.-M. Chung, S. Panwar, and Y. Wang, "Supporting Image and Video Applications in a Multihop Radio Environment using Path Diversity and Multiple Description Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 9, pp. 777–792, 2002.

[29] W.-H. Hsu, Y.-P. Shieh, S.-C. Yeh, and P.-S. Hung, "A Simple Path Diversity Algorithm for Interdomain Routing," in *IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA), 2011*, pp. 291–296, 2011.

[30] M. Jain and C. Dovrolis, "Path Selection using Available Bandwidth Estimation in Overlay-Based Video Streaming," *Computer Networks*, vol. 52, no. 12, pp. 2411–2418, 2008.

[31] A. Razzaq and A. Mehaoua, "Layered Video Transmission Using Wireless Path Diversity Based on Grey Relational Analysis," in *IEEE International Conference on Communications (ICC), 2011*, pp. 1–6, 2011.

[32] A. C. Begen, Y. Altunbasak, O. Ergun, and M. H. Ammar, "Multi-Path Selection for Multiple Description Video Streaming over Overlay Networks," *Signal Processing: Image Communication*, vol. 20, no. 1, pp. 39–60, 2005.

[33] P. Frossard, J. De Martin, and M. Civanlar, "Media Streaming With Network Diversity," *Proceedings of the IEEE*, vol. 96, no. 1, pp. 39–53, 2008.

[34] D. Wu, Y. Hou, W. Zhu, Y.-Q. Zhang, and J. Peha, "Streaming Video over the Internet: Approaches and Directions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 282–300, 2001.

[35] R. Kuschnig, I. Kofler, and H. Hellwagner, "Improving Internet Video Streaming Performance by Parallel TCP-Based Request-Response Streams," in *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*, pp. 1–5, Jan 2010.

[36] M. Hassan and T. Landolsi, "Adaptive Rate Control in IPTV Applications over Wireless Channels," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2008*, pp. 1–5, March 2008.

[37] P. Zhu, W. Zeng, and C. Li, "Joint Design of Source Rate Control and QoS-Aware Congestion Control for Video Streaming Over the Internet," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 366–376, 2007.

[38] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. Begen, and D. Oran, "Probe and Adapt: Rate Adaptation for HTTP Video Streaming At Scale," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, 2014.

[39] M. Hassan and T. Landolsi, "A Retransmission-Based Scheme for Video Streaming over Wireless Channels," in *Wirel. Commun. Mob. Comput.*, vol. 10, p. 511521, 2009.

[40] P. Chou and Z. Miao, "Rate-Distortion Optimized Streaming of Packetized Media," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 390–404, 2006.

[41] Y. Li, Z. Li, M. Chiang, and A. Calderbank, "Content-Aware Distortion-Fair Video Streaming in Networks," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, pp. 1–6, Nov 2008.

[42] M. Hassan and M. Krunz, "A Playback-Adaptive Approach for Video Streaming over Wireless Networks," in *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 6, pp. 5 pp.–3691, Dec 2005.

[43] J. Dong and Y. Zheng, "Content-Based Retransmission for 3-D Wavelet Video Streaming Over Lossy Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 9, pp. 1125–1133, 2006.

[44] Y.-M. Hsiao, C.-H. Chen, J.-F. Lee, and Y.-S. Chu, "Designing and Implementing a Scalable Video-Streaming System using an Adaptive Control Scheme," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1314–1322, 2012.

[45] A. Kashyap and B. Bing, "Efficient HD Video Streaming over the Internet," in *Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon)*, pp. 272–275, 2010.

[46] R. Puri, K.-W. Lee, K. Ramchandran, and V. Bharghavan, "An Integrated Source Transcoding and Congestion Control Paradigm for Video Streaming in the Internet," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 18–32, 2001.

[47] P. Correia, P. Assuncao, and V. Silva, "Multiple Description of Coded Video for Path Diversity Streaming Adaptation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 923–935, 2012.

[48] J. Chakareski, S. Han, and B. Girod, "Layered Coding vs. Multiple Descriptions for Video Streaming over Multiple Paths," in *In Proc. of ACM Multimedia*, pp. 422–431, 2003.

[49] J. Kim, "Layered Multiple Description Coding For Robust Video Transmission Over Wireless Ad-Hoc Networks," *World Academy of Science, Engineering and Technology*, 2006.

[50] P. Correia, L. Ferreira, P. Assuncao, L. Cruz, and V. Silva, "Optimal Priority MDC Video Streaming for Networks with Path Diversity," in *International Conference on Telecommunications and Multimedia (TEMU), 2012*, pp. 54–59, 2012.

[51] I. Ali, S. Moiron, M. Fleury, and M. Ghanbari, "Video Streaming over an Ad Hoc Network using Data Partitioning and Path Diversity," *IEEE Wireless Communications*, vol. 20, no. 3, pp. 105–111, 2013.

[52] H. Egilmez, S. Civanlar, and A. Tekalp, "An Optimization Framework for QoS-Enabled Adaptive Video Streaming Over OpenFlow Networks," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 710–715, 2013.

[53] S. Domoxoudis, S. Kouremenos, A. Drigas, and V. Loumos, "Frame-Based Modeling of H.264 Constrained Videoconference Traffic over an IP Commercial Platform," in *2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006. TRIDENTCOM 2006.*, pp. 6 pp.–221, 2006.

[54] K. Salah, F. Al-Haidari, M. H. Omar, and A. Chaudhry, "Statistical Analysis of H.264 Video Frame Size Distribution," *Communications, IET*, vol. 5, no. 14, pp. 1978–1986, 2011.

[55] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[56] K. Shring, "H.264/AVC Software Coordination," 2014.

[57] M. Hassan, T. Landolsi, K. Assaleh, and H. Mukhtar, "An Adaptive Scheduling Algorithm for Video Transmission over Wireless Packet Networks," *Electronic Notes in Discrete Mathematics*, vol. 36, pp. 1319–1326, 2010.

Vita

Dana Hussein received her B.S. in Electrical Engineering from the American University of Sharjah (AUS) in 2012. After that, she joined the graduate program at the department of Electrical Engineering at AUS as a teaching assistant and currently pursuing her M.S. in Electrical Engineering degree. During this time, she participated in the UKSim 2014, with a paper on her research findings.