ADVANCED LINK AND TRANSPORT CONTROL PROTOCOLS FOR BROADBAND OPTICAL ACCESS NETWORKS

A Dissertation
Presented to
The Academic Faculty

by

Chunpeng Xiao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2006

ADVANCED LINK AND TRANSPORT CONTROL PROTOCOLS

FOR BROADBAND OPTICAL ACCESS NETWORKS

Approved by:

Dr. Gee-Kung Chang, Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. G. Tong Zhou
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Mary Ann Ingram
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. John Copeland
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. George Riley
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Mostafa Ammar
College of Computing
Georgia Institute of Technology

Date Approved: October 27, 2006

To my parents, and my wife Leihong Li

ACKNOWLEDGEMENTS

I would like to express my gratitude to all whose direct and indirect support helped me complete my thesis.

First, I would like to thank my advisor, Prof. Gee-Kung Chang who has continuously supported my work and ideas throughout my studies at Georgia Institute of Technology. I appreciate his close guidance and great efforts to prepare me to become a mature researcher. My special thanks also go to Dr. G. Tong Zhou for her advices that prepare me a lot for my research work.

Next, I would like to thank my thesis committee members: Dr. John Copeland, Dr. George Riley, Dr. Mostafa Ammar, and Dr. Mary Ann Ingram for helping to improve my dissertation with their thoughtful advice and suggestions.

I would like to thank my group members: Benny Bing, Jianjun Yu, Daniel Guidotti, Arshad Chowdhury, Yin-Jung Chang, Yong-Kee Yeo, Joseph Long, Oladeji Alkanbi, Muhammad Haris, Muhammad Faisal, Zhensheng Jia, Jerome Vasseur, Claudio Estevez, David Boivin, Ming-Fang Huang, and Lingbin Kong, for their willingness to help, our insightful discussions and fruitful collaborations.

My sincere thanks also go to my parents for their unconditional love, encouragement, and support.

Last but not least, I would like to express my deepest gratitude to my wife Leihong Li for her company during my graduate study. Without her love, patience, understanding, and support, I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AIMD	Additive Increase and Multiplicative Decrease
APON	ATM PON
ATM	Asynchronous Transfer Mode
BDP	Bandwidth Delay Product
CBS	Committed Burst Size
CIR	Committed Information Rate
CMTS	Cable Modem Termination System
DOCSIS	Data Over Cable Service Interface Specification
DSL	Digital Subscriber Line
DS-MPLS-TE	Differentiated Services-aware MPLS Traffic Engineering
EBS	Excess Burst Size
EFM	Ethernet in the First Mile
EIR	Excess Information Rate
E-Line	Ethernet Line
E-LAN	Ethernet LAN
EPL	Ethernet Private Line
EPON	Ethernet PON

EVPL	Ethernet Virtual Private Line
FTTC	Fiber to the Curb
FTTH	Fiber to the Home
HFC	Hybrid Fiber Coax
IETF	Internet Engineering Task Force
IP	Internet Protocol
LAN	Local Area Network
MAC	Media Access Control
MAN	Metropolitan Area Network
MPLS	Multi-Protocol Label Switching
OLT	Optical Line Terminator
ONU	Optical Network Unit
OSI	Open System Interconnection
PON	Passive Optical Network
QoS	Quality of Service
SAN	Storage Area Network
SLA	Service Level Agreement
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing

UDP	User Datagram Protocol
UNI	User Network Interface
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing

SUMMARY

The objective of this dissertation is to improve the service quality of broadband optical access networks by developing advanced link- and transport- layer protocols. Access networks connect business and residential premises to metropolitan area networks or wide area networks. Current access technologies represent a significant bottleneck in bandwidth and service quality between a high-speed residential/enterprise network and a largely overbuilt core backbone network. Although it is believed that passive optical network (PON) will be the most promising solution to provide truly broadband connections to end users, a suit of protocols are required to provide dedicated bandwidth, controlled delay and jitter, and preferable packet loss characteristics, for quadra play applications including data, voice, video, and interactive multimedia games. This dissertation studies three key techniques for improving quality of service (QoS) performance of broadband access networks, including transport protocol for broadband networks, media access control (MAC) protocol for PON, and delta compression for fast content download.

In this dissertation, we design a new reservation MAC scheme that arbitrates upstream transmission, prevents collisions, and varies bandwidth according to demand and priority. The new access scheme exploits both WDM and TDM to cater for both light and heavy bandwidth requirements. We analyzed, evaluated, and simulated the performance and practicality of the proposed scheme.

Next, we introduce delta compression algorithms as an efficient method for fast content download. We present a theoretical framework for delta compression based on information theory and Markov models, including insights into the compression bounds. We also simulated and implemented a generic delta compression scheme and demonstrated its real-time and non real-time performance.

In the third part of this dissertation, we enhance the transport performance of Ethernet services by addressing the throughput optimization issue at the edge of the network. The performance of TCP over SLA driven Ethernet services is not adequate, because the mechanisms of TCP are designed for best effort networks and cannot utilize the reserved bandwidth efficiently. In this research, an SLA-aware transport protocol is proposed to increase the end-to-end throughput for Ethernet services.

To reduce the impact of random loss to application layer throughput, a resilient transport control scheme dealing with random error is proposed for Ethernet services. Transport layer congestion control is stated as a hypothesis test process making decisions through the ACK feedbacks. It is shown that traditional TCP adopts a very simple decision rule making decisions based on only a single sample. A novel resilient congestion control scheme is proposed with a better decision rule using multiple samples. The new method increases throughput significantly by keeping false-alarm probability under control and adjusting congesting window adaptively.

CHARPTER 1

INTRODUCTION

1.1 Motivation

Access networks connect business and residential premises to metropolitan area networks (MANs) or wide area networks (WANs). Because of the explosive growth of broadband multimedia applications, such as video streaming, high-definition television (HDTV), video on demand, and interactive games, a huge demand for bandwidth has been imposed on the access infrastructure. As DWDM technology was developed for the long haul network and Gigabit Ethernet for the local area network (LAN), access networks tend to be the bottle-neck for end-to-end broadband applications. Today, the two most popular access network solutions are digital subscriber line (xDSL) technologies deployed by telephone companies, and cable modems from cable companies. These access technologies do not have comparable bandwidth capability with Gigabit Ethernet and have limitations in providing high-quality integrated services, including video, voice, and data. Unlike metro and long-haul networks, access networks must serve a more diverse and cost-sensitive customer base. End users may range from individual homes to corporate premises, and services must therefore be provisioned accordingly. Data, voice, and video must be offered over the same high-speed connection with guarantees on quality of service (QoS) and the ability to upgrade bandwidth and purchase content on a needed basis [1]. Therefore, the next-generation access solutions have to be cost efficient when providing more bandwidth.

In the so-called FTTx access networks, optical fiber replaces copper in the distribution network. For example, in fiber to the curb (FTTC) or home (FTTH), the capacity of access networks is sufficiently increased to provide broadband services to subscribers. Because of the cost sensitivity of access networks, passive optical networks (PONs) are considered to be the most promising technology, as they can provide reliable yet integrated data, voice, and video services to end users at bandwidths far exceeding current access technologies. Unlike other access networks, PONs are point-to-multipoint networks capable of transmitting over 20 kilometers of single-mode fiber. PONs can offer symmetrical data transmission on both the upstream and downstream links, allowing the end user to provide Internet services such as music file sharing and Web hosting. In addition to providing a good alternative, PONs represent an excellent evolutionary path for current access technologies such as cable and DSL. By using passive components (such as optical splitters and couplers), PONs reduce the installation and maintenance costs of fiber as well as connector termination space. These costs still require laying fiber, which makes PONs more expensive to install. However, since fiber is loss limited rather than bandwidth limited (as opposed to copper wires, cable, and wireless), the potential performance gains and long-term prospects make PONs well-suited for new neighborhoods or installations.

Enabling technologies for broadband optical access involve the lower four layers of the Open System Interconnection (OSI) reference model, i.e., physical layer, data link layer, network layer, and transport layer. In the physical layer, components for optical access, such as light source, photo detector, splitter/coupler, inter-leaver, and amplifier, are already commercially available. In the data link layer, media access control (MAC) protocol plays an important role for providing dedicated bandwidth, controlled delay and jitter, and preferable packet loss characteristics. Although frameworks have been proposed [2] [3] for optical access, detailed QoS provisioning schemes are still under investigations. In the network layer, protocols to provide QoS with IP, including DifferServ [4], IntServ [5], and DS-MPLS-TE [6], have been standardized by the Internet Engineering Task Force (IETF) and implemented in real networks. In the transport layer, despite the widespread deployment of TCP protocol, current TCP congestion control schemes based on additive increase and multiplicative decrease (AIMD) and slow-start are challenged to improve the efficiency of the link bandwidth utilization as well as fairness among multiple users for next-generation broadband networks with large bandwidth delay product (BDP).

1.2 Objectives

The objective of this dissertation is to improve the service quality of broadband optical access networks by developing advanced link- and transport- layer protocols. Research efforts focus on three areas:

- Media access control (MAC) protocol for PON
- Delta compression for fast content download
- Transport protocol for Ethernet services

Current access technologies represent a significant bottleneck in bandwidth and service quality between a high-speed residential/enterprise network and a largely overbuilt core backbone network. Although it is believed that passive optical network (PON) will be the most promising solution to provide truly broadband connections to end users, a suit of protocols are required to provide dedicated bandwidth, controlled delay and jitter, and preferable packet loss characteristics, for integrated applications including data, voice and video. This research studies several key techniques for improving quality of service (QoS) performance of broadband access networks.

1.3 Outline

The dissertation is organized as follows:

Chapter 2 reviews the literature in the field of broadband access. The existed and emerging solutions of broadband optical access networks are presented. Delta compression is introduced to reduce Internet traffic volume. Several problems with the current TCP in broadband access networks are discussed.

In Chapter 3 [7], we propose a new reservation MAC scheme that not only arbitrates upstream transmission and prevents collisions, but also varies bandwidth according to demand and priority, reduces request delay using pre-allocation, and handles the addition/reconfiguration of network nodes efficiently. The new access scheme exploits both WDM and TDM to cater for both light and heavy bandwidth requirements and supports both Ethernet and ATM packets without segmenting or aggregating them. We analyzed, evaluated, and simulated the performance and practicality of the proposed scheme.

Delta compression algorithms take advantage of the statistical correlations between different files or packets so that common sequences between two files can be encoded using a copy reference. Delta compression forms the core of fast and efficient content download. In this dissertation, we present for the first time, a theoretical framework for delta compression based on information theory and Markov models, including insights into the compression bounds. We also simulated and implemented a generic delta compression scheme and demonstrated its real-time and non real-time performance [8, 9].

In Chapter 5 [10, 11], we aim to enhance the transport performance of Ethernet services by addressing the throughput optimization issue at the edge of the network. Ethernet services convert Ethernet from a best effort technology to a service level agreement (SLA) driven carrier-grade technology, extending the simplicity and flexibility of Ethernet beyond the LAN to the MAN/WAN. The performance of TCP over SLA driven Ethernet services is not adequate, because the mechanisms of TCP are designed for best effort networks and cannot utilize the reserved bandwidth efficiently. In this dissertation, a novel SLA-aware transport protocol is proposed to increase the end-to-end throughput for Ethernet services. The performance of the proposed protocol is compared with traditional TCP through theoretical analysis and simulation.

In Chapter 6 [12], a resilient transport control scheme is proposed for Ethernet services to reduce the impact of random loss to application layer throughput. Transport layer congestion control is stated as a hypothesis test process making decisions through the ACK feedbacks. It is shown that traditional TCP adopts a very simple decision rule making decisions based on only a single sample. A novel resilient congestion control scheme is proposed with a better decision rule using multiple samples. The new method can keep false-alarm probability under control and adjust congesting window adaptively, therefore it increases throughput significantly.

Finally, Chapter 7 concludes the dissertation and provides future research directions.

CHARPTER 2

BACKGROUND

2.1 Broadband Access Solutions

Current access technologies represent a significant bottleneck in bandwidth and service quality between a high-speed residential/enterprise network and a largely overbuilt core backbone network. Backbone networks are provisioned for operation under the worst-case scenarios of link failures, and thus backbone links are lightly loaded most of the time. In addition, high-capacity routers and ultra-high-capacity fiber links have created a true broadband architecture. However, large backbones are not the whole equation; the distribution of that connectivity to individual enterprises and homes is just as critical for meeting the huge demand for more bandwidth (Figure 1). Unfortunately, the cost of deploying true broadband access networks with current technologies remains prohibitive. This in turn makes it difficult to support end-to-end quality of service (QoS) for a wide variety of applications, particularly non-elastic applications such as voice, video, and multimedia that cannot tolerate variable or excessive delay or data loss.

2.1.1 Existing Access Solutions

When it comes to access networks, network operators have a difficult choice among competing technologies – digital subscriber line (DSL), cable, optical, and fixed wireless. Key considerations to the choice include deployment cost and time, service range, and performance. The most widely deployed solutions today are DSL and cable modem networks, which had a combined total of roughly 25 million users by the end of

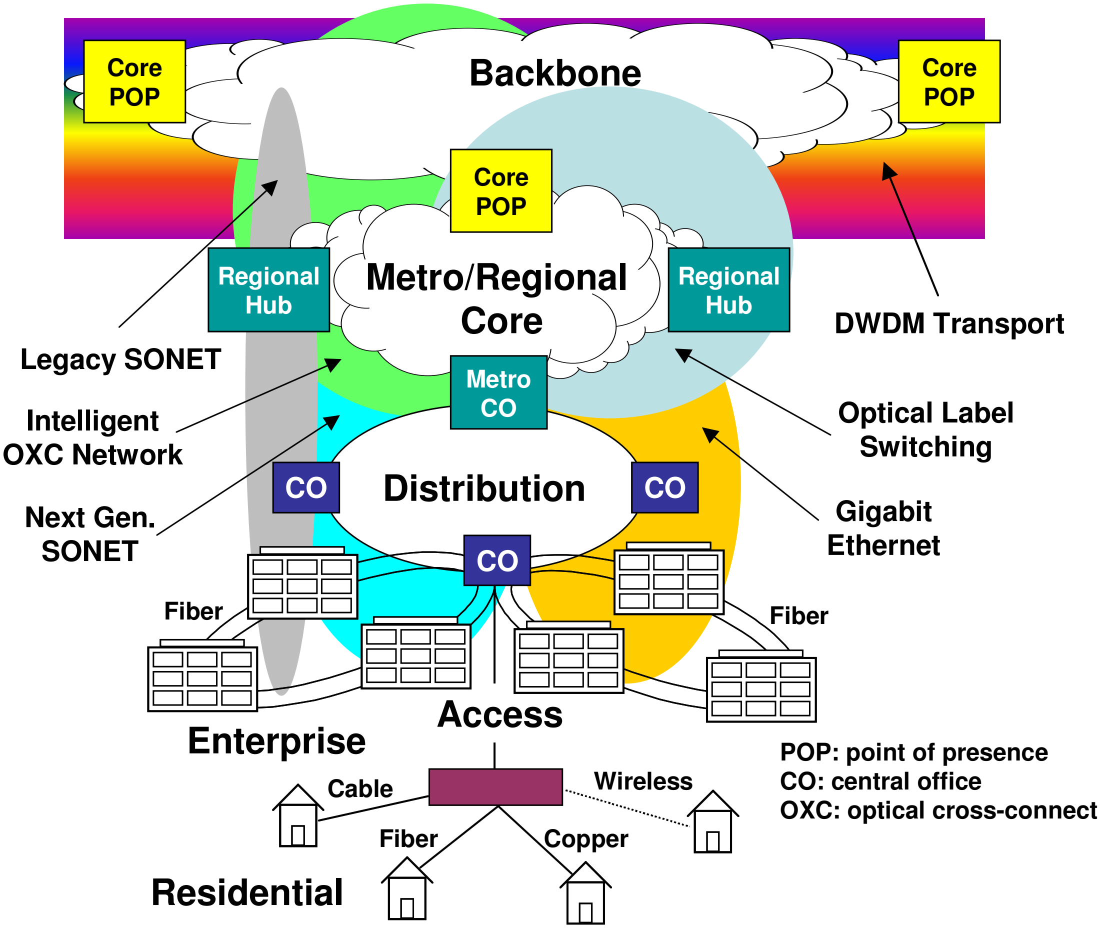


Figure 1: Distributing optical backbone connectivity to enterprises and homes.

2003. Although they offer better performance over 56 Kbit/s dial-up telephone lines, they are not true broadband solutions for several reasons. For instance, they may not be able to provide enough bandwidth for emerging services such as content-rich services, media storage, peer-to-peer services, multiplayer games with audio/video chat to teammates, streaming content, on-line collaboration, high-definition video on demand, and interactive TV services. In addition, fast Web-page download still poses a significant challenge, particularly with rich, engaging, and value-added information involving high-resolution DVD video streaming, multimedia animation, or photo quality images. Finally, only a handful of users can access multimedia files at the same time, which is in stark contrast to direct broadcast TV services. To encourage broad use, a true broadband solution must be scalable to thousands of users and must have the ability to create an ultra-fast Web-page download effect, superior to turning the pages of a book or flipping program channels on a TV, regardless of the content.

A major weakness of both DSL and cable modem technologies is that they are built on top of existing access infrastructures, not optimized for data traffic [13]. In cable modem networks, RF channels that are left over after accommodating legacy analog TV services are dedicated for data. DSL networks do not allow sufficient data rates at required distances because of signal distortion and crosstalk. Most network operators have come to realize that a new, data-centric solution is necessary, most likely optimized over the Internet Protocol (IP) platform. The new solution should be inexpensive, simple, scalable, and capable of delivering integrated voice, video, and data services to the end user over a single network.

2.1.2 Broadband Optical Access

DSL or cable modem access provides the benefits of installed infrastructure, virtually eliminating deployment costs. If fixed wireless access is chosen, network providers gain the benefit of quick and flexible deployment. However, these access methods may suffer bottlenecks in bandwidth-on-demand performance and service range. For example, cable networks are susceptible to ingress noise, DSL systems can be plagued by significant crosstalk, and unprotected broadcast wireless links are prone to security breaches and interference. Furthermore, current DSL and cable deployments tend to have a much higher transmission rate on the downstream link, which restricts Internet applications to mostly Web browsing and file downloads.

While wireless access is excellent for bandwidth scalability in terms of the number of users, optical access is excellent for bandwidth provisioning per user. Furthermore, the longer reach offered by optical access potentially leads to more subscribers. Optical access networks offer symmetrical data transmission on both the upstream and downstream links, allowing the end user to provide Internet services such as music/video file sharing and Web hosting. In addition to providing a good alternative, such networks represent an excellent evolutionary path for current access technologies. These costs still require laying fiber, which makes optical access networks more expensive to install. However, since fiber is loss limited rather than bandwidth limited (as opposed to copper wires, cable, and wireless), the potential performance gains and long-term prospects make optical access networks well-suited for new neighborhoods or installations. In addition, there are innovative solutions for deploying fiber in the last mile, even in established neighborhoods [14].

The passive optical network (PON) is a technology viewed by many as an attractive solution to the last-mile problem as PONs can provide reliable yet integrated data, voice, and video services to end users at bandwidths far exceeding current access technologies. Unlike other access networks, PONs are point-to-multipoint networks capable of transmitting over 20 kilometers of single-mode fiber. As shown in Figure 2, a PON minimizes the number of optical transceivers, central office terminations, and fiber deployment compared to point-to-point and curb-switched fiber solutions. By using passive components (such as optical splitters and couplers) and eliminating regenerators and active equipment normally used in fiber networks (e.g., curb switches, optical amplifiers), PONs reduce the installation and maintenance costs of fiber as well as connector termination space. The general PON architecture consists of the optical line terminator (OLT) on the service provider side and optical network unit (ONU) (or sometimes the optical network terminal) on the user side (Figure 3). The ONUs are connected to the OLT through one shared fiber and can take different FTTx configurations, e.g., fiber to the home (FTTH), fiber to the curb (FTTC), and more recently fiber to the premise (FTTP). The upstream and downstream optical bands specified by ITU-T for dual- and single-fiber PONs are shown in Figure 4.

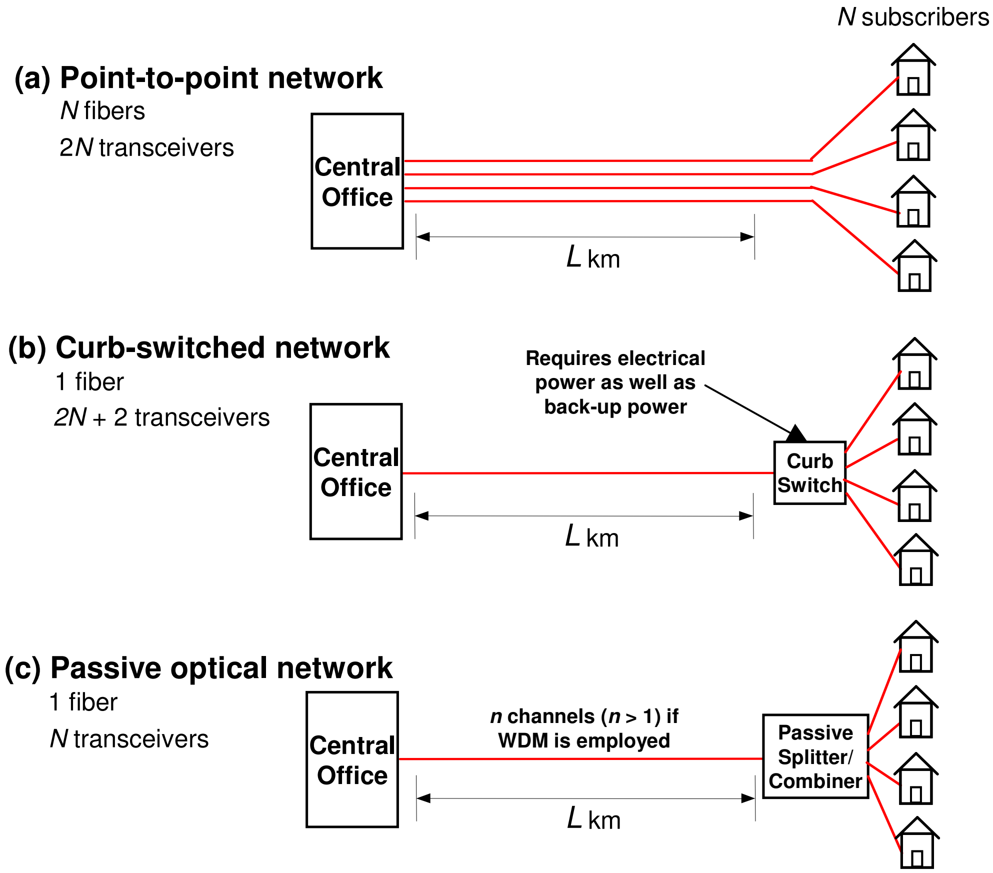
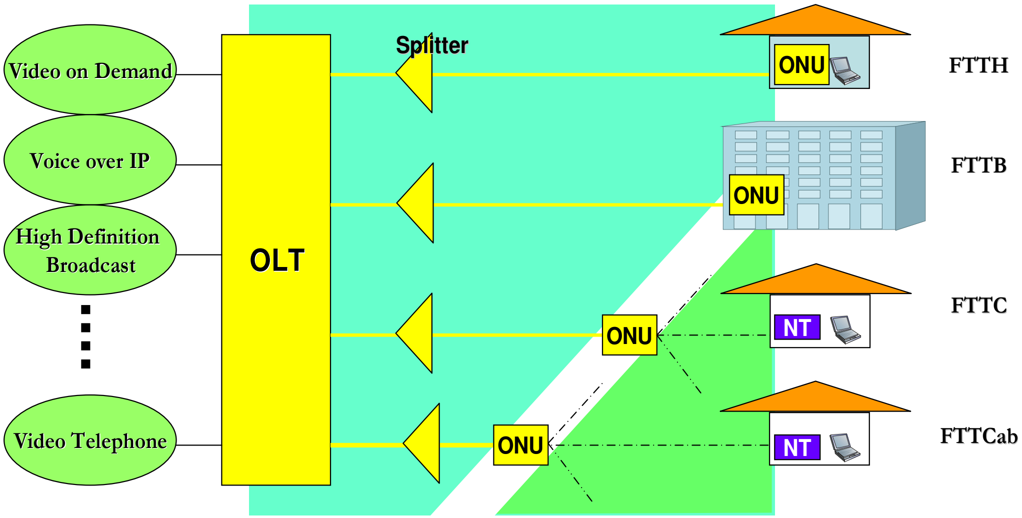

N subscribers
(a) Point-to-point network
N fibers
2N transceivers
Central Office
L km
(b) Curb-switched network
1 fiber
2N + 2 transceivers
Central Office
Requires electrical power as well as back-up power
Curb Switch
L km
(c) Passive optical network
1 fiber
N transceivers
Central Office
n channels (n > 1) if WDM is employed
Passive Splitter/ Combiner
L km

Figure 2: Fiber to the home (FTTH) deployment scenarios.



FTTH: Fiber to the Home FTTB: Fiber to the Building

FTTC: Fiber to the Curb FTTCab: Fiber to the Cabnet

Figure 3: Typical passive optical access network.

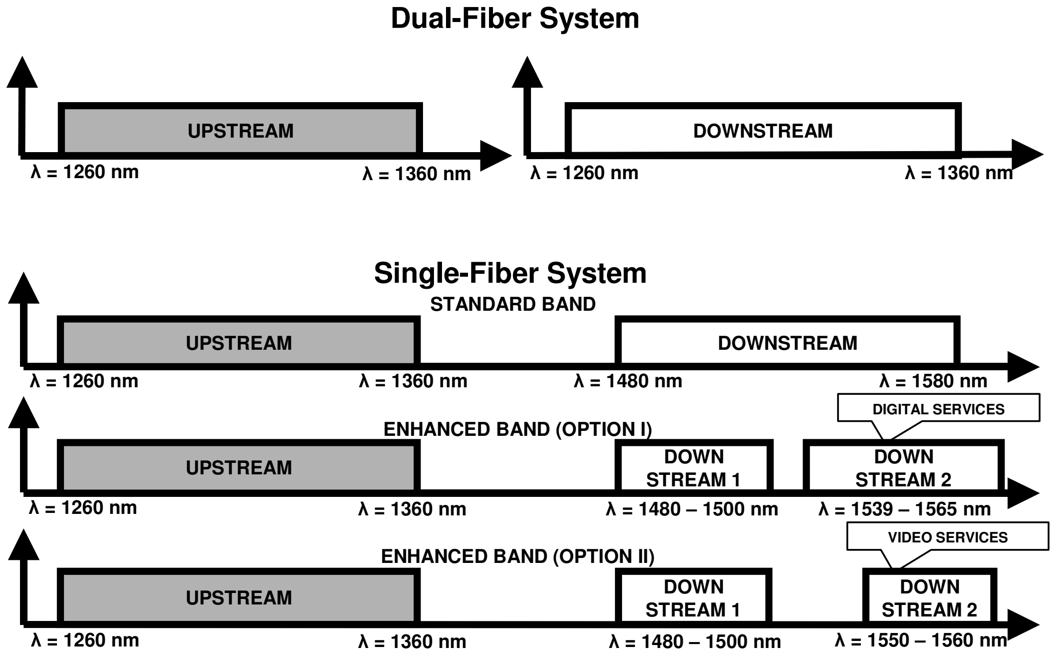


Figure 4: Upstream and downstream optical bands for dual and single-fiber PONs.

PONs typically fall under two groups: ATM PONs (APONs) and Ethernet PONs (EPONs). APON is supported by FSAN and ITU-T because of its connection-oriented QoS feature and extensive legacy deployment in backbone networks [2] [3]. EPON is standardized by the IEEE 802.3ah Ethernet in the First Mile (EFM) Task Force. EPONs leverage low cost, high-performance, silicon-based optical Ethernet transceivers. With the growing trend of GigE and 10GigE in the metro and local area networks, EPONs ensure that IP/Ethernet packets start and terminate as IP/Ethernet packets without expensive and time-consuming protocol conversion, or tedious connection setup.

Wavelength division multiplexing (WDM) is a high-capacity and efficient optical signal transmission technology that is prevalent in long-haul backbone applications, but is now emerging in metropolitan area networks (MAN). WDM uses multiple wavelengths of light, each wavelength corresponding to a distinct optical channel (also known as lightpath or lamda, λ), to transmit information over a single optic fiber simultaneously. Current backbone commercial WDM systems have been increased up to 40 (100GHz spacing),80 (50GHz spacing) in C-band or 160 wavelengths in C+L-band on a single fiber. It is an economical alternative to installing more fibers and a means to dramatically provide higher capacity.

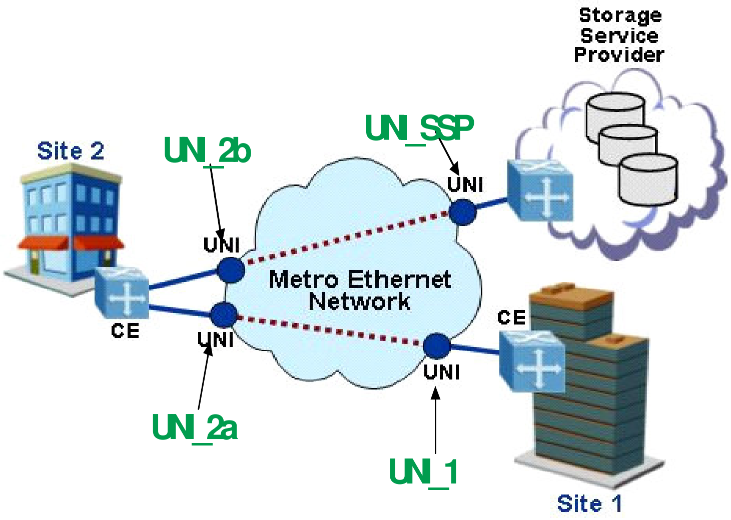
2.1.3 Ethernet for the First Mile and Ethernet Service

Ethernet for the First Mile (EFM) is an effort to extend the Ethernet's reach over the first-mile access link between end users and carriers, and to make Ethernet a low-cost broadband alternative to technologies such as DSL and cable. The motivation for doing this is sound since there are currently more than 500 million Ethernet ports deployed globally and it is advantageous to preserve the native Ethernet frame format rather than terminate it and remap its payload into another layer 2 protocol (e.g., point-to-point protocol, PPP). The EFM specifications are developed by the IEEE 802.3ah Task Force (http://www.ieee802.org/3/efm), which was formed in November 2000. The draft standard (version 3.0 was issued on January 2004) includes *physical layer* specifications for copper, fiber point-to-point, fiber point-to-multipoint topologies, shown in Table 1. The EFM draft standard also defines operations, administration, and maintenance (OAM) aspects of the technology, which local carriers and network operators will use to monitor, manage, and troubleshoot access networks. The same management protocols and architecture work across all EFM topologies.

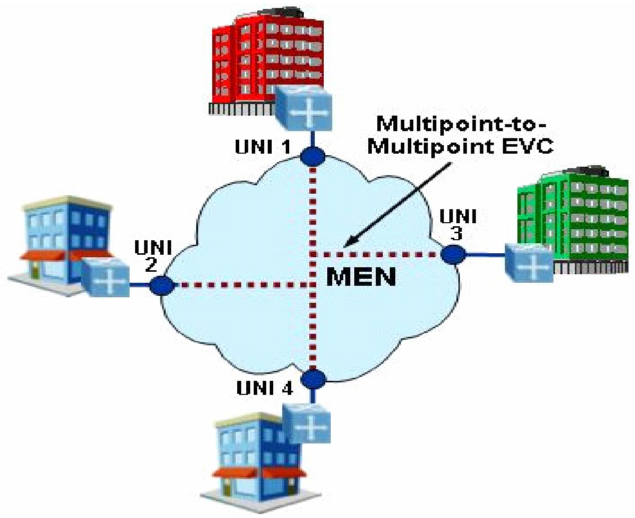
Table 1: EFM topologies.

Topology	**Feature**
EFM Copper	Existing copper wire (Cat 3) at more than 10 Mbit/s for a range of up to at least 750 m
EFM Fiber	Single-mode fiber at 0.1 to 1 Gbit/s for a range of at least 10 km
EFM PON	Optical fiber at 1 Gbit/s for a range of up to 20 km
EFM Hybrid	Topologies that combine the three topologies listed above

The Metro Ethernet Forum (MEF) is an effort to utilize Ethernet as a core protocol for the connectivity from user site to public Internet and connectivity between geographically separate corporate sites (LAN extension). As shown in Figure 5, Ethernet services are classified into either E-Line (point-to-point), or E-LAN (multipoint-to-multipoint). While Ethernet was originally developed as a data-centric protocol, it has evolved to support a full range of services, including real-time services with stringent bandwidth, delay, and loss requirements. The features of a number of existing and future IEEE standards have been employed (e.g., prioritization, virtual LAN or VLAN tagging, traffic shaping, bandwidth management, and resource reservation). Quality of service (QoS) is defined in service level agreement (SLA) and achieved by using a combination of techniques.



E-Line (point-to-point)



(b) E-LAN (multipoint-to-multipoint)

Figure 5: Two types of Ethernet virtual connections.

2.3 Transport Protocol for Broadband Networks

Despite the widespread deployment of the protocol, standard Transmission Control Protocol (TCP) is demonstrated to have a poor performance in the broadband networks [15]. Current TCP implementations rely on packet loss as the only indicator of network congestion, based on the fact that a congested router is the most likely reason for a packet loss in the wired portion of the network. In broadband optical networks with large bandwidth-delay product (BDP), TCP congestion window size will be very large. A very small chance of packet loss will be enough to prevent AIMD-based standard TCP from utilizing available bandwidth efficiently. Therefore, TCP needs to progress from its original low-speed network oriented design and advance to meet the challenges introduced by the broadband using more aggressive congestion control algorithms.

TCP is a connection-oriented transport protocol to provide congestion control and flow control with a sliding window. Four congestion control algorithms are defined in standard TCP [16]: slow start, congestion avoidance, fast retransmit and fast recovery, and congestion is indicated by packet loss. The performance of two implementations of standard TCP (Tahoe and Reno) as well as two modifications of TCP Reno (New-Reno and SACK) are compared in [17]. TCP Tahoe refers to TCP that operates with the slow start, congestion avoidance, and fast retransmission algorithms. TCP Reno refers to TCP that operates with the earlier algorithms plus fast recovery. TCP New-Reno eliminates Reno's wait for a retransmit timer when multiple packets are lost within a window period. In SACK, the receiver reports a non-contiguous set of data that has been received and queued to the sender. Therefore, the case of multiple packet loss can be handled more efficiently. A window-based congestion control mechanism is used in standard TCP. The slow start and congestion avoidance algorithms must be used by a sender. A congestion window is maintained by the sender to control the amount of delivered data into network before getting an acknowledgement from the receiver, where the ideal value is the bandwidth-delay product (BDP is the product of the available bandwidth and the round-trip delay). Another state variable, the slow start threshold (ssthresh), is used to determine whether the slow start or congestion avoidance algorithm is used to control data transmission. In the congestion avoidance phase, TCP probes the available bandwidth by increasing the congestion window slowly, approximately one segment per round-trip delay (RTT), and reduces the congestion window dramatically in the case of congestion, which is indicated by packet loss.

Although the so-called additive increase multiplicative decrease (AIMD) used by TCP successfully solved the stability problem in the 1980s, some disadvantages are exhibited with the increase of network link capacity and diversity of access technologies [18]. TCP has poor link utilization when operated on broadband networks with high bandwidth-delay product because of the conservative AIMD algorithm; long RTT flows obtain less throughput than short RTT flow because the congestion window is increased in the time unit of the RTT.

Besides the simple solution of using multiple TCP connections in parallel, two types of efforts are making for better broadband transport protocols: modifying traditional TCP protocols and redesigning new congestion control schemes based on UDP. Much research has been carried to incrementally modify standard TCP, such as HighSpeed TCP [15], Scalable TCP [19], Explicit Control Protocol (XCP) [20], etc. HighSpeed TCP acts the same as traditional TCP in the case of small BDP, but uses a more aggressive response function for the congestion window in the case of high bandwidth and long delay. In Scalable TCP, a more aggressive MIMD congestion control algorithm is used instead of AIMD. Based on the observation that packet loss is a poor indication of congestion (since this occurs after congestion has occurred), XCP extends the idea of Explicit Congestion Notification and adds a congestion header. Because the per-flow congestion state is carried in packets, intermediate routers can help senders to choose better congestion windows through the feedback header field. However, the requirement of modifying all intermediate routers in the path may not be realistic. Other researchers developed rate-based congestion control algorithms on top of UDP, including SABUL [21], Tsunami [22], Hurricane [23], RBUDP [24], RUNAT [25], etc. These solutions are implemented at the application layer above UDP. Much better bandwidth utilization can be achieved by both types of solutions, whereas the latter one avoids the modifications to operation system kernels, routers, and other network infrastructures.

2.4 Delta Compression for Internet Traffic

The idea of delta compression has been applied pervasively in different forms but optimized for specific applications: MPEG coding of video sequences, software patch update, Web page refresh/caching, storage backups for multiple versions of data, TCP/IP header suppression, and even genetic sequence decoding. Delta compression computes the difference (or delta) corresponding to the new information between two files, packets, or bit strings. The difference is used by the receiving party to construct the new file or packet. This difference is often a few orders smaller than the original and smaller than a direct compression of the original. Delta compression also offers many other advantages over the individual compression of each file version. For example, it facilitates real-time transmission (similar to MPEG encoded video) and is inherently secure: only the client with the original file can successfully generate the new version. Moreover, the hash function used to generate real-time delta files can be integrated with cryptographic hash functions.

Internet traffic can be regarded as a sequence of content units (packets or files). A strong correlation exists among adjacent files. For example, the web pages from the same web site may share related content and similar frame layout. Conventional compression algorithms, such as LZ [26], are designed to remove the redundancy inside a single file, whereas delta compression algorithms can be more efficient to reduce the correlation among adjacent files. Many delta compression algorithms proposed in the past (e.g., Xdelta [27], Vdelta [28], Zdelta [29], VCDIFF [30]) are copy-based algorithms that replace the prefix of the encoded string by a reference to an identical substring in the previous file. Some efforts have been made to include delta compression in the http 1.1 protocol [31]. Compression ratio is the most important metric for a compression algorithm. Although it is proven that the greedy differencing algorithm finds the most compact delta with fixed-length encoding in [32], there is no theoretical analysis for the best compression ratio that can be achieved by delta compression algorithms. This research aims to validate the effectiveness of delta compression using theoretical analysis rather than heuristics.

CHARPTER 3

A RESERVATION MAC PROTOCOL FOR WDM PON

3.1 Media Access Control for Conventional PONs

For conventional PONs, two kinds of upstream/downstream channel separation methods are available. In space-division multiplexing, two separate optical fibers are used, one for the upstream and the other for the downstream. However, the most popular solution is to use the 1550 nm wavelength for downstream and 1310 nm wavelength for upstream. In this case, only one optical fiber is used.

Because of possible contention in the upstream, the media access control (MAC) protocol is needed to resolve these conflicts and in addition, facilitate statistical multiplexing and provision multiple services for different types of traffic. CSMA/CD, WDM, and TDM are the three main types of MAC protocols, as shown in Table 2. In the CSMA/CD protocol for EPONs, a portion of the upstream signal power from any ONU to the OLT is redirected to the downstream and broadcast to all ONUs [33]. An additional receiver at the wavelength of the upstream is needed at each ONU for collision detection. A distributed random backoff algorithm is activated by each ONU to resolve collisions. In WDM PONs, different ONUs use different wavelengths. Therefore, no collision occurs in the upstream. In the TDM scheme, the upstream is divided into multiple time slots and only one ONU can transmit in any time slot. Two types of TDM slot allocation methods are static-TDMA, where a fixed slot is allocated to an ONU in each fixed-length TDM frame cycle, and dynamic-TDMA, where the OLT polls the status of each ONU and allocates a variable-length time slot to ONUs based on the request and grant mechanism [34] [35].

Table 2: MAC Protocols of PONs.

Name	Feature	Advantage	Disadvantage
CSMA/CD	Collision based, upstream signal loop back to all ONUs at splitter, distributed collision backup algorithm	Simpler OLT function	Additional receiver for upstream wavelength is needed at each ONU
WDM	Collision free	Simple, high bandwidth	Expensive
Static-TDMA	Collision free	Inexpensive, simple, small access delay	Low bandwidth utilization
Dynamic-TDMA	Collision free	Inexpensive, high bandwidth utilization	Complex dynamic bandwidth allocation, larger access delay

Currently, dynamic TDMA is the most promising solution, and a polling protocol based on MPCP (multi-point control protocol) has been designed by the IEEE 802.3ah Task Force group. MPCP carries bandwidth polling, bandwidth allocation, auto-discovery, and ranging functions. As shown in Figure 6, dynamic bandwidth allocation is achieved through two control messages, REPORT and GATE. An ONU sends the REPORT message with its buffer status to OLT for requesting bandwidth; the OLT sends a GATE message to an ONU with the beginning time and time slot size for transmission. A REPORT message could be sent together with data packets at either the beginning or the end of the allocated time interval. If an ONU has an empty buffer, a small time slot will still be allocated to guarantee sending the REPORT message in the next TDMA cycle. The ONU buffer consists of multiple queues with different priorities to provide differentiated services. Eight queues can be supported in the REPORT message and four time slots can be granted in one GATE message [36] [37].

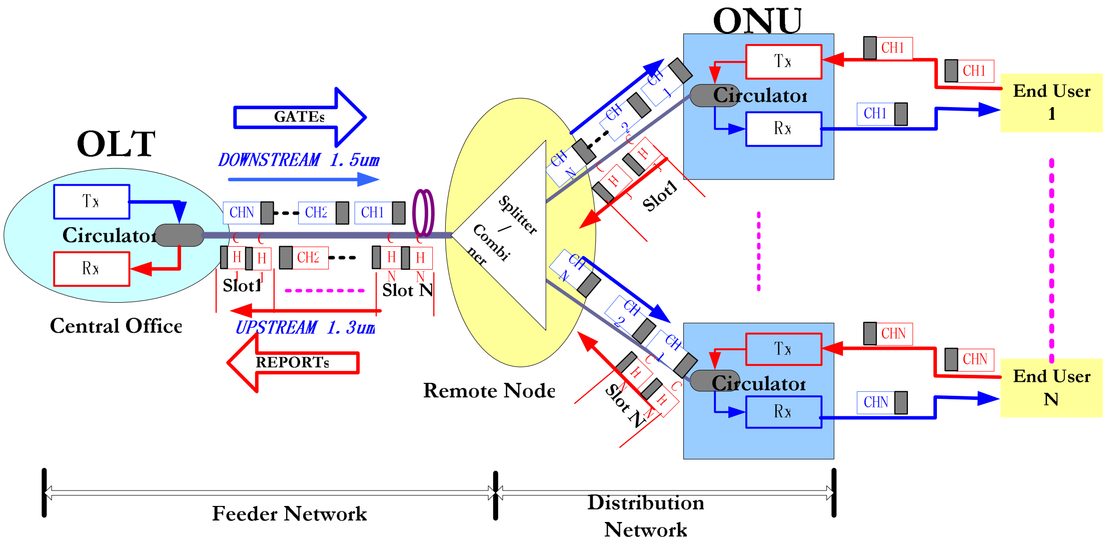
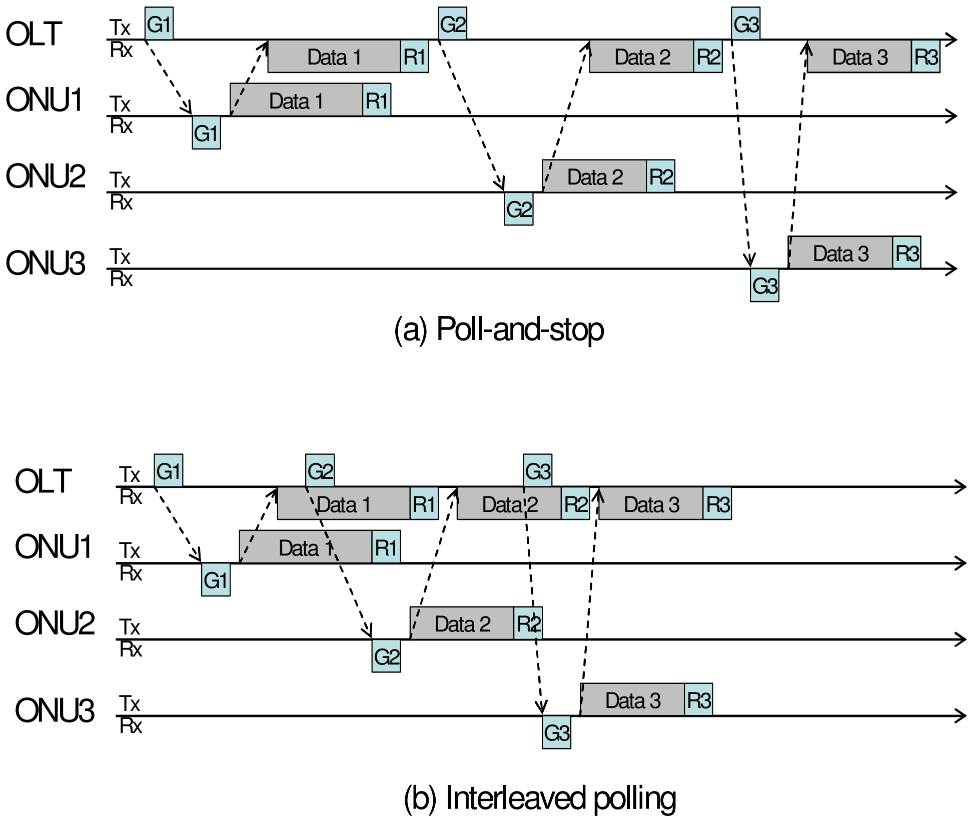


Figure 6: The system architecture of EPON.

As the centralized controller, the OLT polls ONUs one by one to arbitrate upstream bandwidth. The conventional polling method, so-called poll-and-stop polling, is OLT stop sending GATE message to the second ONU until getting the data and REPORT message from the first ONU. Because of the round-trip time (RTT) delay of the GATE message, there is a wasted gap between two ONU's data transmission, as shown in Figure7a. An efficient polling method, interleaved polling, is proposed in [38], that allows the OLT to poll the next ONU before obtaining all data packets from the first ONU. The OLT keeps the RTT information for all ONUs and sends a GATE message in advance, as shown in Figure 7b, so that the time gap between two ONUs' transmissions can be minimized. A significant improvement of bandwidth utilization and access delay can be achieved by interleaved polling.



(a) Poll-and-stop

(b) Interleaved polling

Figure 7: Polling policies of EPON.

3.2 A Novel Reservation MAC Protocol for WDM PONs

The benefits of PONs can be combined with WDM, giving rise to WDM PONs that provides increased bandwidth and allows scalability in bandwidth assignment. Key metrics in the physical layer performance of WDM PONs include latency, link budget, transmitter power and passband, receiver sensitivity, number of serviceable wavelengths, and distance reachable. Several APON and EPON MAC protocols have been suggested in the literature [5 – 8]. The APON protocols are mostly TDMA-based with variations involving dynamic bandwidth allocation and others. EPON MAC protocols are in general polling-based. Unfortunately, none of the proposals integrate both ATM and Ethernet on a WDM PON. This research focuses on the QoS improvement of access network. A MAC protocol for WDM PONs based on dynamic TDMA and pre-allocation is proposed and evaluated through OPNET simulation.

A new contention-free MAC protocol that has flexible controls for both legacy APONs and emerging EPONs is proposed in this research. In addition, the protocol can support WDM, QoS, and dynamic bandwidth allocation. The proposed protocol employs synchronous TDM with the network synchronized at the byte level, allowing the data slot size to be efficiently varied from one byte to a full-length data packet. The various parameters are specified as follows.

QoS Levels

We consider 3 main classes of traffic although more classes can be included as needed (as shown in Table 3).

Table 3: QoS levels of the proposed MAC protocol.

Class	Traffic	Priority
CBR	Voice	1
VBR-rt	Video	2
Bursty	Web	3
User-defined	User-defined	4

Note: priority 1 > 2 > 3 > 4

Link Speed

The link speed between OLT and ONU is shared by the MAC protocol. The lowest speed is assumed to be 155 Mbps. This is scalable to 622 Mbps, 1.2 Gbps, 2.4 Gbps, and higher speeds.

TDM Frame Length

We assume a maximum access distance of 15 km, which gives a maximum one-way propagation delay of 75 μs. The TDM frame length corresponds to the 2-way propagation delay (i.e., 150 μs).

Request Packet

Request packets (format shown in Figure 8) are transmitted on the minislots of the TDM frame (each minislot of size 3 bytes) and 1 minislot is assigned per user. Thus, the total size of the reservation interval will be 3*N* bytes, where *N* is the number of active users. Here, we assign 3 bits for the CBR class to accommodate for 8 simultaneous channels of voice traffic.

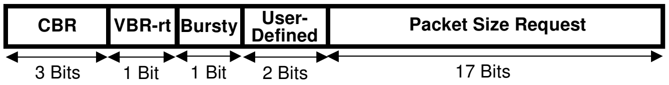


Figure 8: Request format of the proposed MAC protocol.

Grant Packet

This is sent by OLT to ONU and specifies the ONU identity number and the number of data slots allotted. Note that the grant packet is received on the downstream channel by users in different locations at different times but within the TDM frame time.

Packet Header

This contains ranging, TDM frame and byte synchronization, and ONU identity bits.

The operation of the MAC protocol is shown in Figure 9 and summarized as follows:

- Users send request packets (RP_i) in their assigned minislots at start of the i^{th} TDM frame (F_i);
- ONUs then transmit their respective data packets in accordance with the OLT grants received in the $(i-1)^{\text{th}}$ TDM frame (F_{i-1});
- Once RP_i is received by OLT, it computes the grants and broadcasts it back to the ONUs in the $(i+1)^{\text{th}}$ TDM frame (F_{i+1}).

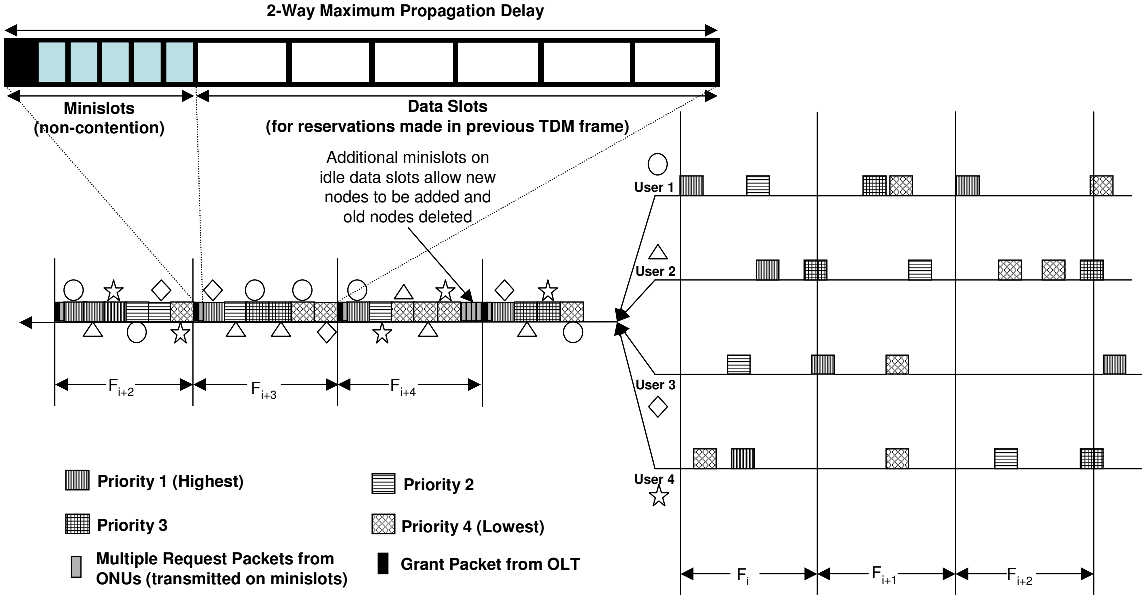

2-Way Maximum Propagation Delay
Minislots (non-contention)
Data Slots (for reservations made in previous TDM frame)
Additional minislots on idle data slots allow new nodes to be added and old nodes deleted
User 1
User 2
User 3
User 4
F_{i+2}
F_{i+3}
F_{i+4}
F_i
F_{i+1}
F_{i+2}
Priority 1 (Highest)
Priority 2
Priority 3
Priority 4 (Lowest)
Multiple Request Packets from ONUs (transmitted on minislots)
Grant Packet from OLT

Figure 9: Operation of the proposed MAC protocol in single wavelength for 4 users.

Figure 9 also shows the physical flow of packets arriving at different users and transmitted on the link according to the proposed MAC protocol. For simplicity, we show 4 users generating fixed-size packets of priorities 1 to 4, where priority 1 is the highest and priority 4 is the lowest. The MAC protocol serves priority 1 packets of all the users in a cyclic way, then priority 2, and so on. As such, it may be treated as a class-based cyclic server. For the case when ONUs with the same priority request for more bandwidth exceeding one TDM frame time, the available bandwidth will be allocated to each ONU in proportion to the requested bandwidth. The ONUs will then need to submit another reservation in a subsequent TDM frame to request for additional bandwidth. A packet is only considered for transmission if it is completely generated or has reached the buffer of the user before the end of the TDM frame. This is because the grant packet is sent at the start of next TDM frame and cannot be included by the user unless it is complete. Hence, a data packet goes through a mandatory delay of one TDM frame due to the request and grant mechanism. We can reduce this delay by pre-allocating a minimum number of data slots to the users, which can increase in subsequent TDM frames if the users make reservations in their minislots.

Figure 10 shows how multiple wavelengths are multiplexed in the WDM PON. Multiple TDM frames of different wavelengths with different priorities are transmitted in parallel using WDM, and each TDM frame may come from a single heavy user or shared by several light users in the time domain.

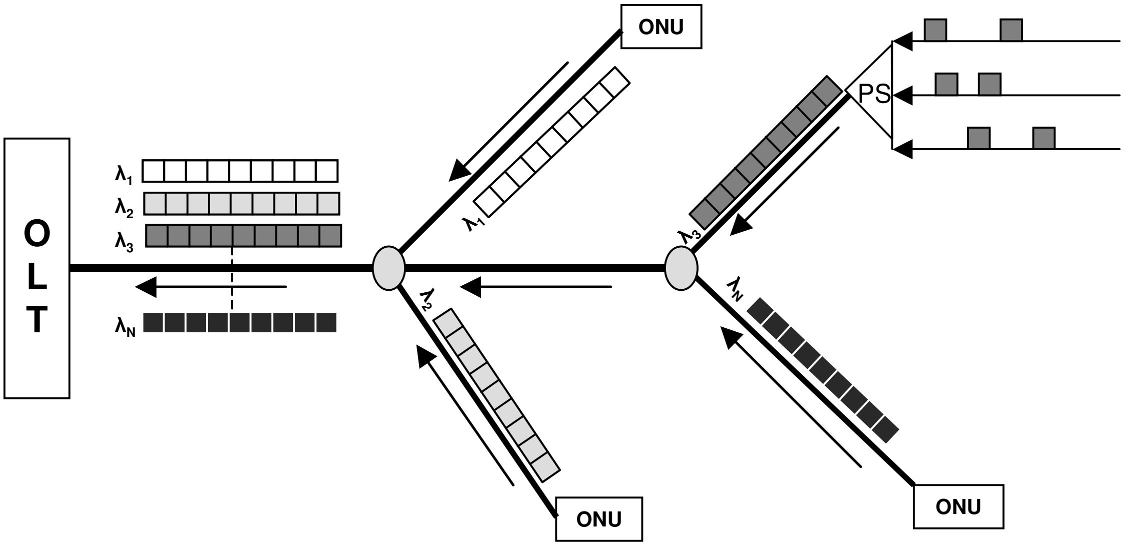


Figure 10: Multiple wavelength (WDM) upstream bandwidth sharing.

Most PON protocols either use reservation protocols (where minislots are used to reserve for larger data slots) or static time division multiple access (TDMA) protocols (where each ONU is statically assigned a fixed number of data slots). The access scheme that we present here combines the advantages of reservation and static TDMA by preallocating a minimum number of data slots for the ONU, which can be increased dynamically in subsequent TDM frames through reservation minislots on a needed basis. By preallocating data slots, a data packet can be transmitted immediately instead of waiting for a duration equivalent to the two-way propagation delay associated with the request and grant mechanism in reservation protocols. However, the number of data slots that can be pre-allocated must be kept to a minimum since if more slots are preallocated, then there is a possibility that slots could become wasted when the network load is low and this is the main disadvantage of static TDMA. Other novelties of the scheme are that it not only arbitrates upstream transmission and prevents collisions, but also varies bandwidth according to QoS demand and priority, accounts for varying delays caused by physical fiber length difference, and handles the addition/reconfiguration of network nodes efficiently [7].

3.3 Performance Evaluation

The performance of the proposed MAC protocol is analyzed and extensively simulated using OPNET. We consider both ATM and Ethernet traffic types at each ONU. In the case of ATM traffic, an arrival corresponds to a single ATM cell whereas for Ethernet, we modeled each variable-size Ethernet packet as a batch (bulk) arrival of ATM cells. In each case, the arrival probability distribution is Poisson. We employ simulation parameters that are consistent with the ITU-T standards and they are described as follows. For the single arrival ATM case, the link rate is chosen to be 155.52 Mbit/s shared by 10 ONUs. A data cell comprises 3 header bytes and 53 data bytes, and a TDM frame comprises 53 data slots (equivalent to the duration of 53 data cells of 56 bytes each) with 1 data slot allocated for minislot reservation (i.e., 52 slots allocated for data cells). For a system with more ONUs, a TDM frame with more minislots can be used for reservation. This gives a TDM frame duration of 53 x 56 x 8 bits / 155.52 Mbit/s = 152.67 μs and a slot time of 152.67 μs / 53 = 2.8807 μs. Note that since the TDM frame duration is typically designed for the maximum round-trip (two-way propagation) delay, our TDM frame duration is equivalent to a maximum distance of (152.67 x 10^{-6} s) x (2 x 10^{8} m/s) / 2 or roughly 15 km. The traffic load refers to the ratio of the data generation rate over the link rate. The batch arrival interval in each ONU is exponential distributed with mean 10 x (53 x 8)/(link rate x load x average batch size), while single arrival is a special case with batch size of 1. The bandwidth is dynamically allocated using the request and grant mechanism, and no prioritization is considered. For the batch arrival Ethernet case, the TDM frame and network settings are the same as before. The only change is the traffic pattern where we have simulated Poisson arrivals at each ONU with batch size uniformly distributed between 1 and 30 data cells (30 data cells is roughly equivalent to one maximum-length Ethernet packet of 1,500 bytes).

As can be seen in Figure 11, the proposed MAC protocol is able to accommodate both traffic types very efficiently, achieving a high normalized throughput of 0.928. The average queue length (Figure 12), indicating the number of buffered packets, remains low up till a load of about 0.8 after which it increases dramatically due to heavy traffic. The average delay is depicted in Figure 13.

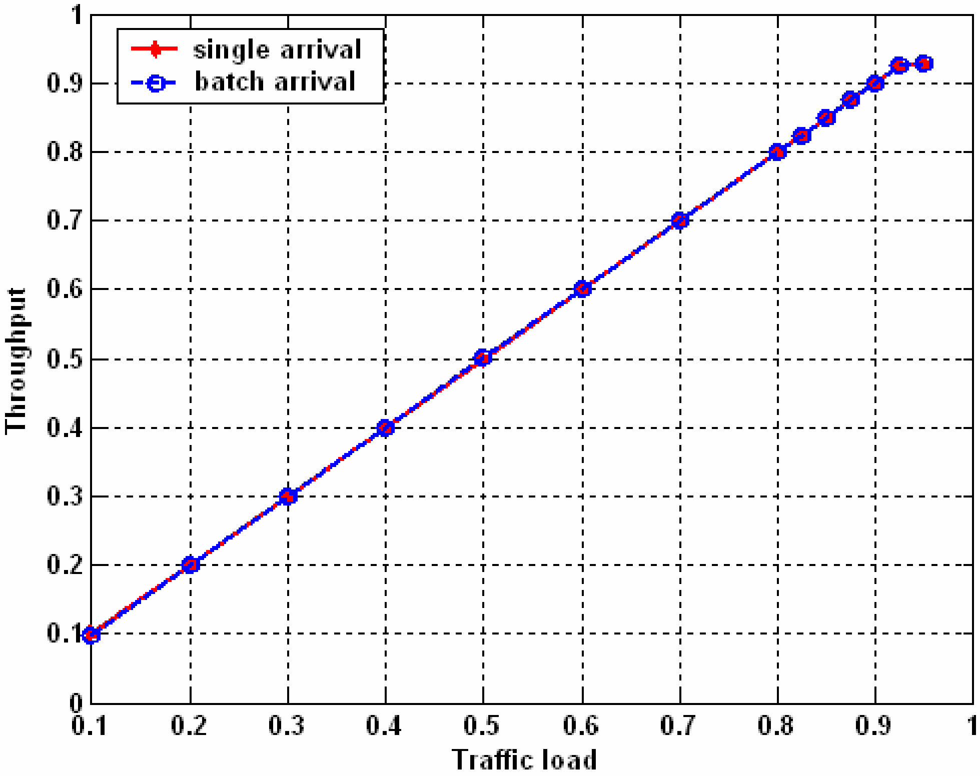


Figure 11: Comparison between single arrival and batch arrival: normalized throughput versus offered traffic load.

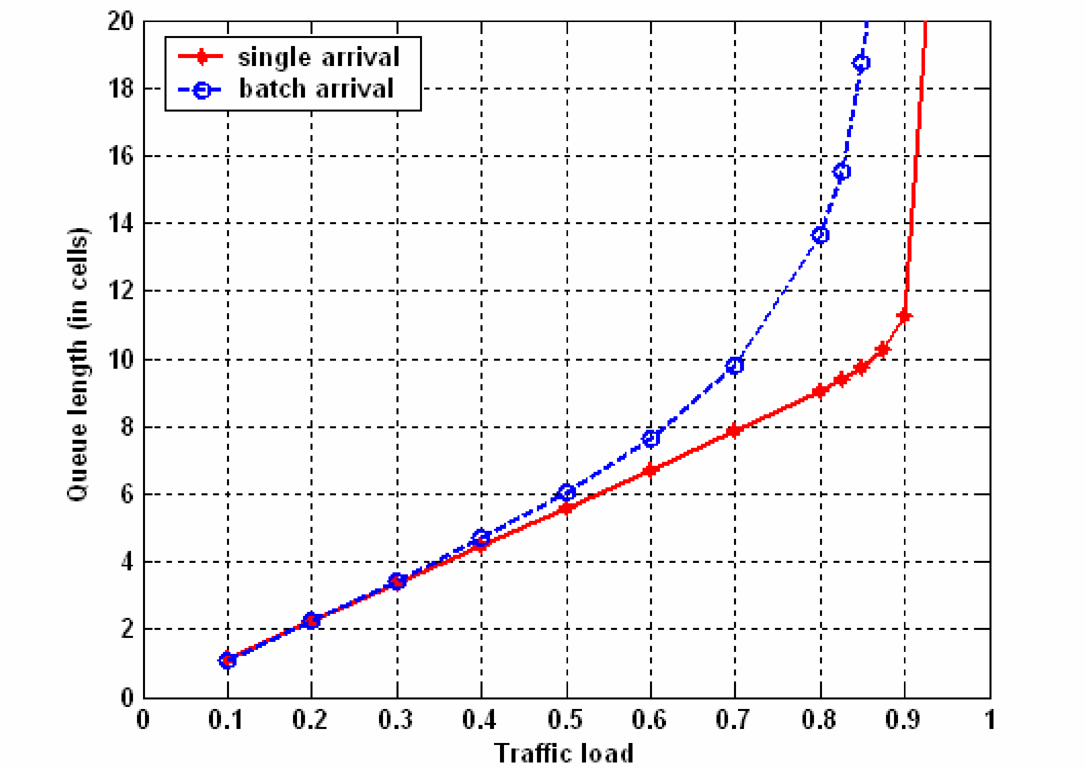


Figure 12: Comparison between single arrival and batch arrival: average queue length versus offered load.

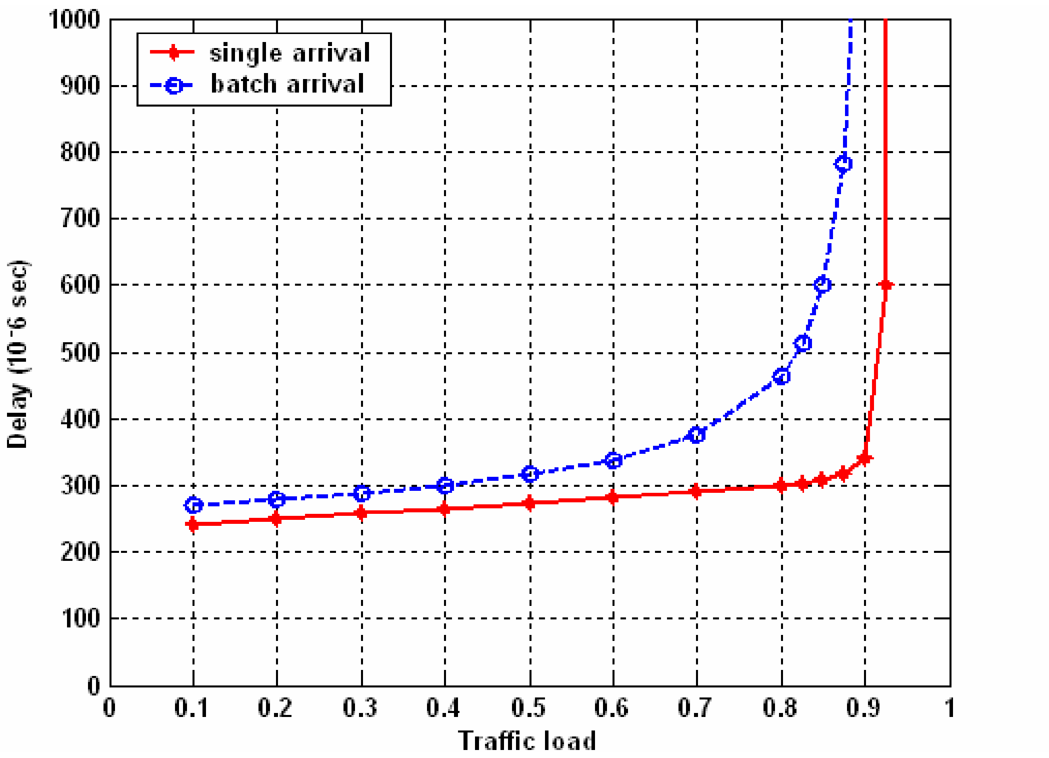


Figure 13: Comparison between single arrival and batch arrival: average delay versus offered load.

3.3.1 Queue Length and Delay Analysis of the Proposed MAC Protocol

In the Appendix, we modeled the average queue length and delay using the discrete-time, bulk-service D/D/1 model. Figure 14 to 17 show both upper and lower bounds of the analytical equations yielded a close match against the simulation.

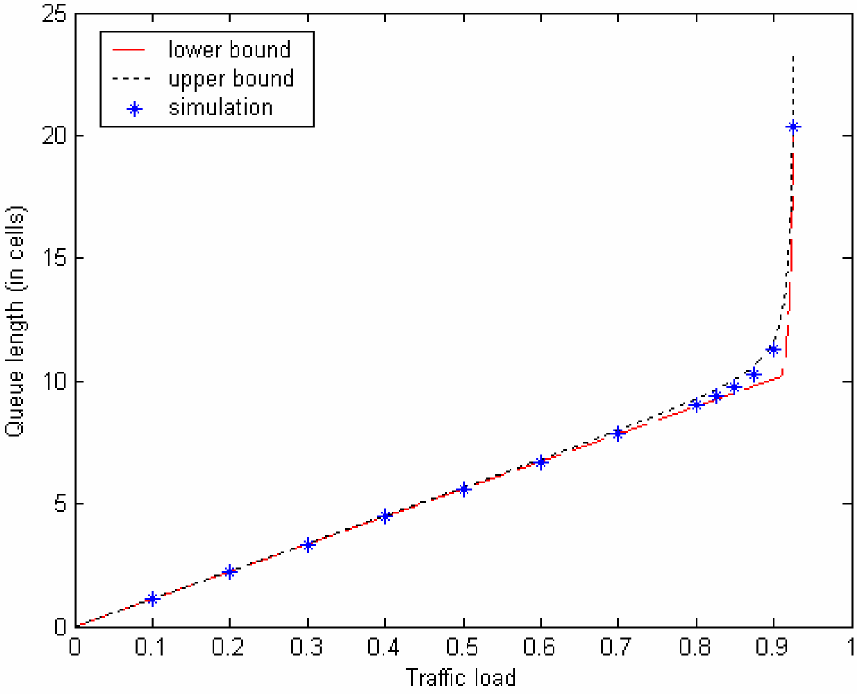


Figure 14: Comparison between analysis results and simulation results: average queue length versus offered load (single arrival).

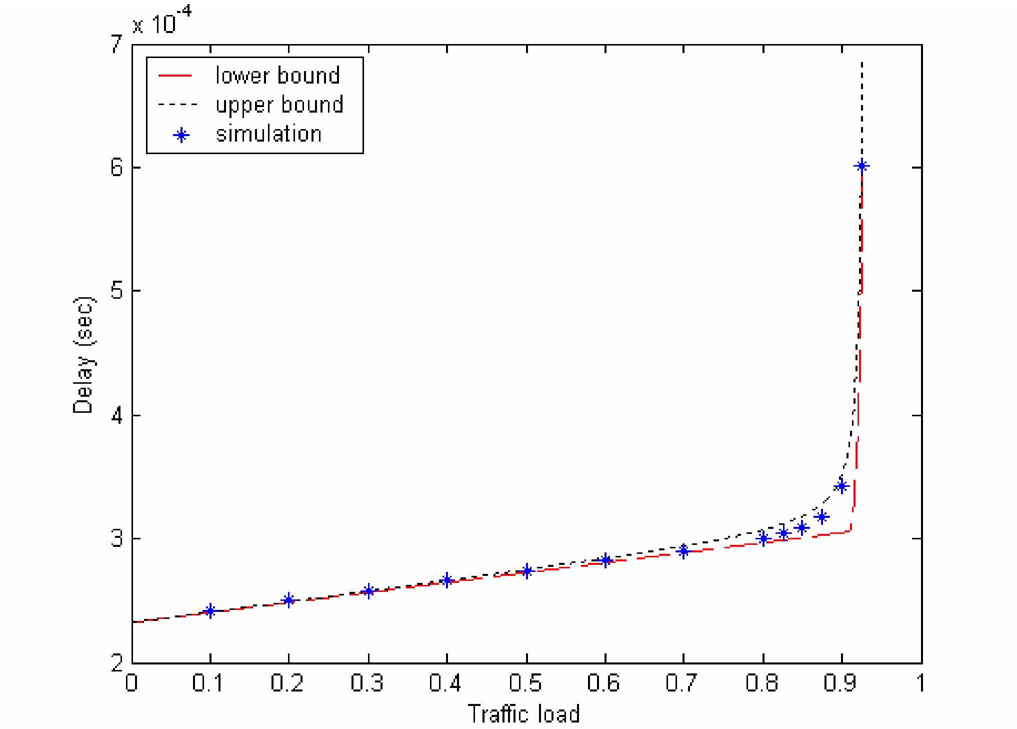


Figure 15: Comparison between analysis results and simulation results: average delay versus offered load (single arrival).

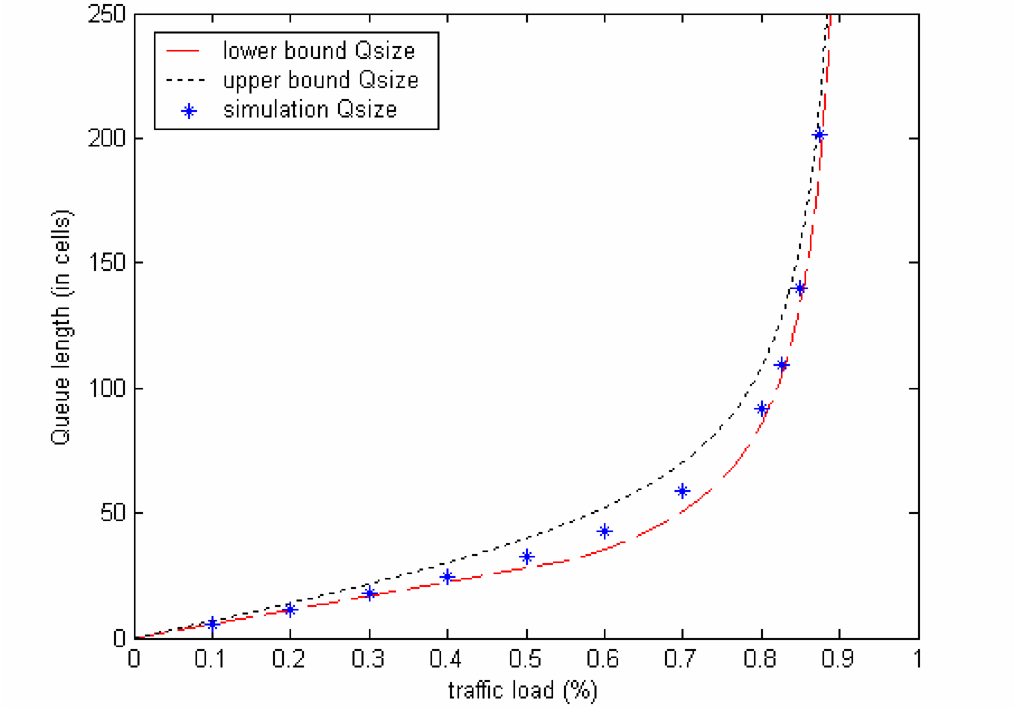


Figure 16: Comparison between analysis results and simulation results: average queue length versus offered load (batch arrival).

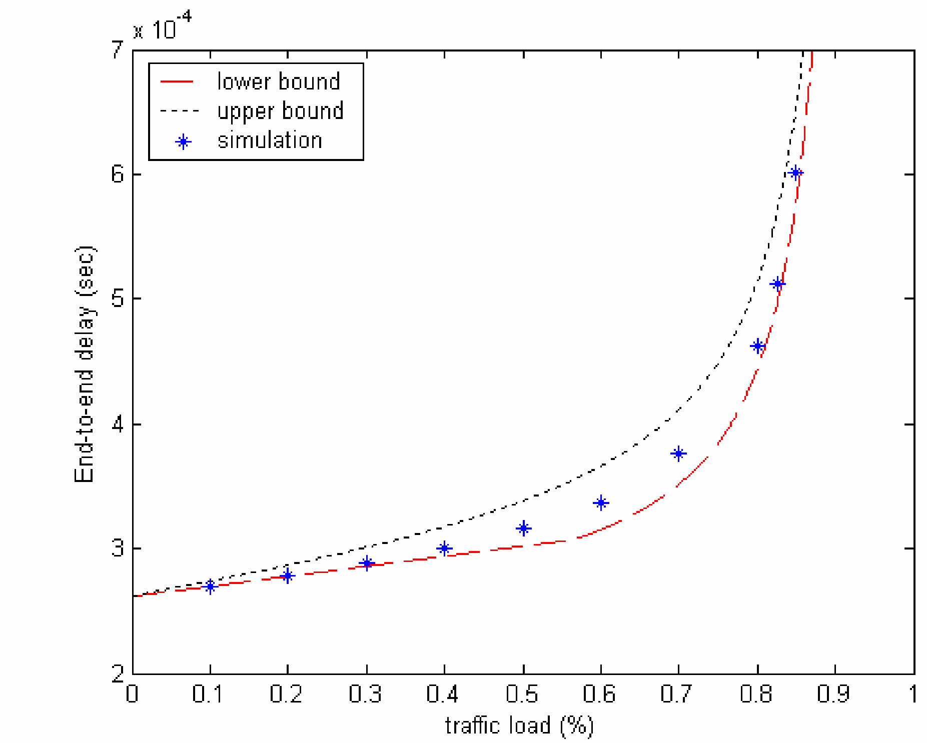


Figure 17: Comparison between analysis results and simulation results: average delay versus offered load (batch arrival).

3.3.2 Effect of Data Slot Pre-allocation on the Proposed MAC Protocol

In this section, we demonstrate the beneficial effect of data slot pre-allocation on the performance of the proposed MAC protocol. The parameters for the simulation are the same as the previous section except that there are no batch arrivals. For the pre-allocation case, each ONU is pre-assigned a minimum of one data slot. A request packet transmitted on a minislot increases this minimum number to a number indicated in the request packet. For the case without pre-allocation, all data slots are assigned only after a request is made on a minislot.

From Figure 18, it is evident that the throughput of the pre-allocation is as efficient as the case without pre-allocation, indicating that a pre-assignment of data slots will not lead to a possible sacrifice in bandwidth when these data slots are idle when users do no have data packets to transmit. Figure19 and 20 show the more interesting case on the average queue length and delay performance. Pre-allocation of data slots clearly reduces the average queue length and delay, even when the network is heavily loaded. Under extreme high loads, the performance of both protocols converge which is expected since all data slots in a TDM frame tend to become filled continuously. In Figure 18 to 20, we also illustrate the performance of static TDMA where 5 cells are allocated for each ONU with the remaining 3 slots of the TDM frame are wasted. Although static TDMA performs best under low load, it is due to the single arrival assumption. With batch arrivals, static TDMA performs the worst, as shown in the next section.

To simulate Ethernet PON, traffic is generated as a batch arrival pattern with batch size uniformly distributed between 1 and 30 cells and exponentially distributed inter-batch arrival time. As can be seen from Figure 21 to 23, static TDMA performs poorly because the average batch size is 15.5 cells per packet, which implies an average of 3 TDM frames are needed to transmit all cells in a packet. Since each packet arrives randomly, in static TDMA, users are not able to request for more bandwidth appropriate for the batch size. Note that the performance of the pre-allocation algorithm is now slightly better than without pre-allocation under light load. Under heavy load, the pre-allocation algorithm incurs slightly more overheads.

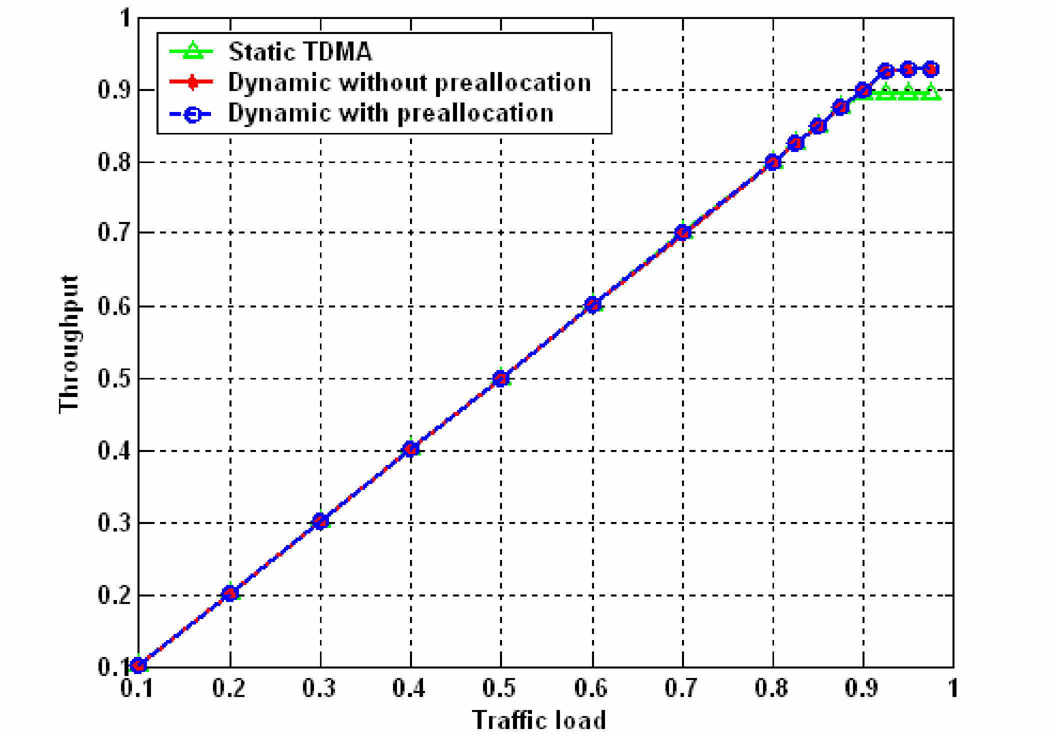


Figure 18: Comparison among different allocation schemes: normalized throughput versus traffic load (single arrival).

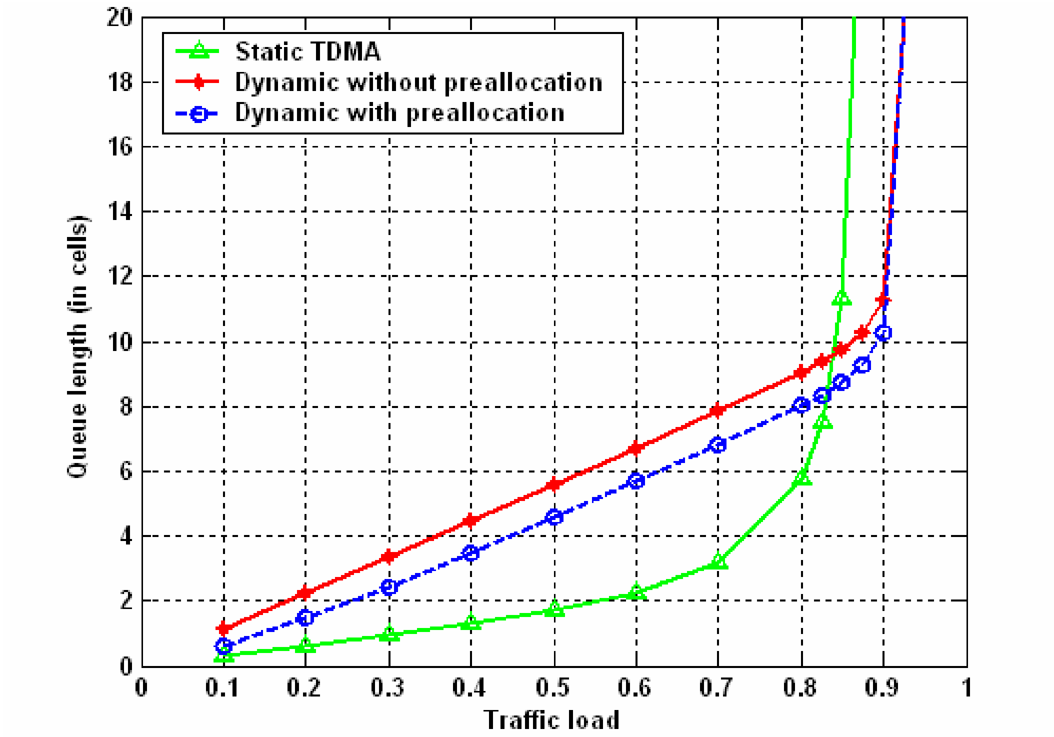


Figure 19: Comparison among different allocation schemes: average queue length versus traffic load (single arrival).

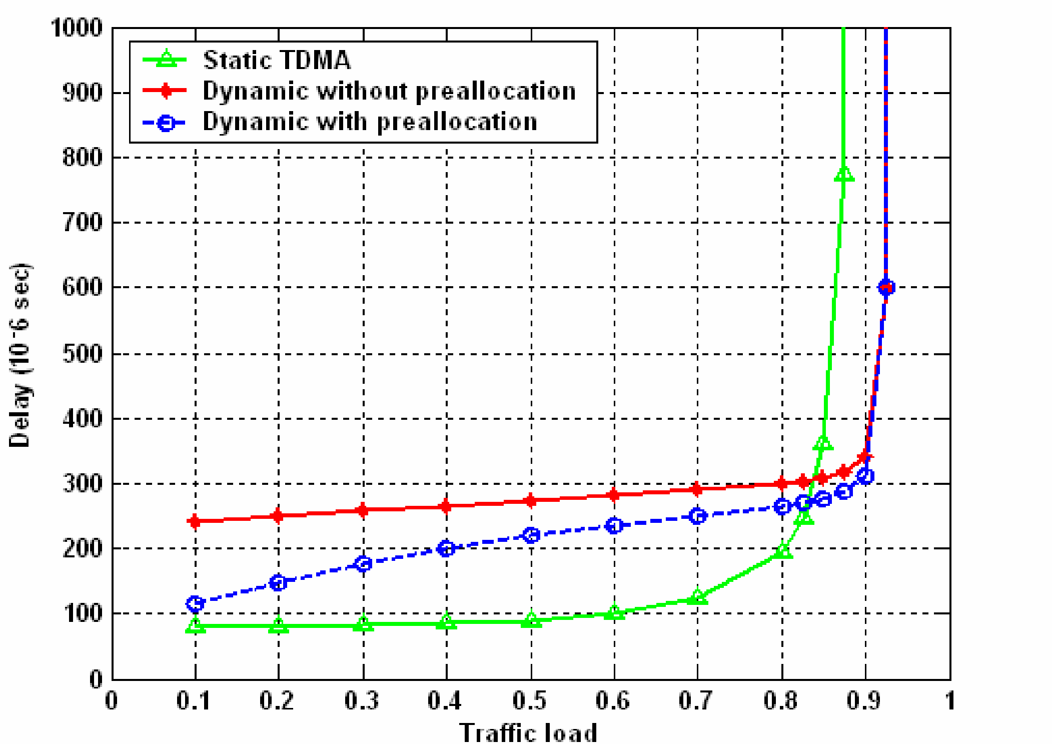


Figure 20: Comparison among different allocation schemes: average delay versus traffic load (single arrival).

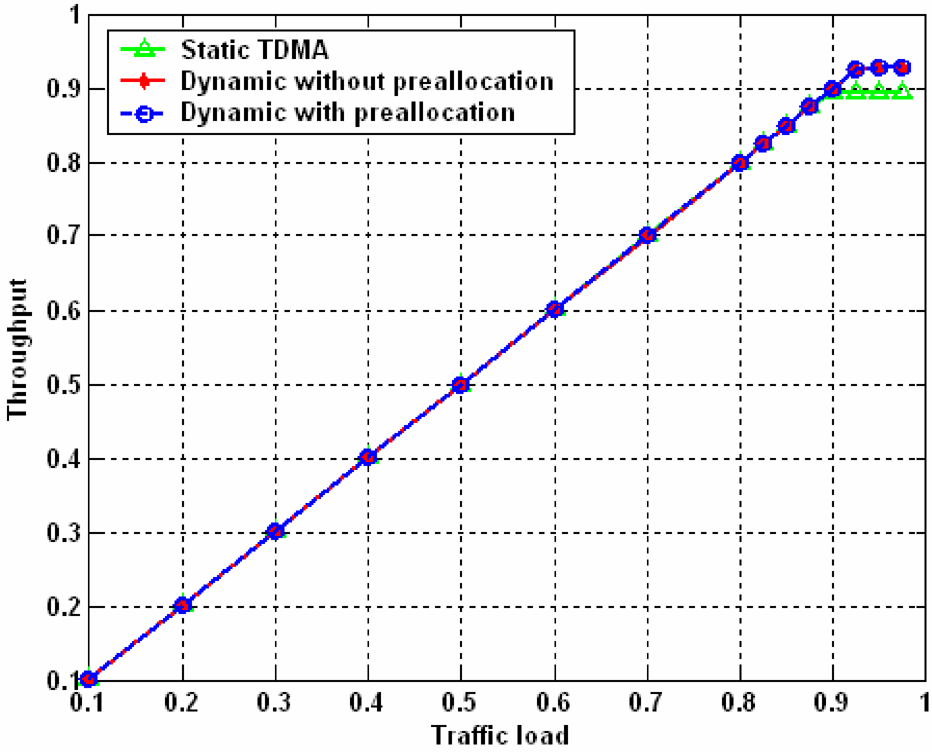


Figure 21: Comparison among different allocation schemes: normalized throughput versus traffic load (batch arrival).

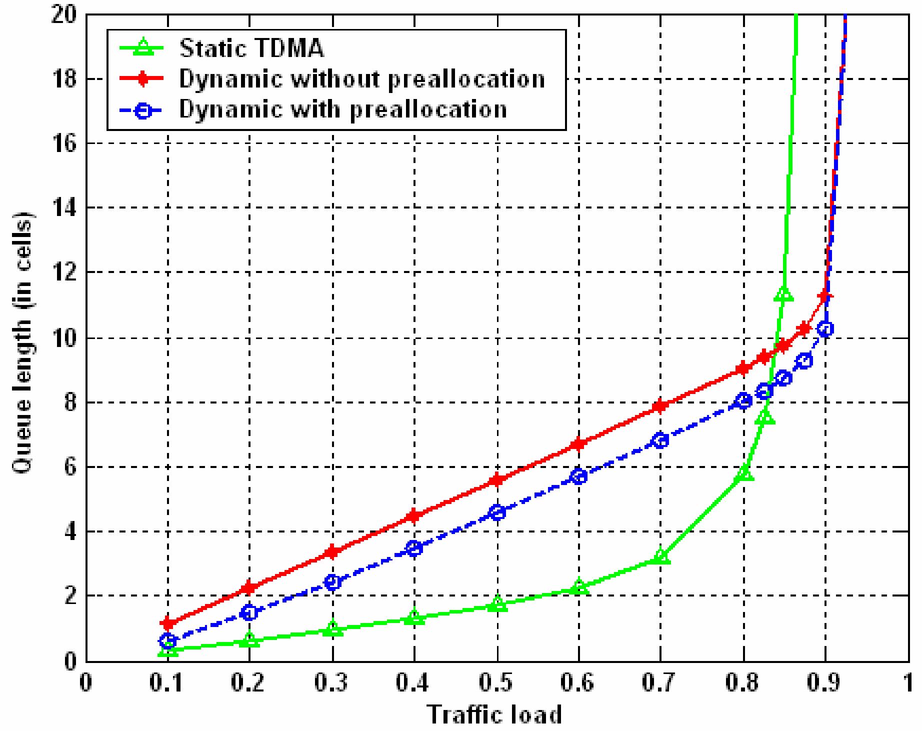


Figure 22: Comparison among different allocation schemes: average queue length versus traffic load (batch arrival).

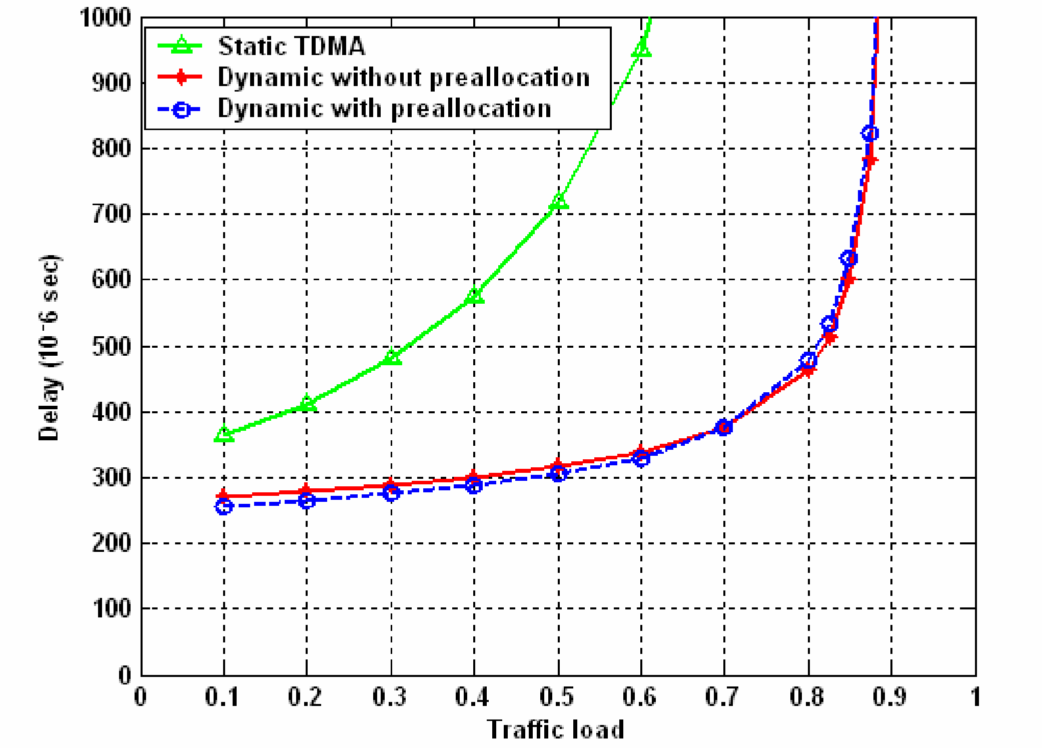


Figure 23: Comparison among different allocation schemes: average delay versus normalized traffic load (batch arrival).

CHAPTER 4

DELTA COMPRESSION FOR FAST CONTENT DOWNLOD

4.1 Problem Definition

In the previous chapter, we proposed a new MAC protocol for WDM PON with pre-allocation, which provides QoS with tremendous bandwidth. In this chapter, we investigate how to improve application performance by reducing traffic redundancy. Internet traffic can be regarded as a sequence of packets or files. The redundancy among adjacent files or packet can be removed by delta compression algorithms. Although there are some heuristics about how to do delta compression, there is no theoretical analysis for the best compression ratio that can be achieved by delta compression algorithms. In this research, based on a two-state Markov content generation model, two closed form expressions are derived for the compression ratios of the delta compression with fixed-length coding and variable-length coding. It is also shown that the lower bound for delta compression ratio is the conditional entropy. This research aims to validate the effectiveness of delta compression using theoretical analysis and measurement rather than heuristics.

The delta compression problem can be stated more precisely as: Given the previous packet X_k as a reference, design an invertible, lossless compression algorithm F, such that the length of compressed packet $F(X_{k+1})$ is minimized. X_k is called the reference packet and X_{k+1} is called the target packet. Let us denote the $(k+1)$th packet as a *byte* (i.e., 8-bit) sequence $X_{k+1} = \{X_{k+1}(1), X_{k+1}(2), \dots , X_{k+1}(N)\}$ with length N. Hence, $X_{k+1}(i)$ represents the ith byte of packet X_{k+1}. Let $L(X_{k+1})$ denote the packet length of X_{k+1}. Let

X_k(*position*, *len*) denote a substring of X_k. The variable "*position*" is the location of this substring in X_k, and "*len*" is the length of this substring. For example: if X_k = [0101100], then $L(X_k) = 7$, $X_k(2, 4)$ = [1011]. Let $\tilde{X}_{k+1}$ denote the compressed packet of X_{k+1}. Then

$$\tilde{X}_{k+1} = F(X_{k+1}) \, , \; X_{k+1} = F^{-1}(\tilde{X}_{k+1}) \, ,$$

$$compression\_ratio = \frac{L(\tilde{X}_{k+1})}{L(X_{k+1})} \, .$$

In the classical Lempel-Ziv (LZ) algorithm, a single, individual packet X_{k+1} is compressed. The average entropy per bit of X_{k+1} is $H(X_{k+1})/8N$, which is the lower bound of the compression ratio (defined as compressed packet size/uncompressed packet size). In delta compression, the previous packet X_k is known to both sender and receiver. As a result, $H(X_{k+1} | X_k)/8N$, the average conditional entropy per bit of X_{k+1}, becomes the lower bound of the compression ratio. Because $H(X_{k+1} | X_k)$ is not greater than $H(X_{k+1})$, a higher compression ratio can be achieved by delta compression compared to classical LZ.

3.2 Analytical Model

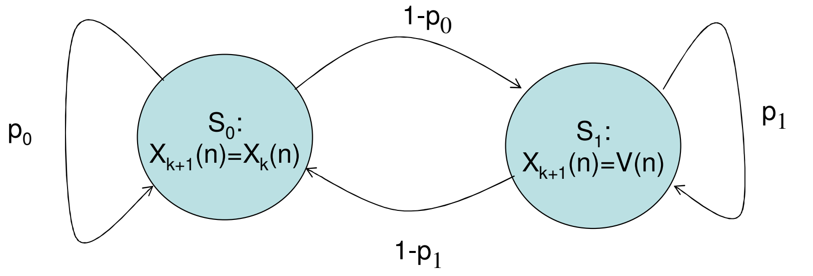
3.2.1 Content Generation Model

Some fundamental limits on the bandwidth savings achievable with delta compression are provided in this section. Delta compression can, in general, improve the download of Internet traffic as long as inter-packet correlation exists. However, the complexity of real Internet traffic does not permit a close form expression for the compression ratio. To simplify our analysis, we assume all packets have the same size of

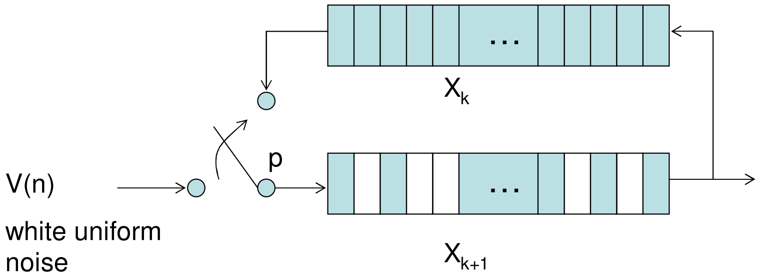
N bytes. Suppose packet X_{k+1} is generated from the previous packet X_k (i.e., old information already sent) and an *N*-byte packet *V* (can be considered new information, independent of old information, to be included in packet X_{k+1}) according to a two-state Markov model as shown in Figure 24(a). When the process is in state S_0, a byte of old information is copied from $X_k(n)$ to $X_{k+1}(n)$ and a byte of new information is copied from $V(n)$ to $X_{k+1}(n)$ in state S_1. The transition matrix is $\begin{bmatrix} p_0 & 1-p_0 \\ 1-p_1 & p_1 \end{bmatrix}$, where $0 \le p_0 \le 1$, $0 \le p_1 \le 1$.

By setting $p = p_0 = 1 - p_1$, a simplified model can be expressed as:

$$X_{k+1}(n) = \begin{cases} X_k(n), & \textit{with probability} \quad p \\ V(n), & \textit{with probability} \quad 1-p \end{cases} \quad , \quad \text{where n} = 1, 2, \ldots, \text{N}.$$



(a) State transmission diagram for packet generation



(b) Simplified packet generation model

Figure 24: Packet generation model.

As shown in Figure 24(b), X_{k+1} is generated by copying a byte from X_k with probability p or copying a byte from V with probability $1 - p$. Subsequently, we assume each byte of X_k and V is generated as an independent and identically distributed (i.i.d) random variable with a discrete uniform distribution between 0 and 255.

3.2.2 Analysis of Entropy Rate

Based on the simplified packet generation model, the conditional entropy of X_{k+1} can be calculated. Because $X_k(n)$ and $V(n)$ (where $n = 1, 2, \ldots, N$) are i.i.d. random variables with a discrete uniform distribution between 0 and 255, $X_{k+1}(n)$ (where $n = 1, 2, \ldots, N$) is also an i.i.d. random variable with a discrete uniform distribution between 0 and 255. Since

$$
\begin{aligned}
&P[X_{k+1}(n) = j \mid X_k(n) = i] \\
&= \begin{cases} p + (1-p)P[V(n) = j], & i = j \\ (1-p)P[V(n) = j], & i \neq j \end{cases} \\
&= \begin{cases} p + \dfrac{1-p}{256}, & i = j \\ \dfrac{1-p}{256}, & i \neq j \end{cases}
\end{aligned}
$$

Denoting $q = p + (1 - p)/256$ as the probability that $X_{k+1}(n)$ is equal to $X_{k+1}(n)$, this gives the average conditional entropy per bit as

$$
\begin{aligned}
&\frac{H[X_{k+1} \mid X_k]}{8N} = \frac{H[X_{k+1}(n) \mid X_k(n)]}{8} \\
&= \frac{\sum_{i=0}^{255} P[X_k(n) = i] H[X_{k+1}(n) \mid X_k(n) = i]}{8} \\
&= -\frac{1}{8}\left[q\log(q) + 255\left(\frac{1-q}{255}\right)\log\left(\frac{1-q}{255}\right)\right] \\
&= -\frac{1}{8}\left(\frac{1+255p}{256}\right)\log(\frac{1+255p}{256}) - \frac{255}{8}\left(\frac{1-p}{256}\right)\log\left(\frac{1-p}{256}\right).
\end{aligned}
$$

For comparison purposes, note that average entropy per bit is

$$\frac{H[X_{k+1}]}{8N} = \frac{H[X_{k+1}(n)]}{8} = 1$$

For a sequence of packets $\{X_1, X_2, \ldots, X_k, X_{k+1}\}$, assuming it is a Markov process and stationary, the entropy rate of the content traffic process is

$$H_0 = \lim_{k\to\infty} \frac{H(X_1, X_2, \cdots, X_{k+1})}{k+1} = \lim_{k\to\infty} H(X_{k+1} \mid X_k)$$

3.2.3 Compression Ratio Analysis of Fixed-length Coding

The average conditional entropy per bit is the lower bound of delta compression ratio, which may only be approached by highly efficient coding of the delta file. In this section, a closed form expression for the compression ratio with fixed length coding is derived. According to the simplified content generation model in Figure 24(b), the new packet X_{k+1} is a mixture of old information bytes from X_k and new information bytes from V. The distance between two adjacent new information bytes can be modeled as a Geometric distributed random variable W: $P(W = n) = C(1 - C)^{n-1}$, where n=1, 2, … and $C = 1 - q = 255(1 - p)/256$. The substring between the two adjacent new information bytes in X_{k+1} is copied from X_k with a length of W-1. The compression ratio for a fixed-length coding scheme is:

$$\eta = \frac{L_1 + 0 \cdot P(W=1) + \min(L_1, L_2) \cdot P(W=2) + L_2 \cdot P(W>2)}{E(W)}$$

where L_1 is the coded length of a new information byte copied from V, and L_2 is the coded length of a multiple-byte substring copied from X_k. The specific values for L_1 and L_2 depend on the coding scheme. Because a one-byte substring copied from X_k can also be treated as a new information byte, it can be coded either by L_1 or L_2 bytes. When $L_1 < L_2$, the closed form expression is

$$\eta = \frac{L_1 + C(1-C)L_1 + (1-C)^2 L_2}{1/C}$$

3.2.4 Compression Ratio Analysis of Variable-length Coding

As mentioned in the previous section, the difference sequence generated by the differencing algorithm is a mixture of "copy" and "add" instructions. According to the simplified content generation model shown in Figure 24(b), the "copy" and "add" instructions are interleaved one after the other. The difference information is contained in the "copy" length parameter, the "add" length parameter, and the added new bytes. Therefore, by performing entropy coding to those three parts will yield an optimal compression ratio. Note that while the length of copied substring can be modeled as a Geometric distributed random variable Y: $P(Y=n)=A(1-A)^{n-1}$, where A = 1-q, the size of added substring can also be modeled as a Geometric distributed random variable Z: $P(Z=n)=B(1-B)^{n-1}$, where n = 1, 2, … and B = q. The lower bound of the compression ratio η for an entropy coding scheme can be derived:

$$\begin{aligned}
\eta &\geq \frac{H(Y)+H(Z)+\log(255)E(Z)}{8[E(Y)+E(Z)]} \\
&= \frac{-\frac{[A\log(A)+(1-A)\log(1-A)]}{A}-\frac{[B\log(B)+(1-B)\log(1-B)]}{B}+\frac{\log(255)}{B}}{8\left(\frac{1}{A}+\frac{1}{B}\right)} \\
&= -\frac{1}{8}[q\log q+(1-q)\log(1-q)]+\frac{\log(255)}{8}(1-q) \\
&= -\frac{1}{8}\left[q\log q+(1-q)\log\left(\frac{1-q}{255}\right)\right] \\
&= \frac{H(X_{k+1} \mid X_k)}{8N}
\end{aligned}$$

3.3 Implementation

Using C++, a fast content downloading application is implemented based on the delta compression scheme with fixed-length coding. The scheme gives an excellent compression gain (and yields a short delta by using greedy differencing algorithm) but not necessarily the best delta creation time. Three bytes are used to code each substring: two for the location of a substring and one for the length of a substring. Therefore, we enforce the maximum size of one matched substring to be 256 bytes. We employ a set of HTML Web pages with different content and random lengths from a forum website, and save them at a content server in both delta compressed and uncompressed formats. The client downloads and decompresses the web pages one after the other in real time over a 1 Mbps 802.11b wireless link.

As shown in Figure 25, the average real-time throughput improvement using delta compression is about 4 times more than the case when there is no compression. This takes into account the overheads needed to transmit and decode the delta file. The average static compression gain is also more than twice that of the commercial WinZip™ software.

	Uncompressed	Delta compressed	WinZip compressed
File size	44 KB	3 KB	7 KB
Equivalent Throughput	303 Kbps	1,171 Kbps	N/A

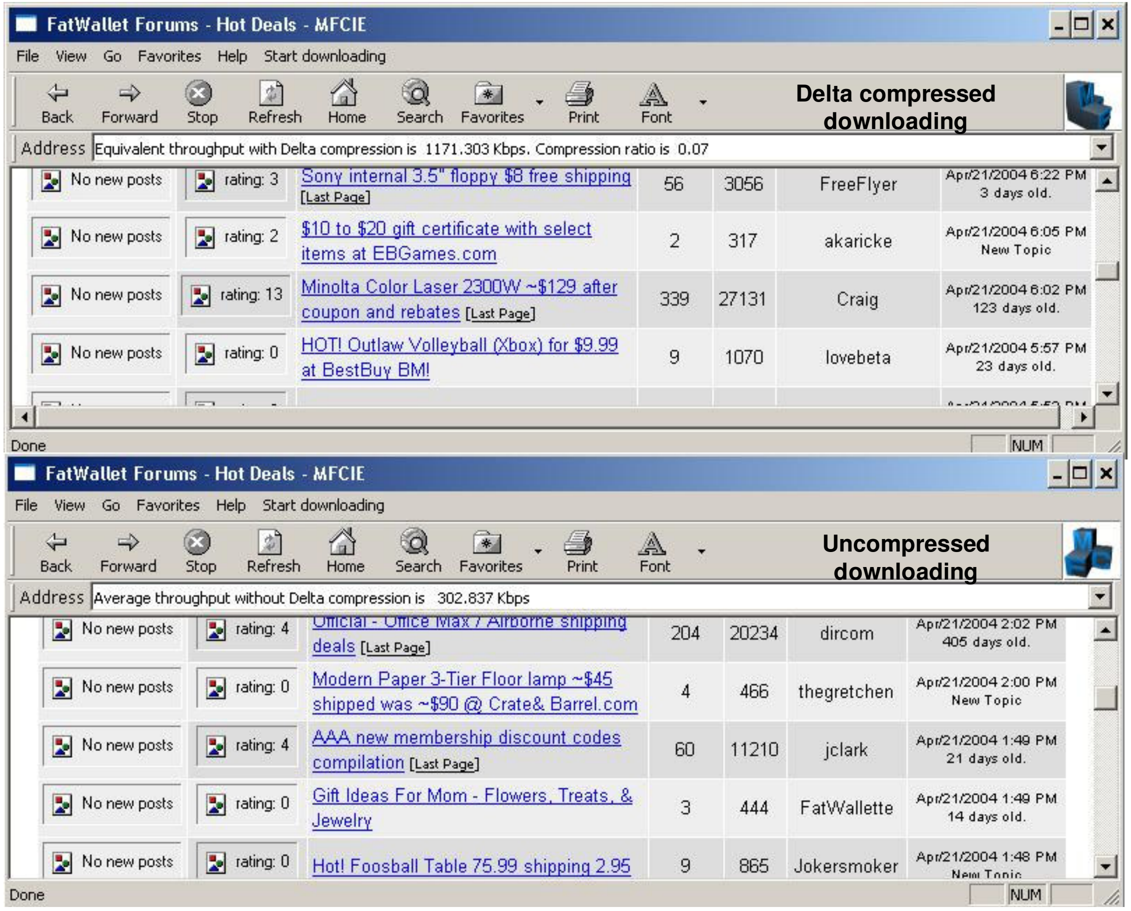
non real-time

real-time

Compression ratio, η = Compressed_size/Uncompressed_size

Average throughput, *T* =Transmitted_bits/Time_interval

Equivalent throughput = *T*/η


FatWallet Forums - Hot Deals - MFCIE
File View Go Favorites Help Start downloading
Back Forward Stop Refresh Home Search Favorites Print Font
Delta compressed downloading
Address
Done
NUM
FatWallet Forums - Hot Deals - MFCIE
File View Go Favorites Help Start downloading
Back Forward Stop Refresh Home Search Favorites Print Font
Uncompressed downloading
Address
Done
NUM

Equivalent throughput with Delta compression is 1171.303 Kbps. Compression ratio is 0.07

No new posts	rating: 3	Sony internal 3.5" floppy $8 free shipping [Last Page]	56	3056	FreeFlyer	Apr/21/2004 6:22 PM 3 days old.
No new posts	rating: 2	$10 to $20 gift certificate with select items at EBGames.com	2	317	akaricke	Apr/21/2004 6:05 PM New Topic
No new posts	rating: 13	Minolta Color Laser 2300W ~$129 after coupon and rebates [Last Page]	339	27131	Craig	Apr/21/2004 6:02 PM 123 days old.
No new posts	rating: 0	HOT! Outlaw Volleyball (Xbox) for $9.99 at BestBuy BM!	9	1070	lovebeta	Apr/21/2004 5:57 PM 23 days old.

Average throughput without Delta compression is 302.837 Kbps

No new posts	rating: 4	Official - Office Max / Airborne shipping deals [Last Page]	204	20234	dircom	Apr/21/2004 2:02 PM 405 days old.
No new posts	rating: 0	Modern Paper 3-Tier Floor lamp ~$45 shipped was ~$90 @ Crate& Barrel.com	4	466	thegretchen	Apr/21/2004 2:00 PM New Topic
No new posts	rating: 4	AAA new membership discount codes compilation [Last Page]	60	11210	jclark	Apr/21/2004 1:49 PM 21 days old.
No new posts	rating: 0	Gift Ideas For Mom - Flowers, Treats, & Jewelry	3	444	FatWallette	Apr/21/2004 1:49 PM 14 days old.
No new posts	rating: 0	Hot! Foosball Table 75.99 shipping 2.95	9	865	Jokersmoker	Apr/21/2004 1:48 PM New Topic

Figure 25: Delta compression performance.

The access delay is an important metric for content downloading, which is defined as the time difference between the instance when the client sends a request message to the server and the instance when the client retrieves the original content from the server. The access delay has several components: transmission delay, which is defined as packet size divided by link rate; propagation delay, which is the speed of light; queuing delay, which depends on traffic load at network nodes; and processing delay at application layer. Moreover, HTML files are delivered by TCP protocol, which involves congestion control delay by TCP window control mechanism. The effects of compression to those delay components are listed in Table 4. Denoting coding/decoding delay as the total time used to compress content at the server and the time to decompress at the client, denoting transfer delay as the summation of other delay components other than coding/decoding delay, we trade computing power and coding/decoding delay for network bandwidth and transfer delay, by using delta compression. In order to obtain a shorter access time, the transfer delay gained from compressed content should be larger than the additional coding/decoding delay of delta compression.

In order to improve the speed of delta compression, hash table can be used for substring matching. The flow chart of setting up a hash table based on the reference file is shown in Figure 26. The measured delay components are shown in Table 5. The greedy differencing algorithm is used. The computer used for measurement is a Dell laptop with one P4 2GHz CPU.

Table 4: Effects of compression to delay components.

	No compression	With compression
File size	Larger	Smaller
Processing delay	Smaller	Larger
Transmission delay	Larger	Smaller
Propagation delay	Same	Same
Queuing delay	Larger	Smaller
TCP control delay	Larger	Smaller

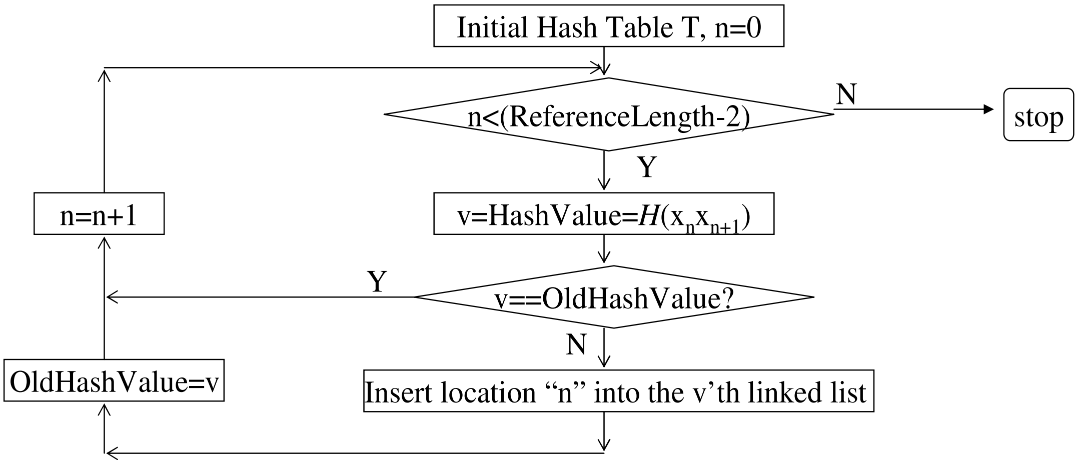

Initial Hash Table T, n=0
n<(ReferenceLength-2)
N
stop
Y
n=n+1
v=HashValue=$H(x_n x_{n+1})$
Y
v==OldHashValue?
N
OldHashValue=v
Insert location “n” into the v’th linked list

Figure 26: Hash table set up.

Table 5: Measured delay components.

	Transfer delay (Sec)	Compression delay (Sec)	Decompression delay (Sec)
Without delta compression	1.1	0	0
Delta compression without Hash method	0.2	2.7	0.05
Delta compression with Hash method	0.2	0.1	0.05

*File size is about 44kB

CHAPTER 5

AN SLA-AWARE TRANSPORT PROTOCOL FOR ETHERNET SERVICES

5.1 QoS in Ethernet Services

Carrier-grade Ethernet is currently regarded as one important driving forces for the growth of the telecommunication market [39]. The high capital and operational expenditure of today's MAN/WAN is due to the legacy IP/ATM /SONET/WDM overlaid architecture. As shown in Figure 27, The new transport solution consolidates the lower three layers and replaces them by an optical Ethernet transport [40]. Ethernet services convert Ethernet from a best effort technology to a service level agreement (SLA) driven carrier-grade technology, extending the simplicity and flexibility of Ethernet beyond the LAN to the MAN/WAN [41].

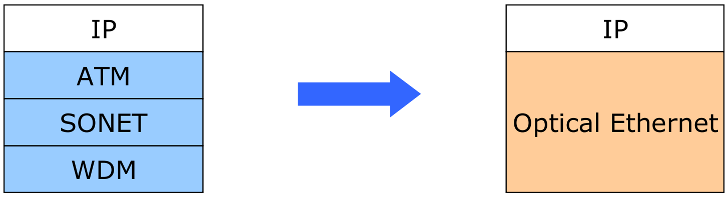


Figure 27: Evolution of transport networks.

The concept of Ethernet service is defined by Metro Ethernet Forum (EFM). A Metro Ethernet Network (MEN) is a network that connects geographically separated enterprise LANs in a metro area using Ethernet. Customer Equipment (CE) attaches to a MAN at the User-Network Interface (UNI) using a standard 10M/100M/1G/10G Ethernet interface. The concept of Ethernet virtual connection (EVC) is defined as an association of two or more UNIs (subscriber sites) enabling the transfer of Ethernet frames between them. Ethernet services can be point-to-point (called Ethernet Line) and multipoint-to-multipoint (called Ethernet LAN). EFM defined two types of Ethernet line services: Ethernet private line (EPL) and Ethernet virtual private line (EVPL). EPL is the Ethernet equivalent to the leased line. Bandwidth is dedicated to each customer with excellent quality of service (QoS). In EVPL, Multiple EVCs are statistical multiplexed in a UNI. Therefore bandwidth can be shared among customers with high bandwidth utilization. Traffic policing has to be used to provide QoS, similar with Frame Relay and ATM. Bandwidth profile can be defined per class of service (CoS), per EVC, or per ingress UNI in an SLA. A bandwidth profile includes at least four parameters: CIR, CBS, EIR, and EBS. CIR conformant traffic will be delivered with high priority, as required by SLA. EIR conformant traffic will be delivered with best of effort if spare bandwidth is available. The traffic out of EIR profile will be discarded [42]. In [43], a centralized admission control scheme is proposed for Ethernet services. In [44], the bandwidth endowed spanning tree problem is investigated in detail to ensure each hop have sufficient capacity to satisfy bandwidth demands. Verizon's experience with deploying Ethernet services is introduced in [45].

5.2 EVPL Channel Capacity

According to information theory, the capacity of a channel, which is defined as the mutual-information between sender and receiver, is the upper bound of the data rate that can be delivered through the channel without error [46]. In order to evaluate the channel capacity of EVPL, an erasure channel model is proposed. As shown in Figure 28, the traffic from the sender is policed according to the SLA at the egress point. The packet that falls into the CIR profile will be marked with high priority and delivered with guaranteed QoS. A packet that falls out of the CIR profile, but still in the EIR profile, will be marked with low priority and delivered in a best-effort manner. A packet that falls out of the EIR profile will be discarded immediately. Within the transport network, a high priority packet suffers a loss probability of α, which is defined in the SLA. A best-effort packet suffers a loss probability of β, which depends on background traffic, and can be any value between 0 and 1. For a prioritized channel as shown in Figure28, the capacity can be expressed as:

$$C = CIR \cdot (1 - \alpha) + EIR \cdot (1 - \beta) \qquad (1)$$

Given that $0 \leq \beta \leq 1$, the lower bound of EVPL channel capacity can be derived as:

$$C = CIR \cdot (1 - \alpha) \qquad (2)$$

which is equal to the capacity of a EPL channel with the same CIR and loss rate parameters.

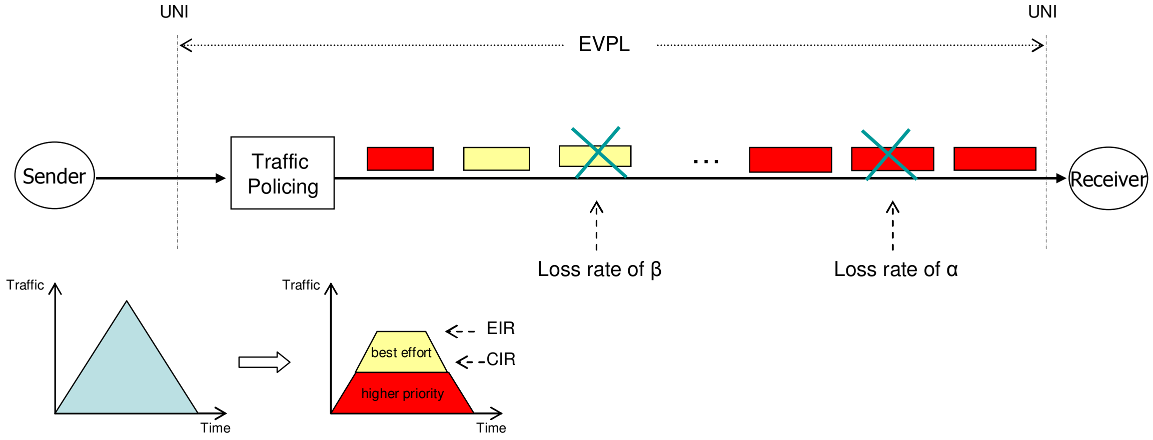


Figure 28: Channel model for Ethernet Virtual Private Line.

5.3 Constrains of Traditional TCP over EVPL

In this dissertation, TCP-Reno and TCP-Sack are considered as two examples of traditional TCP protocols. TCP-Reno reduces the number of timer-out and the slow start occurrences by using fast retransmission [16]. TCP-Reno retransmits a packet after the number of duplicate acknowledgments (ACKs) for the packet exceeds a threshold. The threshold is normally three. Two state variables are kept at the TCP sender: slow start threshold (*ssthresh*) and the size of congestion window (*Cwnd*).

TCP-Sack has a better lost recovery algorithm in that the receiver can inform the sender about all TCP segments that have arrived successfully [47]. TCP senders need only retransmit the segments that are actually lost and can recover more effectively from the case of multiple losses in one RTT. Both TCP-Reno and TCP-Sack share the AIMD feature of TCP congestion control. In the congestion avoidance phase, the *Cwnd* is incremented by one segment per RTT if there is no loss, but decreased by a half after a single loss event indicated by triple duplicate ACKs.

When all packet losses are indicated by triple duplicate ACKs, the TCP throughput *B* can be expressed as:

$$B = \frac{\sqrt{3}}{RTT\sqrt{2bp}} \times Packet\_size \qquad (3)$$

where *RTT* is the round trip time, *p* is the packet loss rate, and *b* is the number of packets that are acknowledged by a received ACK [48]. In many TCP implementations, *b* is 2.

TCP cannot fully utilize the reserved bandwidth when RTT is large, because TCP throughput is inversely proportional to RTT. This problem can be relieved temporally by using jumbo frames. But the inverse proportional decay of TCP throughput is a fundamental constraint. TCP was designed for best-effort packet switching networks without previous knowledge about available bandwidth. The TCP sender has to probe the available bandwidth using the acknowledgement feedback from the receiver. As RTT increases, TCP estimates available bandwidth less accurate and responds to congestion slower. However, with the previous knowledge of SLA, the channel capacity of an EVPL service is lower bounded by CIR. Comparing (2) and (3), it is clear that for EVPL services with a certain loss rate, the TCP throughput will deviate from the channel capacity.

5.4 SLA-Aware Transport Protocol Design

There are two approaches to improve transport throughput. One approach is to start with a new slate and design a brand new transport protocol. This method is very complex and may not match upper layer applications or lower layer infrastructure very well. In this research, we extend the window control mechanism of existing TCP. Our approach targets general TCP, not limited to specified TCP flavors.

The mismatch between the channel capacity and the throughput of TCP over EVPL is due to the lack of SLA information at the TCP sender. Using packet loss as the only congestion indicator, traditional TCP protocols cannot distinguish the reasons for packet losses. However, in EVPL services with SLA, it is reasonable to assume that the losses in CIR traffic (the high priority traffic) are due to random loss, and the losses in EIR traffic (the best effort traffic) are due to congestion. Because the minimum available bandwidth is CIR and the minimum RTT is the two-way propagation delay, a new state variable *Cwnd_Min* is proposed as: $Cwnd\_Min = 2 \cdot Delay_{propagation} \cdot CIR$, which is the amount of data that EVPL can handle with high priority in one RTT interval with traffic shaping.

The modification here is replacing the multiplicative decrease behavior of traditional TCP,

$$Cwnd = Cwnd / 2 \qquad (4)$$

by a shifted multiplicative decrease function:

$$(Cwnd - Cwnd\_Min) = (Cwnd - Cwnd\_Min)/2 \qquad (5)$$

Therefore, the congestion window *Cwnd* is lower bounded by *Cwnd_Min* in the congestion avoidance phase. Comparing (4) and (5), it is easy to find that the proposed SLA-aware transport protocol is a shifted version of the traditional TCP multiplicative decrease mechanism. In another words, the traditional TCP is a special case of the proposed protocol with *Cwnd_Min* equals zero, which means no QoS guarantee.

From (4) and (5), it is clear that the proposed protocol deliver Cwnd_Min plus a variable amount of data in one RTT. Assuming EIR>CIR and traffic shaping is used, the throughput of the proposed protocol can be expressed as:

$$B = CIR + \frac{\sqrt{3}}{RTT\sqrt{2bq}} \cdot Packet\_size \qquad (6)$$

in which q is defined as the ratio of the number of lost packets divided by the total number of transmitted best effort packets. Denoting the loss rate of CIR traffic by p_1 and the loss rate of best effort traffic by p_2, q can be expressed as:

$$q = \frac{CIR \cdot p_1 + (B - CIR) \cdot p_2}{B - CIR} \qquad (7)$$

Combining (6) and (7),

$$p_2 \cdot (B - CIR)^2 + CIR \cdot p_1 \cdot (B - CIR) - \frac{3 \cdot Packet\_size^2}{2 \cdot b \cdot RTT^2} = 0$$

B can be solved,

$$B = \left(1 - \frac{p_1}{2p_2}\right) CIR + \sqrt{\left(\frac{p_1}{2p_2}\right)^2 \cdot CIR^2 + \frac{3 \cdot Packet\_size^2}{2b \cdot p_2 \cdot RTT^2}} \qquad (8)$$

In a special case, when there is no QoS guarantee and CIR is 0, the throughput of the proposed protocol will be:

$$B = \sqrt{\frac{3 \cdot Packet\_size^2}{2b \cdot p_2 \cdot RTT^2}},$$

which is the same with the throughput of traditional TCP, as expected.

5.5 Simulation

5.5.1 Simulation Setup

OPNET simulations are carried to evaluate the performance of the proposed SLA-aware transport protocol over EVPL services. In the simulation, we implement our approach based on TCP_Sack. The simulation setup is shown in Figure29.

A 200 Mbyte file is delivered from the sender to the receiver using FTP at the application layer through an EVPL across a WAN. The traffic controller is an OPNET node module programmed to simulate EVPL service behavior, as shown in Figure 30, including traffic policing/shaping, loss events generation, and delay/jitter insertion. The end-to-end throughput at application layer is calculated as *file_size*/*FTP_time*. When the packet loss rate is small and retransmission is rare, the difference between application layer throughput and link layer throughput is marginal. Traffic shaping using a dual-token bucket [49] is used in addition to policing, so that the traffic is smoothed with less packet loss. RTT is modeled by propagation delay plus a small random delay jitter. EIR is set to be twice as large as CIR. The EVPL simulation parameters are shown in Table 6, with QoS parameters set to the bronze service class in [50]. Here, high priority CIR traffic enjoys a lower loss rate of 0.1%. Best-effort traffic suffers a higher loss rate of 1%.

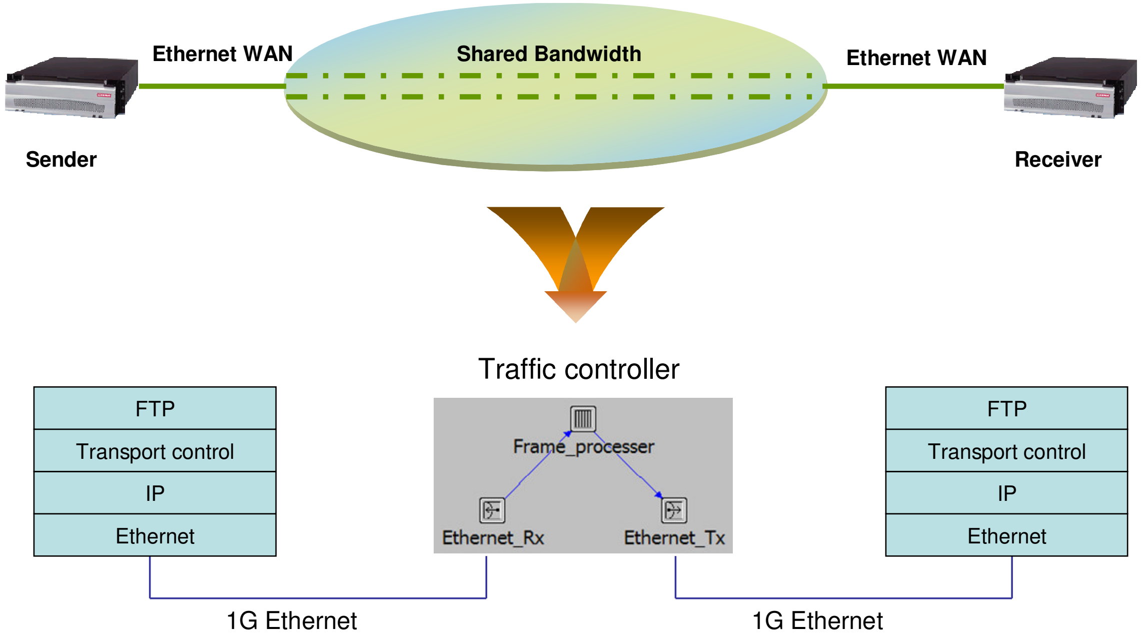


Figure 29: Simulation setup for EVPL services.

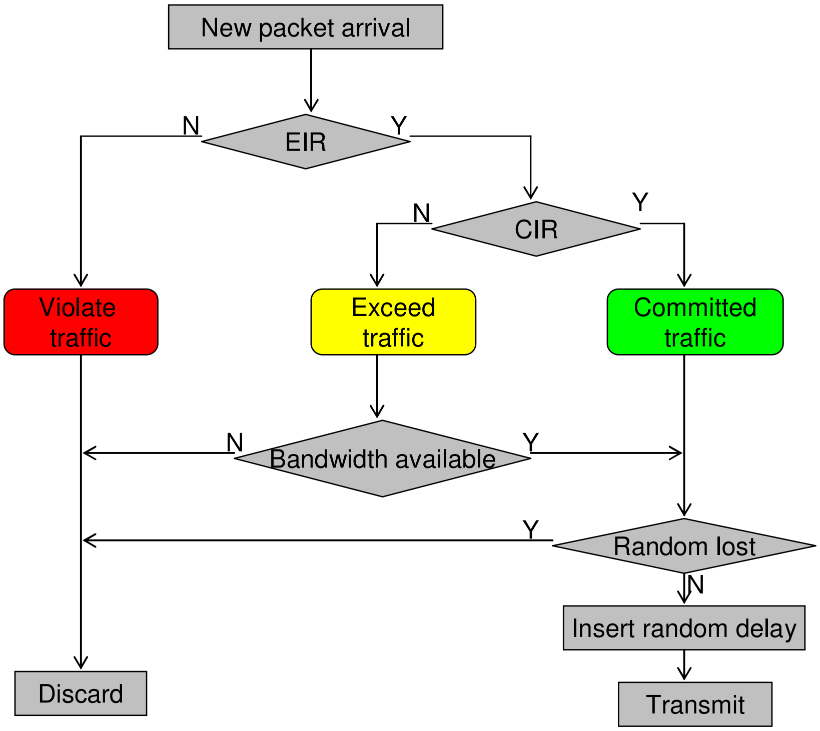


Figure 30: Functions of the traffic controller module.

Table 6: Simulation parameters for SLA-aware transport protocol.

CIR	100Mbps
CBS	25Kbytes
CIR traffic loss rate	0.1%
EIR	200Mbps
EBS	50Kbytes
Best-effort traffic loss rate	1%
Round-trip propagation delay	[1, 10] ms
Jitter	0.1ms
Traffic shaping buffer size	125 Kbytes
Receiver buffer size	250Kbytes
Packet size	1.5Kbytes

5.5.2 Simulation Results

The throughput of the proposed protocol is shown in Figure 31. The values of the simulation results are slightly lower than the analytical results, because the analytical throughput includes retransmissions and is for the congestion avoidance phase only, whereas the simulation throughput is pure goodput and is for the whole FTP session, including both slow start and congestion avoidance phases.

For comparison, the throughput of TCP_Sack in the same network configuration is shown in Figure 32. Because the loss rate for high priority traffic is 0.1% and the loss rate for low priority (best-effort) traffic is 1%, the average packet loss rate is between

0.1% and 1%. The simulation results are lower bounded by the analytical results with a loss rate of 1% and upper bounded by the analytical results with a loss rate of 0.1% in (3). For smaller RTT, the throughput is higher and more packets are delivered in a best-effort manner, so the simulation results are closer to the analytical curve with a loss rate of 1%. For larger RTT, the throughput is smaller and more packets fall into the CIR profile. The simulation results are closer to the analytical curve with a loss rate of 0.1%.

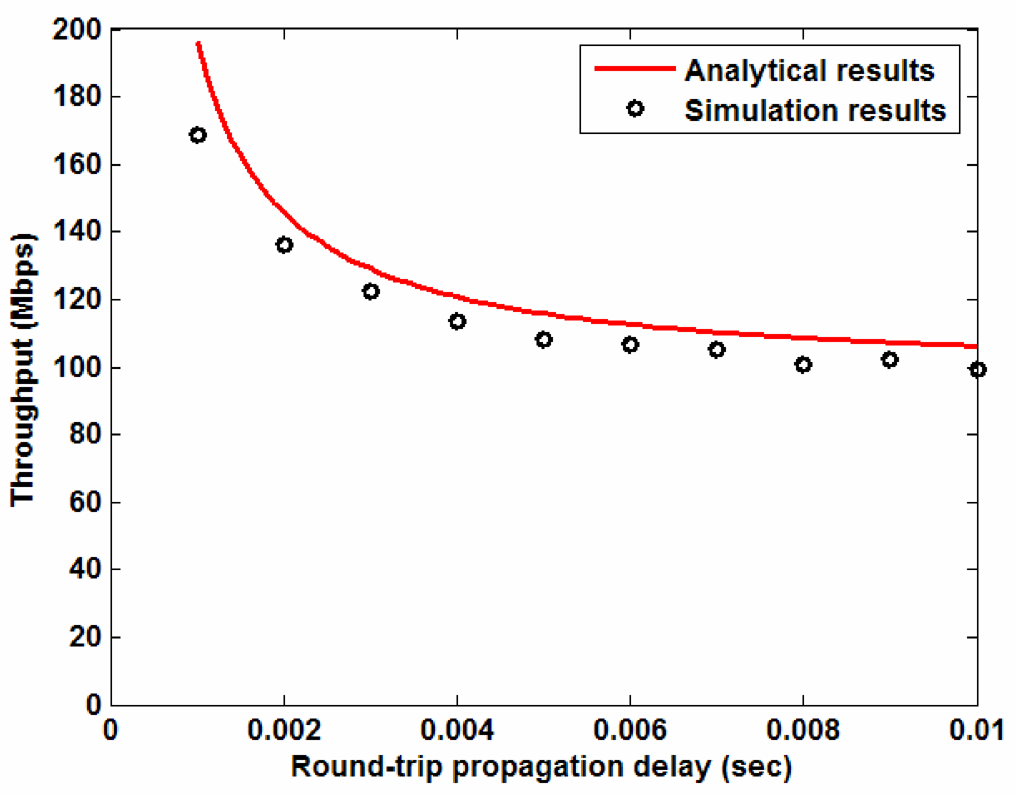


Figure 31: Throughput of the SLA-aware transport protocol.

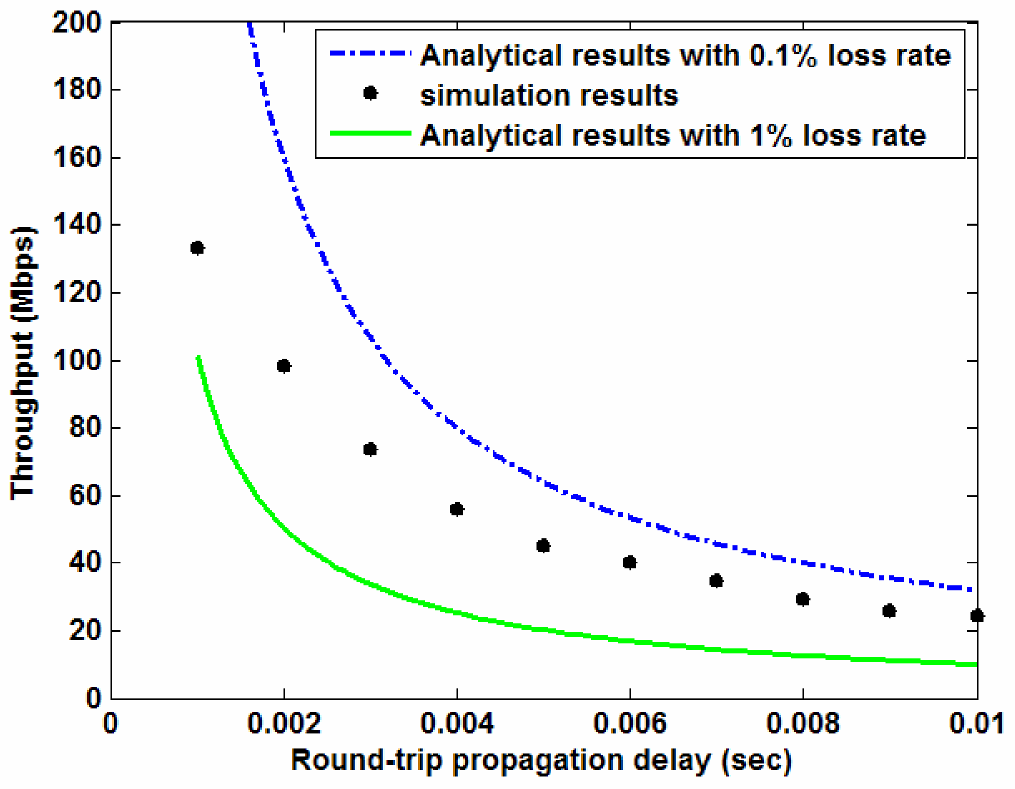


Figure 32: Throughput of TCP_Sack.

To compare the congestion window control behaviors, a snapshot of the *Cwnd* state variable trace is shown in Figure 33 for the proposed protocol, and in Figure 34 for traditional TCP_Sack, in the case that RTT is 4ms (*Cwnd_Min* is 50Kbytes). It is clear that the proposed protocol has a larger congestion window size than traditional TCP. A larger congestion window leads to higher throughput for a window-based control protocol like TCP. Because the proposed protocol has larger throughput, it suffers more packet loss events per unit time. However, the proposed protocol reduces its congestion window less severely than tradition TCP, and its congestion window size is always larger than *Cwnd_Min* in the congestion avoidance phase.

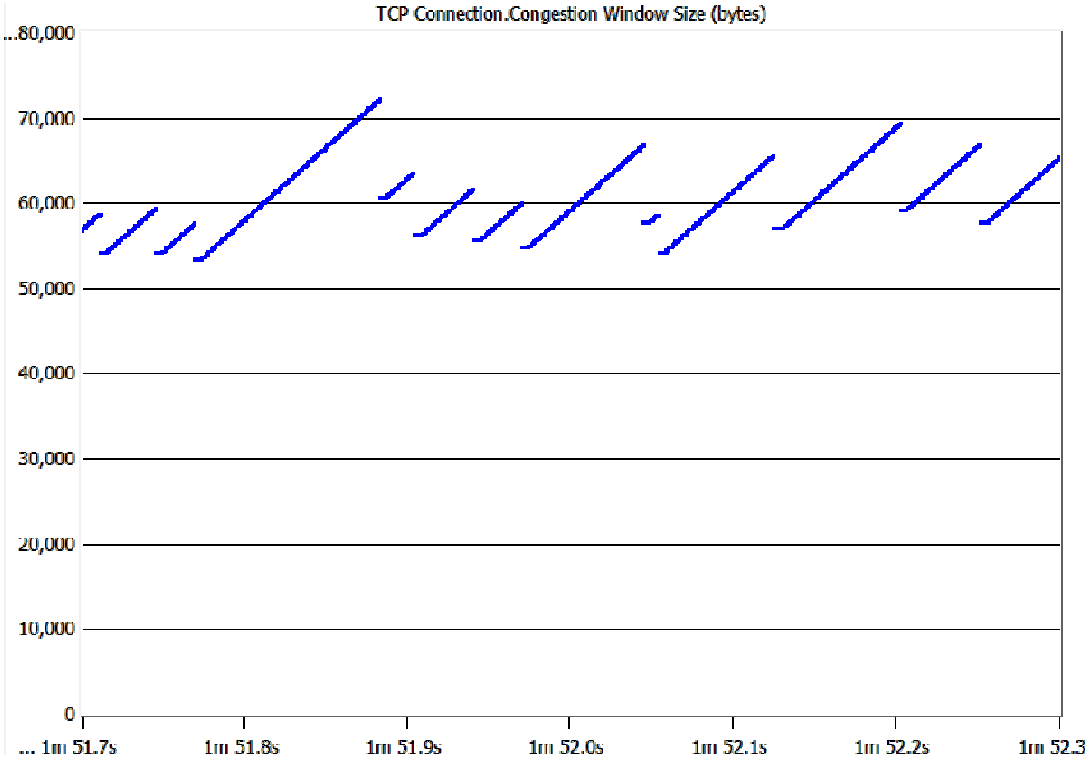


Figure 33: Congestion window trace of SLA-aware transport protocol.

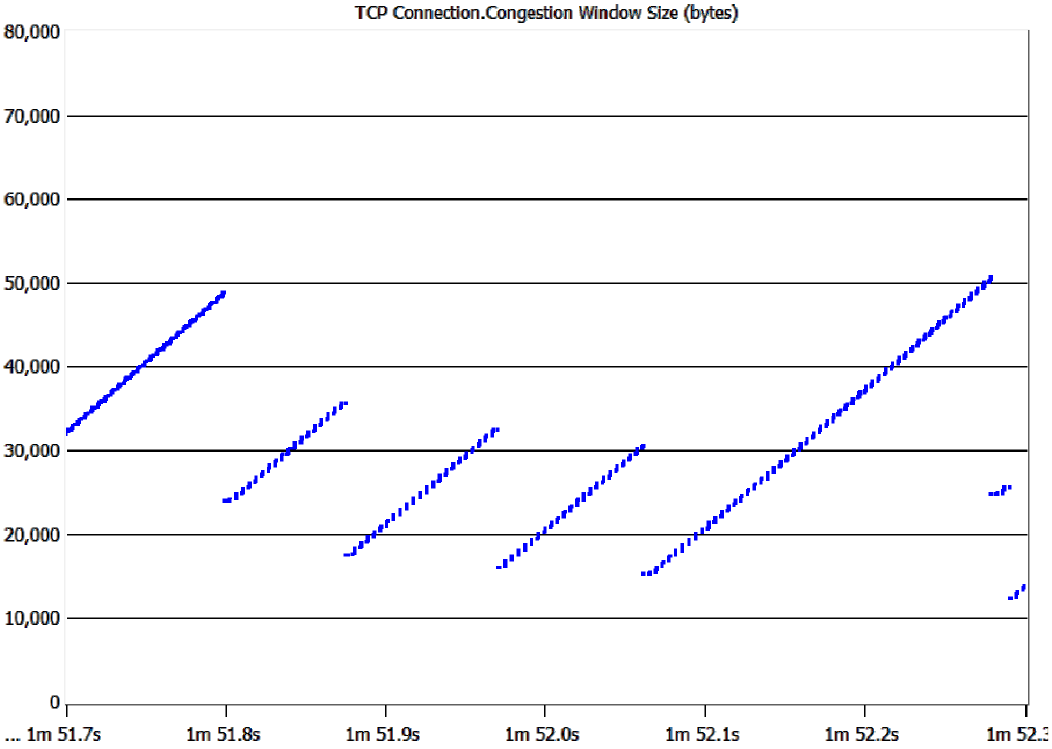


Figure 34: Congestion window trace of TCP-Sack.

CHAPTER 6

A RESILIENT TRANSPORT SCHEME FOR ETHERNET SERVICES

6.1 Reasons of Packet Loss in Ethernet Services

Congestion and random loss are two reasons of packet loss in Ethernet services. Although QoS mechanisms at link layer can allocate dedicated bandwidth for each user and keep loss rate small enough as the requirement in SLA, because of the stochastic nature of Ethernet networks, a small but non-zero random loss rate is inevitable. This random loss can be tolerated by Ethernet services, but not by TCP. Traditional TCP was designed for best-effort networks, assumed all packet losses are because of congestion. Therefore, TCP cannot utilize available bandwidth efficiently in a broadband link when there is no congestion but rare random loss present.

In the previous chapter, it was shown that traditional TCP has a poor performance in high speed Ethernet services environments. It is proven that the channel capacity of Ethernet services is lower bounded by the reserved bandwidth based on an erasure channel model, and the traditional TCP is not able to approach that bound. An SLA-aware transport protocol was proposed using a shifted AIMD algorithm [10]. Although that protocol can fully utilize the reserved bandwidth and achieve much better throughput performance than conventional TCP, it still cannot utilize the available bandwidth very efficiently.

In this chapter, a new channel model is proposed for Ethernet services by considering both SLA and available bandwidth information. As shown in Figure 35, the traffic from the sender is policed according to the SLA at the egress point. The packet that falls into the CIR profile will be marked with high priority. A packet that falls out of the CIR profile, but still in the EIR profile, will be marked with low priority and delivered in a best-effort manner. A packet that falls out of the EIR profile will be discarded immediately. Transport layer congestion control is expressed as a hypothesis test process making decisions through the ACK feedbacks.

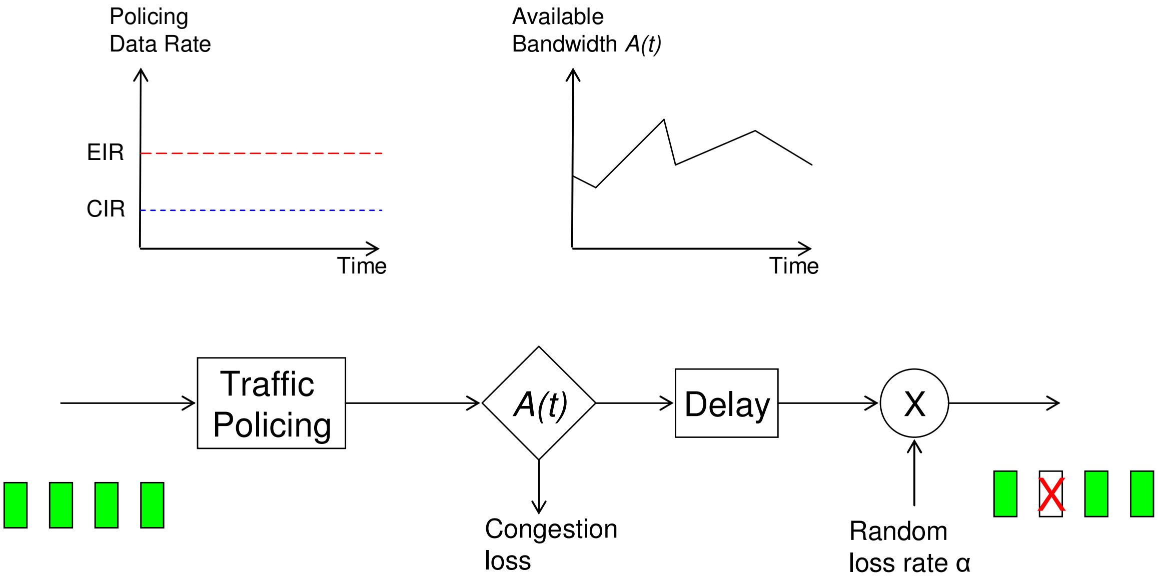


Figure 35: Channel model of EVPL with available bandwidth and random loss.

Within the transport network, three parameters are used to describe the network condition: time-varying available bandwidth *A(t)*, random packet loss rate α, and a time delay. Due to the bursty nature of Internet traffic, the instant value of background traffic depends on the time scale [51]. In this research, we assume the value of available bandwidth *A(t)* is calculated with a time granularity comparable with or larger than RTT, and *A(t)* should be larger than CIR. Packet losses in the transport network are caused by two reasons: congestion and random loss. We define the congestion suffered by a flow as the status that the available bandwidth *A(t)* is smaller than the sending rate. We define random loss as the packet loss due to physical layer random error or transient burst of background traffic (in the time scale smaller than RTT).

According to the erasure channel model in Figure 35, where packets will be either delivered successfully or discarded, the average channel capacity of EVPL can be expressed as

$$\begin{aligned} C &= \frac{1}{T}\int_0^T \min(A(t), EIR)dt \cdot (1-\alpha) \\ &= CIR \cdot (1-\alpha) + \frac{1}{T}\int_0^T [\min(A(t), EIR) - CIR]dt \cdot (1-\alpha) \end{aligned}$$

Given that $A(t) \geq CIR$ and $EIR \geq CIR$, the same lower bound of EVPL channel capacity can be derived as

$$C = CIR \cdot (1-\alpha)$$

that is the same with (2) in the previous chapter.

6.2 Adaptive AIMD Algorithms

In this dissertation, we want to distinguish the causes of packet loss and adjust the congestion window size adaptively. Several protocols have been proposed to improve the throughput in large bandwidth environments by modifying the AIMD behavior dynamically.

HighSpeed TCP, a modification to TCP's congestion control mechanism for use with connections with large congestion windows, is proposed in [15]. As an evolutional alternative to traditional TCP, HighSpeed TCP behaves identical to conventional TCP when the congestion window size is small. For the case that the congestion window size is large, HighSpeed TCP increases its congestion window faster when there is no packet loss, and decreases its congestion window less aggressively when there is a packet loss, comparing with traditional TCP.

Instead of changing AIMD behavior dynamically based on the current congestion window size, an adaptive AIMD congestion control algorithm is proposed based on whether there is a packet loss in the previous RTT period [52]. For each round of RTT, the additive increase parameter will be increased cumulatively and the multiplicative decrease parameter will be reset to the default value in the case of no packet loss present in the previous RTT. On the other hand, the multiplicative decrease parameter will be increased cumulatively and the additive increase parameter will be reset to the default value in the case of some packet losses present in the previous RTT. Although this algorithm is adaptive in heuristic, it cannot estimate the loss rate value, because the sender does not count the number of successfully delivered packets and the number of lost packets.

In this chapter, a stochastic model is proposed to capture the random nature of the network environments. For each packet loss indicated by triple duplicate ACKs, a hypothesis test is carried to estimate the reason of the packet loss, due to congestion or random factors. We limit this research by utilizing packet loss information only, without any queuing model for the network or any additional function in the router.

6.3 Transport Control for Ethernet Services Based on Hypothesis Test

TCP was designed for best-effort packet switching networks without previous knowledge about available bandwidth. The TCP sender has to probe the available bandwidth using the acknowledgement feedback from the receiver. The congestion control and error control functions of TCP are bonded together, because it is difficult to distinguish triple duplicate ACKs because of congestion from triple duplicate ACKs because of random loss in a best-effort network. However, in the Ethernet services, the QoS parameters in SLA offer valuable information. The guaranteed loss rate for CIR traffic can be used as a threshold to distinguish network conditions, between the state of congestion and the state of non-congestion. However, the network status is an inherent property inside the network, which is not directly known by the TCP sender. We model the congestion control algorithm as a system with a binary feedback indicating congestion and random loss.

6.3.1 A Stochastic Model for TCP Loss Events in Ethernet Services

We model the network conditions with a state variable of available bandwidth *A(t)*, a constant of random loss rate α, and the round trip delay *RTT*. TCP throughput is a function of the network conditions. Because the TCP sender is located outside the network, the network is a black-box from the perspective of the TCP sender, and the TCP sender cannot observe the internal states of the network directly in real time. The TCP sender can only probe the network conditions by sending data packets and waiting for ACK feedbacks after a time interval of *RTT*, as shown in Figure 36. The measurement results of TCP probing can be expressed as a binary sequence $L = [L_1, L_2, L_3, \dots , L_\mathrm{j}]$, in which $L_j = 0$ if the *j*'th packet is acknowledged successfully, and $L_j = 1$ if the *j*'th packet is lost (indicated by either timer-out or triple duplicate ACKs). The binary sequence *L* is a measurement result at the TCP sender, and is a stochastic function of the available bandwidth *A(t)* and the random loss rate α. In this paper, α is assumed to be equal to the guaranteed loss rate in SLA, which is used as a model parameter to derive the threshold to differentiate congestion from non-congestion network status.

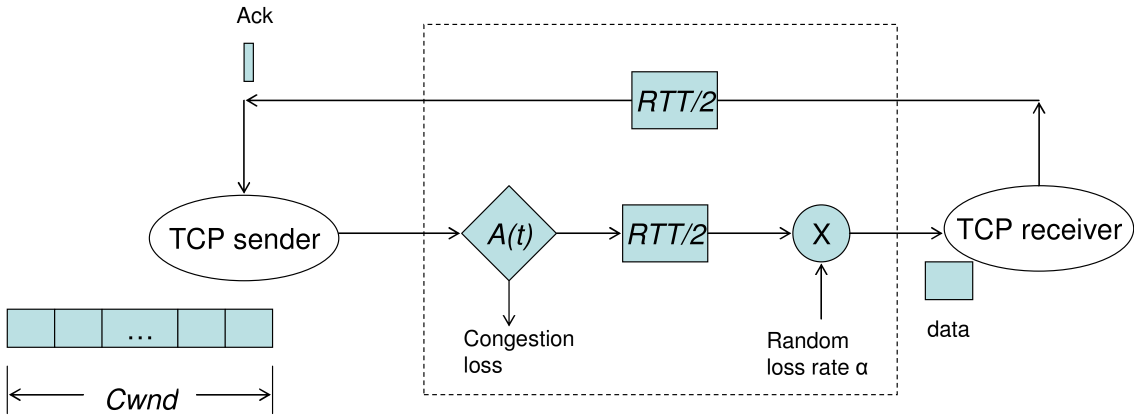


Figure 36: A stochastic model for TCP congestion detection.

For Ethernet services, QoS is provided in layer 2 as follows: for CIR traffic, the loss rate is guaranteed to be less than or equal to α; for EIR traffic, the loss rate could be any value between 0 and 1, depending on network conditions. In transport layer, *P(L$_j$)*, The loss probability of the *j*'th TCP segment, can be expressed as:

$$P(L_j = 1) = \left\langle \begin{array}{ll} \dfrac{Cwnd / RTT - A(t)}{Cwnd / RTT} + \alpha, & A(t) < \dfrac{Cwnd}{RTT} \\ \alpha, & A(t) \geq \dfrac{Cwnd}{RTT} \end{array} \right. \qquad (9)$$

The response function of the conventional TCP to a packet loss, for example, the measurement result of $L_j = 1$, is:

$$Cwnd = Cwnd / 2 \, .$$

In Ethernet services, we model the network status with two states: congestion and non-congestion. In the state of congestion, available bandwidth *A(t)* is smaller than the average sending rate, *Cwnd / RTT*, and the packet loss probability is larger than α. Whereas in the state of non-congestion, available bandwidth *A(t)* is larger than or equal to the average sending rate, and packet loss probability is smaller than or equal to α.

6.3.2 A Simple Hypothesis Test Adopted by Traditional TCP

In mathematical statistics, approaches are available to address the problem of making a definite decision with respect to an uncertain hypothesis which is known only through its observable consequences. A statistical hypothesis test, or more briefly, hypothesis test, is an algorithm to state the alternative (for or against the hypothesis) that minimizes certain risks [53]. The mapping from observable measurements to the decision of accepting/rejecting a certain hypothesis is called the decision rule. In this research, we define the hypothesis as follows:

$$\begin{aligned} &H_0: \quad A(t) \geq \frac{Cwnd}{RTT}, \quad non-congestion \\ &H_1: \quad A(t) < \frac{Cwnd}{RTT}, \quad congestion \end{aligned} \tag{10}$$

Combining (9) and (10), we derive the loss probabilities conditioned on different hypothesis:

$$\begin{aligned} &P(L_j = 1 \mid H_0) = \alpha \\ &P(L_j = 0 \mid H_0) = (1-\alpha) \\ &P(L_j = 1 \mid H_1) = \frac{Cwnd / RTT - A(t)}{Cwnd / RTT} + \alpha \\ &P(L_j = 0 \mid H_1) = \frac{A(t)}{Cwnd / RTT} - \alpha \end{aligned} \tag{11}$$

Traditional TCP adopts a very simple decision rule which is only based on a single measurement result. For the measurement that the *j*'th packet is acknowledged successfully ($L_j = 0$), since $P(L_j = 0|H_0)$ is larger than $P(L_j = 0|H_1)$, H_0 is accepted by TCP.

For the event that there are triple duplicate ACKs for the *j*'th packet ($L_j = 1$), since $P(L_j = 1|H_1)$ is larger than $P(L_j = 1|H_0)$, H_1 is accepted by TCP. As a result, TCP will deduce that congestion happened and will halve its congestion window size.

In detection theory, false-alarm probability P_F and detection probability P_D are the two metrics to evaluate the performance of a hypothesis test. False-alarm probability P_F is defined as the probability that the hypothesis H_1 is declared to be in effect when H_0 is actually in effect. Detection probability P_D is defined as the probability that the hypothesis H_1 is declared to be in effect when H_1 is actually in effect. As shown in Table 7, in this dissertation, P_F is the probability that the transport protocol declares congestion by mistake, when there is no congestion in the network; P_D is the probability that the transport protocol declares the congestion presence correctly, when there is congestion in the network.

Table 7: Probabilities related with congestion detection.

	Definition in detection theory	Meaning in congestion control
P_F	the probability that we declare hypothesis H_1 to be in effect when H_0 is actually in effect	False alarm of congestion detected when there is no congestion present in the network
1-P_F	the probability that we declare hypothesis H_0 to be in effect when H_0 is actually in effect	Correct detection when there is no congestion present in the network
P_D	the probability that we declare hypothesis H_1 to be in effect when H_1 is actually in effect	Correct detection when there is congestion present in the network
1-P_D	the probability that we declare hypothesis H_0 to be in effect when H_1 is actually in effect	Miss probability that the congestion present in the network is overlooked

In practice, it is desirable to find a decision rule that can keep P_F as small as possible and increase P_D as large as possible. Unfortunately, P_F and P_D are correlated in a way that P_F and P_D are either increased or decreased together when we adjust the decision rule. Depending on the optimization criteria, a certain trade-off between P_F and P_D has to be made. Traditional TCP has a very rigid decision rule: given the current event of $L_j = 1$ (the *j*'th packet is lost), congestion is declared for sure. This rule leads to $P_F = P_D = 1$, which means that both false-alarm and detection probability are maximized.

As a result, TCP decreases its congestion window aggressively for each packet loss, and the congestion window size oscillates severely during a TCP session. As shown in [15], conventional TCP cannot utilize available bandwidth efficiently in a broadband link when there is no congestion but rare random losses.

6.3.3 A Hypothesis Test to Distinguish Congestion from Random loss in Ethernet Services

Traditional TCP adopts a very simple decision rule by making decision based on only a single measurement sample, which is inaccurate in networks with stochastic nature. We propose a novel hypothesis test with a decision rule using multiple recent measurements. For the current measurement of $L_j = 1$ (get triple duplicate ACKs for the *j*'th packet), a new decision variable *X* is defined as the number of recent packets delivered continuously and successfully before the *j*'th packet is lost. Denote *p* as the loss probability present in the network, each packet delivery is modeled as a Bernoulli trial and *X* is a Geometric distributed random variable:

$$P(X = n, p) = p \cdot (1 - p)^n \tag{12}$$

where *p* is the loss probability described in (9). A threshold γ is chosen to distinguish congestion and random (non-congestion) loss. A general decision rule is: congestion will be declared if *X* is smaller than γ; non-congestion will be declared if *X* is larger than γ. Combining (9) and (12), we derive the probabilities conditioned on different hypothesis:

$$
\begin{aligned}
&P(X \le \gamma \mid H_0) = 1-(1-\alpha)^{\gamma}\\
&P(X > \gamma \mid H_0) = (1-\alpha)^{\gamma}\\
&P(X \le \gamma \mid H_1) = 1-\left(\frac{A(t)}{Cwnd\ / RTT}-\alpha\right)^{\gamma}\\
&P(X > \gamma \mid H_1) = \left(\frac{A(t)}{Cwnd\ / RTT}-\alpha\right)^{\gamma}
\end{aligned}
\tag{13}
$$

From the definitions of false-alarm probability P_F and detection probability P_D, under the condition that the current packet is lost, we have

$$
\begin{aligned}
&P_F = P(X \le \gamma \mid H_0) = 1-(1-\alpha)^{\gamma}\\
&P_D = P(X \le \gamma \mid H_1) = 1-\left(\frac{A(t)}{Cwnd / RTT}-\alpha\right)^{\gamma}
\end{aligned}
\tag{14}
$$

The comparison of traditional TCP decision rule and the new decision rule is shown in Table 8:

Table 8: Performance comparison of two decision rules.

	Decision rule of traditional TCP	The new decision rule
P_F	1	$1-(1-\alpha)^{\gamma}$
P_D	1	$1-(\frac{A(t)}{Cwnd / RTT}-\alpha)^{\gamma}$

The decision rule of traditional TCP is a special case of the proposed decision rule, when γ equals to ∞ (which means congestion always be declaimed after a packet loss). By using the proposed decision rule instead of the simple rule adopted by conventional TCP, we can make a trade-off between false-alarm probability P_F and detection probability P_D to optimize the throughput performance. Moreover, because available bandwidth $A(t)$ is smaller than $Cwnd / RTT$ in the presence of congestion, the improvement of P_F is larger than the degrading of P_D as shown in Table 8.

To alleviate the degrading of P_D, moderate congestion control can be carried for the declaiming of non-congestion. For example, the congestion window can be reduced to 75%, instead of 50%, in the shifted AIMD. The heuristic of a new response function for packet loss is show in Figure 37. Because moderate congestion control is used instead of aggressive congestion control, when non-congestion is declared for a packet loss, the congestion window oscillates much less than conventional TCP. Therefore, the proposed scheme can tolerate random losses in Ethernet service and is able to achieve better throughput performance than conventional TCP.

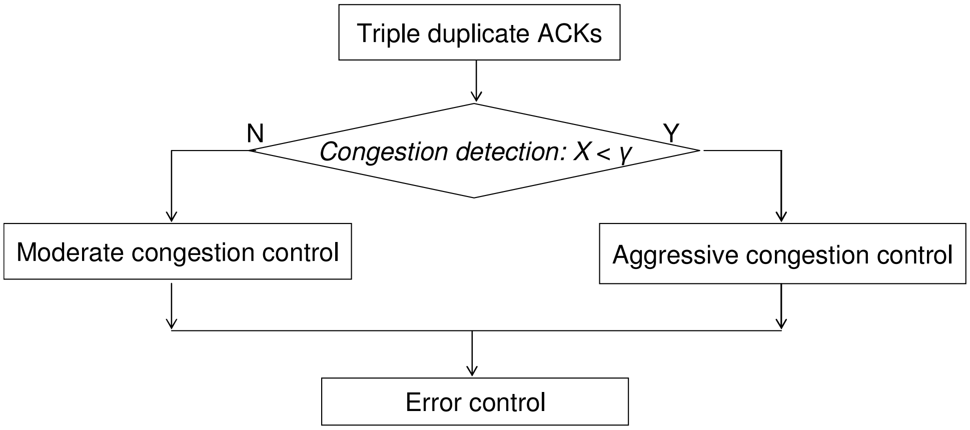

Triple duplicate ACKs
N
Y
Congestion detection: X < γ
Moderate congestion control
Aggressive congestion control
Error control

Figure 37: Heuristic of the proposed response function.

6.4 Simulation Results

Simulations are carried using OPNET to evaluate the performance of the proposed transport scheme over EVPL services. In the simulation, we implement our approach based on TCP-Sack. The simulation setup is shown in Figure 38.

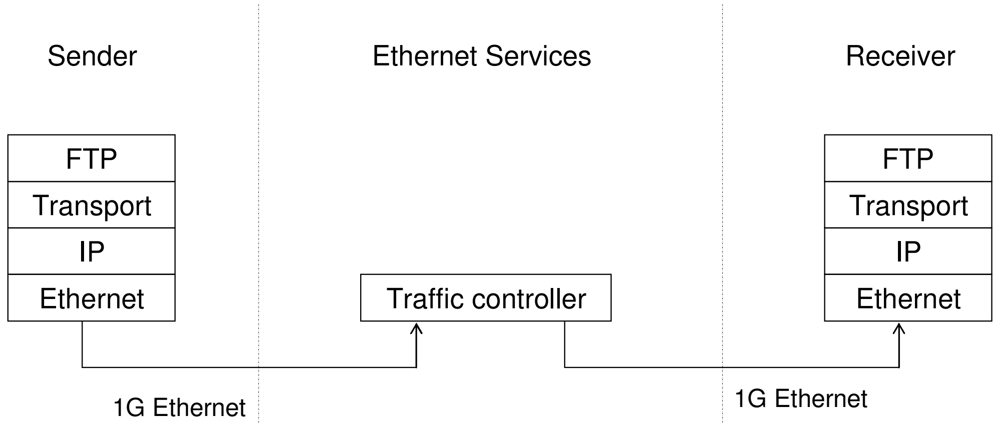


Figure 38: Simulation setup for EVPL services.

A 400 MB file is delivered from the sender to the receiver using FTP at the application layer through an EVPL connection. The traffic controller is an OPNET node module programmed to simulate EVPL service behavior, including traffic policing/shaping, loss events generation, and delay/jitter insertion. The end-to-end throughput at the application layer is calculated as *file_size / FTP_time*. RTT is modeled by propagation delay plus a small random delay jitter. EIR is set to be twice as large as CIR. The EVPL simulation parameters are shown in Table 9, with QoS parameters set to the bronze service class in [50]. The packet loss rate of EIR traffic is periodically time-varying, as shown in Figure 39. During the non-congestion periods (4 second per period), EIR traffic enjoys a low loss rate of 0.1%, the same with CIR traffic. During the congestion periods (1 second per period), EIR traffic suffers a high loss rate of 10%.

Table 9: Simulation parameters for hypothesis test based transport control.

CIR	100 Mbps
CBS	25 kB
CIR traffic loss rate	0.1%
EIR	200 Mbps
EBS	50 kB
Best-effort traffic loss rate	See Figure 39
Round-trip propagation delay	[1, 10] ms
Jitter	0.1 ms
Traffic shaping buffer size	125 kB
Receiver buffer size	250 kB
Packet size	1.5 kB

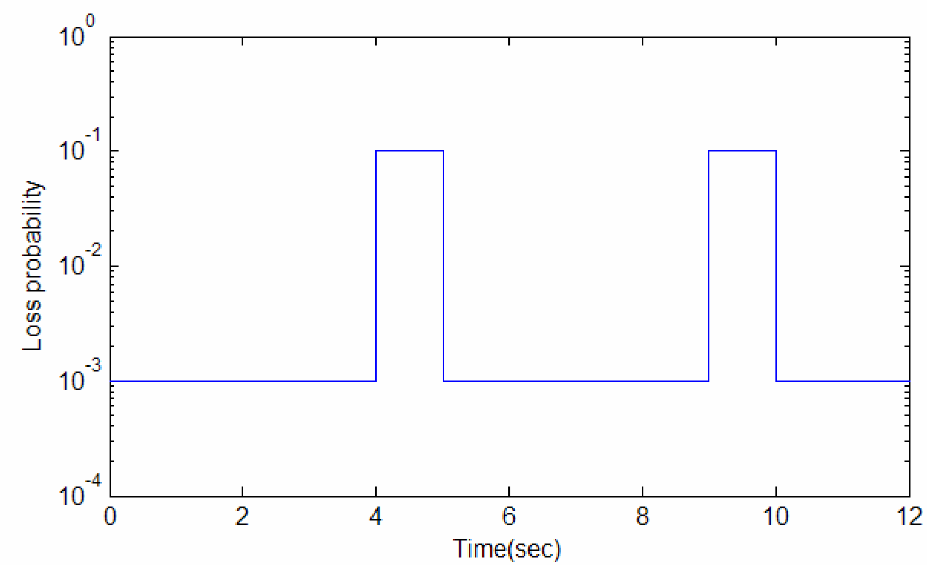


Figure 39: Loss rate pattern for EIR traffic.

The threshold γ is calculated dynamically to satisfy the following condition: the probability without congestion in (12) (p = 0.001) equals the probability with light congestion in (12) (the product of available bandwidth *A(t)* and *RTT* is one segment less than congestion window *Cwnd*):

$$P(X = \gamma, 0.001) = P(X = \gamma, \frac{1}{Cwnd / Packet\_size} + 0.001)$$

The throughputs of the proposed scheme based on the hypothesis test, the SLA-aware transport protocol in the previous chapter, and the traditional TCP-Sack are shown in Figure 40.

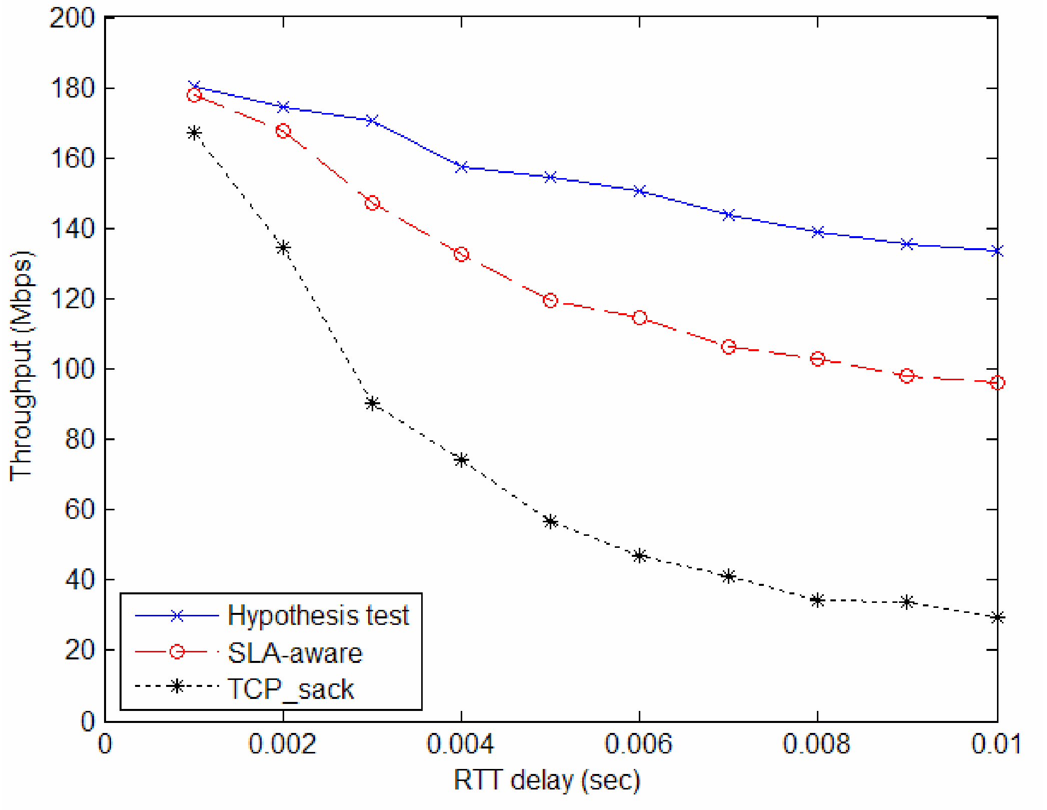


Figure 40: Throughput comparison.

The proposed transport scheme based on a hypothesis test can increase throughput significantly, because the congestion window size is adjusted adaptively based on the test result. When there is a packet loss and the current loss rate is small, random loss is declaimed, and the congestion window will be reduced moderately, which helps to improve end to end throughput. Congestion is declaimed when there is a packet loss and the current loss rate is high. The congestion window will be reduced by half, which is back compatible with TCP.

CHAPTER 7

CONCLUSIONS

In this dissertation, we first studied the media access control in broadband optical access. We then explored the delta compression algorithm that can improve content download performance. We investigated the transport control protocol for Ethernet services. Two novel approaches: an SLA-aware transport protocol and a resilient transport scheme based on hypothesis test are proposed.

7.1 Contributions

Primary contributions of this dissertation are summarized here:

- Designed an efficient MAC protocol with pre-allocation for WDM PONs. The proposed MAC protocol supports prioritized traffic flows and scalable bandwidth granularity by combing WDM and TDM. A pre-allocation scheme is proposed to reduce access delay.
- Evaluated the performance of the proposed MAC protocol in details through both queuing model analysis and OPNET simulation.
- Analyzed and implemented a delta compression scheme for fast content download. For the first time, a theoretical analysis for delta compression algorithms is carried based on information theory. Implementation and measurement results are presented to verify the efficiency of delta compression in real network environments.

- Proposed, analyzed, and simulated an SLA-aware transport protocol for Ethernet services. It is shown that the proposed protocol outperforms traditional TCP in the terms of throughput. Traditional TCP can be regarded as a special case of the proposed protocol without QoS information.
- Proposed a resilient congestion control scheme for broadband metro Ethernet services. The congestion window is controlled adaptively with current congestion status, which is estimated by a hypothesis test. It is demonstrated that the proposed scheme can increase the end-to-end throughput significantly.

In addition, I have investigated DOCSIS 2.0 protocol for hybrid fiber coax (HFC) networks in details as a reference to WDM PON research. A test bed is set up based on an Arris DOCSIS 2.0 CMTS, 10 Arris DOCSIS 2.0 cable modems, and a Spirent SmartBits 6000 traffic generator. The end-to-end delay and throughput performance of DOCSIS 2.0 is measured by feeding Ethernet traffic flows through the DOCSIS 2.0 cable network with different priorities. Peer-to-peer file sharing and video streaming applications are also investigated based on this test bed.

Following is the list of publications related to the work presented in this dissertation.

1. Chunpeng Xiao, Claudio Estevez, Benny Bing, Georgios Ellinas, and Gee-Kung Chang, “A Resilient Transport Control Scheme for Metro Ethernet Services Based on Hypothesis Test,”, *IEEE ICC 2007*, submitted.

2. Chunpeng Xiao, G.K. Chang, Benny Bing, “An SLA-aware Transport Protocol for High Throughput Wide Area Ethernet Services,” *IEEE GLOBECOM 2006*, to appear

3. Chunpeng Xiao, Claudio Estevez, G.K. Chang, “Performance Evaluation of an SLA-Aware Transport Control Protocol for Ethernet Services,” *OPNETWORK 2006*, Washington DC, Aug. 28-Sep. 1, 2006

4. Claudio Estevez, Chunpeng Xiao, G.K. Chang, “Simulation Study of TCP Acceleration Mechanisms for Broadband Access Networks,” *OPNETWORK 2006*, Washington DC, Aug. 28-Sep. 1, 2006

5. Chunpeng Xiao, Benny Bing and G. K. Chang, “Delta Compression for Fast Wireless Internet Download,” *IEEE Globecom 2005*, St. Louis, MO, Nov. 28-Dec. 2, 2005

6. Chunpeng Xiao, Benny Bing and G. K. Chang, “An Efficient Reservation MAC Protocol with Preallocation for High-Speed WDM Passive Optical Networks,” *IEEE INFOCOM 2005*, Miami, FL, March 13-17, 2005

7. Chunpeng Xiao, “Next Generation Broadband Access Networks for Integrated Services,” *IEEE INFOCOM 2005 student workshop*, Miami, FL, March 13-17, 2005

8. Hua Qian, Chunpeng Xiao, Ning Chen, and G. Tong Zhou, "Dynamic Selected Mapping for OFDM," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2005)*, Philadelphia, PA, March 19-23, 2005

9. Chunpeng Xiao and Benny Bing, “Delta Compression with Fixed-Length Substring Coding for Fast Content Download*,” IEEE Communication Letters*, vol. 9, no. 3, pp. 243- 245, March 2005

10. Chunpeng Xiao and Benny Bing, "Measured QoS Performance of the DOCSIS Hybrid-Fiber Coax Cable Network," *IEEE Workshop on Local and Metropolitan Area Networks*, pp. 23-27, San Francisco, CA, April 25-28, 2004

11. Chunpeng Xiao, Raviv Raich, and G. Tong Zhou, "QoS Constrained Statistical Resource Reservation for Wireless Networks," *IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 1713-1717, Pacific Grove, CA, Nov. 7-10, 2003

7.2 Suggestions for Future Research

This dissertation can be extended in a number of directions, including:

- Designing a MAC protocol to achieve end-to-end QoS for wireless/PON hybrid access networks, where a wireless access point is integrated in an ONU. Wireless is used in local area and PON is used in the first mile.
- Designing a WDM PON test bed with MAC control logic to deliver integrated applications, including voice, video, and data.
- Investigating the TCP acceleration technology that can improve TCP throughput using a proxy at center office.

- Combining the SLA-aware and hypothesis test based approaches proposed in this dissertation with other broadband TCP solutions, to further improve transport efficiency and fairness in Ethernet services

APPENDIX A: QUEUING ANALYSIS OF THE PROPOSED PON MAC PROTOCOL

We wish to analyze the upstream connection from 10 ONUs to 1 OLT in a logical bus architecture (Figure 41). At each ONU, the traffic arrival is Poisson and an independent and identically distributed (i.i.d.) Poisson. Each arrival can be either single packet or a batch of packets. We define T_f = TDM frame time and T_s = slot time. The access delay due to the request and grand mechanism of ONU 2 is shown in Figure42.

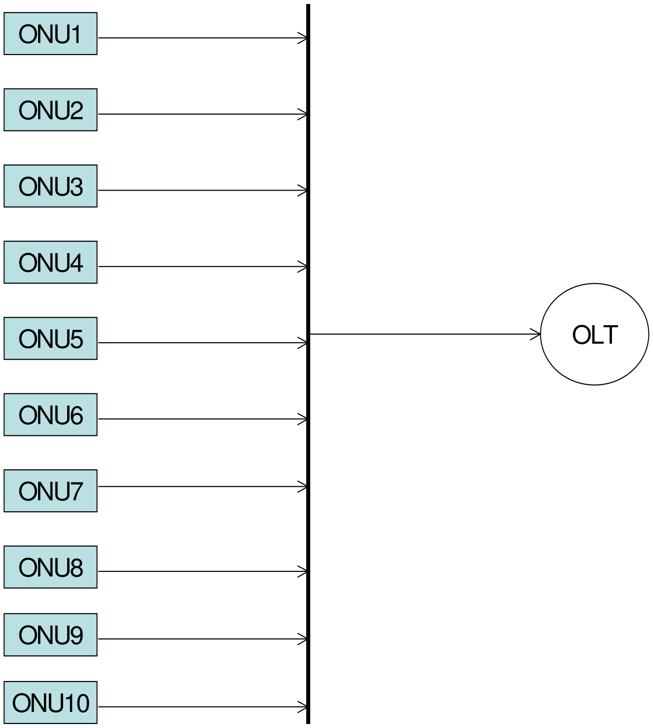
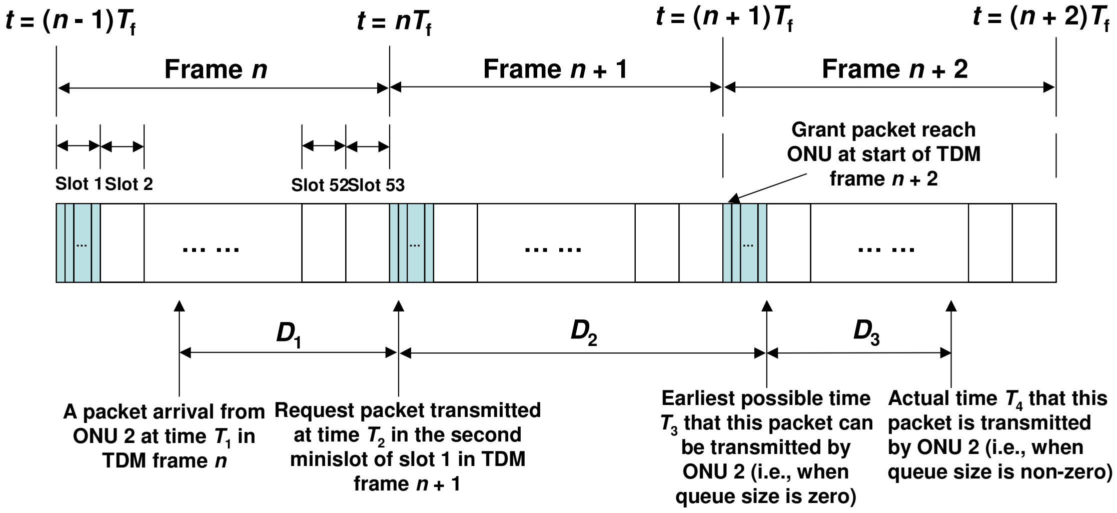


Figure 41: PON upstream channel model.



$t = (n - 1)T_f$
$t = nT_f$
$t = (n + 1)T_f$
$t = (n + 2)T_f$
Frame n
Frame $n + 1$
Frame $n + 2$
Slot 1
Slot 2
Slot 52
Slot 53
Grant packet reach ONU at start of TDM frame $n + 2$
… …
D_1
D_2
D_3
A packet arrival from ONU 2 at time T_1 in TDM frame n
Request packet transmitted at time T_2 in the second minislot of slot 1 in TDM frame $n + 1$
Earliest possible time T_3 that this packet can be transmitted by ONU 2 (i.e., when queue size is zero)
Actual time T_4 that this packet is transmitted by ONU 2 (i.e., when queue size is non-zero)

Figure 42: Access delay for ONU 2.

Without propagation delay, the average end-to-end delay can be viewed as the sum of the access delay ($D_1 + D_2 + D_3$) and the packet transmission time (D_4):

$$\overline{D} = \overline{D_1} + \overline{D_2} + \overline{D_3} + \overline{D_4} \quad \text{(A.1)}$$

where

D_1 = Waiting time for request slot (i.e., the time between the arrival of a packet and transmission of a request packet).

D_2 = Interval between transmission of request packet (from ONU) and the first data slot in the following TDM frame.

D_3 = Queuing delay to model multiple packet arrivals in one TDM frame as well as possible backlog in previous TDM frames.

D_4 = Packet transmission delay (which equals packet size/link rate).

Note: The bar on top of each delay component represents average time.

The access delay can be computed using the queuing model shown in Figure43. A private individual queue of each ONU is modeled for D_1. The departure rate $\mu_p(t)$ of the private queue is time varying and is an impulse time sequence with deterministic interval T_f (Figure 44):

$$\mu_p(t) = \delta(t - nT_f - O_i) \qquad \text{(A.2)}$$

where O_i = time offset of request minislot from beginning of current TDM frame for the i^{th} ONU. After the TDM frame reservation time (modeled using a delay line), packets from different ONUs are added into a single common (global) queue served by the OLT.

For single arrival case, the arrival process of the i^{th} ONU to the common queue is a deterministic interval process with batch size represented by a Poisson distributed random variable $Y_i(n)$. The overall arrival process A of the common queue is the aggregation of the individual arrival processes from each ONU. Because all request minislots are in the first slot of each TDM frame, we have $A(n) = \sum_{i=1}^{10} Y_i(n)$, where

$$P[A(n) = m] = \frac{\lambda^m e^{-\lambda}}{m!}, \qquad and \qquad \lambda = \sum_{i=1}^{10} \lambda_i \qquad \text{(A.3)}$$

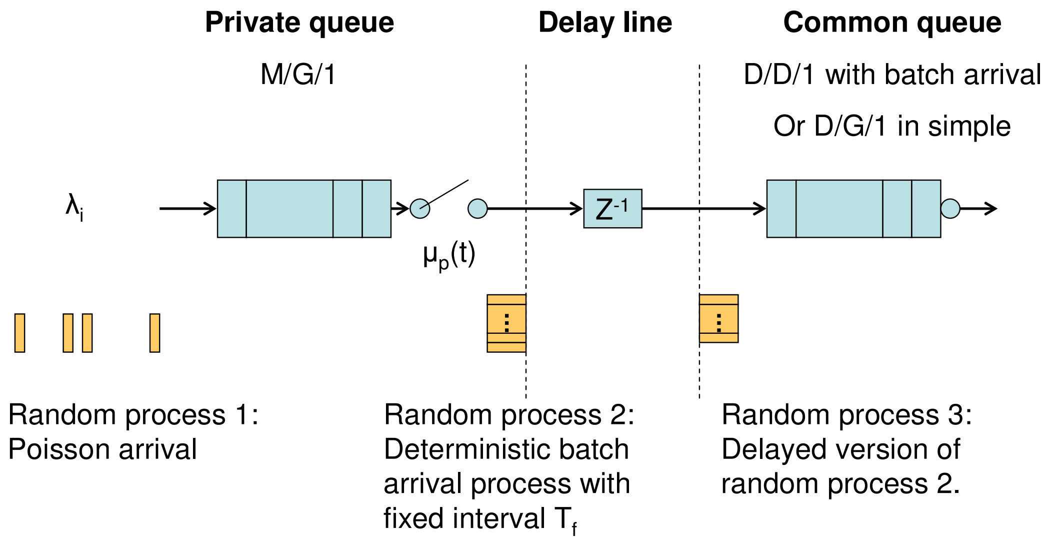


Figure 43: Access delay queuing model.

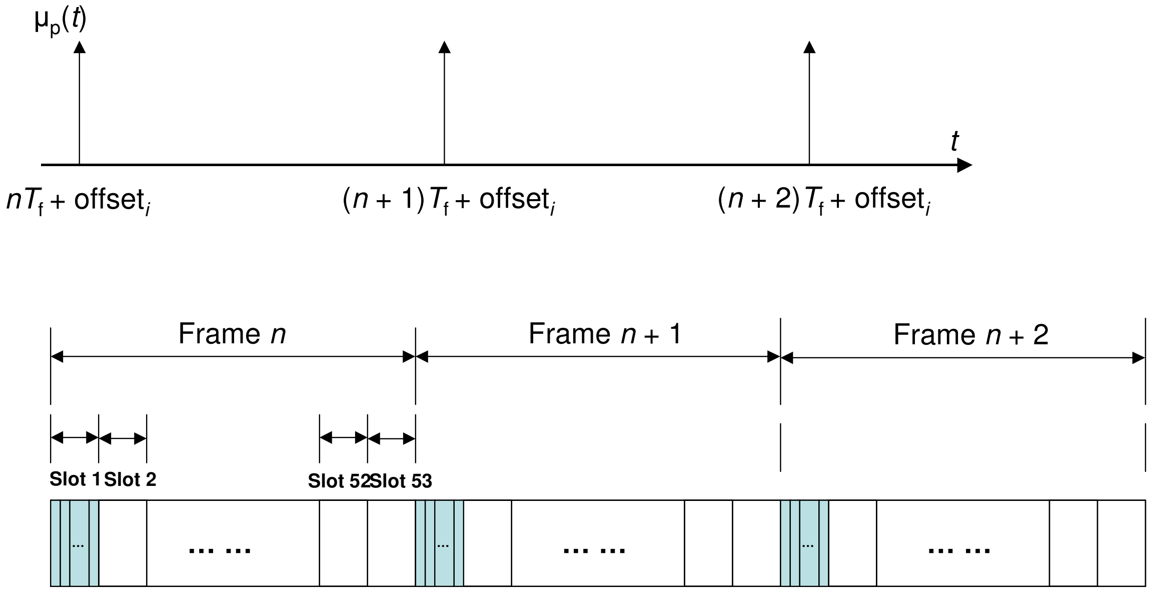


Figure 44: Time-varying departure rate $\mu_p(t)$ of the private queue.

For the batch arrival case, we define

Y = Number of Poisson batches arriving in one T_f.

B_i = Number of packets in the ith batch: an i.i.d. random variable with a discrete uniform distribution between 1 and 30 (a specific case based on our assumption), for $i = 1 \ldots N$.

A = Number of packets arriving at the common queue in one T_f.

$$A = \sum_{i=1}^{Y} B_i, \qquad P(Y = n) = \frac{\lambda^n e^{-\lambda}}{n!} \quad . \qquad \text{(A.4)}$$

The mean and variance of A (i.e., μ_A, σ^2_A) can be calculated as follows:

$$\begin{aligned}
\mu_A &= E(A) = \sum_{n=0}^{\infty} (\mu_A \mid n) P(Y = n) = \frac{31}{2} \sum_{n=0}^{\infty} (nP(Y = n)) = \frac{31}{2} \mu_Y = \frac{31}{2} \lambda \\
E(A^2) &= \sum_{n=0}^{\infty} E(A^2 \mid n) P(Y = n) = \frac{961}{4} \sum_{n=0}^{\infty} (n^2 P(Y = n)) + \frac{899}{12} \sum_{n=0}^{\infty} (nP(Y = n)) \\
&= \frac{961}{4} (\mu_Y^2 + \sigma_Y^2) + \frac{899}{12} \mu_Y = \frac{961}{4} (\lambda^2 + \lambda) + \frac{899}{12} \lambda \\
&= \frac{961}{4} \lambda^2 + \frac{1891}{6} \lambda
\end{aligned}$$

$$\sigma_A^2 = E((A - \mu_A)^2) = E(A^2) - \mu_A^2 = \frac{961}{4} \lambda^2 + \frac{1891}{6} \lambda - \frac{961}{4} \lambda^2 = \frac{1891}{6} \lambda$$

The common queue can be modeled as a discrete-time, bulk service $D/D/1$ queue with batch arrivals of data cells. Alternatively, each batch can be treated as a varying size packet in continuous time, giving rise to a $D/G/1$ queue [54] for the common queue, where the service time is Poisson distributed (this model is not considered in this paper).

We now derive the discrete-time queue size analysis of the common queue. Time is assumed to be slotted with a given slot duration, in our case, this duration is TDM frame time T_f. In [55], the discrete-time queuing model with bulk service is defined by the recursion: $X_{n+1} = \max\{X_n - s, 0\} + A_n$, where X_n denotes the queue length at the beginning of slot n, A_n denotes the number of new arrivals during slot n, which is the same with A in (A.3) and (A.4), and s denotes the fixed number of customers that can be served during one slot time (corresponding to one TDM frame).

The mean value of X_n can be solved using a numerical method. However, this is computational complex and lack connection to the first and second order moments of A_n. Instead, we can employ the bounding techniques [55] to derive the upper bound and lower bound of the average discrete-time queue:

$$\frac{\sigma_A^2}{2(s-\mu_A)}+\frac{\mu_A}{2}\le\mu_X\le\frac{\sigma_A^2}{2(s-\mu_A)}+\frac{\mu_A}{2}+\frac{\min\{\mu_A,s-1\}}{2} \quad . \quad \text{(A.5)}$$

where μ_A is the mean of A_n, $\sigma_A{}^2$ is the variance of A_n, and μ_X is the mean of X_n. Another lower bound can be derived:

$$\mu_X = E(\max\{X_n - s, 0\}) + \mu_A \le \mu_A . \qquad \text{(A.6)}$$

It can be shown that for Poisson distributions, A_n, the lower bound in (A.6), is tighter than (A.5) except under high traffic load. In this paper, the lower bound is expressed as:

$$\mu_X \le \max\left(\mu_A, \frac{\sigma_A^2}{2(s-\mu_A)}+\frac{\mu_A}{2}\right) \qquad \text{(A.7)}$$

We are now in a position to derive all the delay components associated with equation A.1. As the arrival process is random,

$$\overline{D_1} = \frac{T_f}{2}$$

The average delay from the i^{th} request minislot of TDM frame n to the first data slot of TDM frame $n + 1$ is computed as follows. In the simulation, the first slot (56 bytes) is divided into minislots for transmission of request packets and each request packet (and minislot) is 4 bytes. Hence, the first 40 bytes are used for request packet transmission, giving:

$$\overline{D_2} = T_f + (1 - 40/56/2)T_s$$

The average queuing delay for the common queue is

$$\overline{D_3} = average\_occupancy\_upon\_arrival \times average\_service\_time$$

If the length of common queue is less than s (i.e., $X_n \leq s$), all queued packets can be transmitted in the current TDM frame, so *average_service_time* = $T_s = T_f/53$. However, if the common queue length is larger than s (i.e., $X_n > s$), some queued packets will be transmitted in a later TDM frame, so $T_s = T_f/53 \leq$ *average_service_time* $\leq T_f/52$. Thus, the lower and upper bounds for the mean of D_3 are:

$$\overline{D}_{3\_lower\_bound} = \left[\left(lower\_bound\_queue\_size - \mu_A \right) + \frac{E\left(A \cdot \dfrac{(A-1)}{2} \right)}{E(A)} \right] \times \frac{T_f}{53}$$

$$\overline{D}_{3\_upper\_bound} = \left[(upper\_bound\_queue\_size - \mu_A) + \frac{E\left(A \cdot \frac{(A-1)}{2} \right)}{E(A)} \right] \times \frac{T_f}{52}$$

In our simulation, s = 52. In single arrival case, the relation between the normalized traffic load (0 ≤ ρ < 1) and packet arrival rate λ is λ = ρ x 53 x 56/53 = 56ρ. For a stable queuing system, the arrival rate must be less than the departure rate, so λ < s = 52 or equivalently 0 ≤ ρ < 0.928. For 10 ONUs, the average queue size in each ONU is (μ_X +λ)/10, where λ/10 is the average number of packets in the private queue and delay line.

REFERENCES

[1] B. Bing and G. K. Chang, “Chapter 2: The Last Mile, the Edge, and the Backbone,” in Broadband Last-Mile Technologies, N. Jayant (ed.), CRC Press, USA, 2005, ISBN: 0824758862.

[2] ITU-T Standards G.983.1 to G.983.8, “Broadband optical access systems based on passive optical networks (PONs)”.

[3] ITU-T Standards G.984.1 to G.984.2, “Gigabit-capable passive optical networks (GPON)”.

[4] B. Davie, A. Charny, J.C.R. Bennett, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, “An expedited forwarding PHB (Per-Hop Behavior),” RFC 3246, March 2002.

[5] J. Wroclawski, “The use of RSVP with IETF integrated services,” RFC 2210, September 1997.

[6] F. Le Faucheur, W. Lai, “Requirements for support of differentiated services-aware MPLS traffic engineering,” RFC 3564, July 2003.

[7] C. Xiao, B. Bing and G. K. Chang, “An Efficient Reservation MAC Protocol with Preallocation for High-Speed WDM Passive Optical Networks,” *IEEE INFOCOM 2005*, Miami, FL, March 13-17, 2005

[8] C. Xiao and B. Bing, “Delta Compression with Fixed-Length Substring Coding for Fast Content Download,” *IEEE Communication Letters*, vol. 9, no. 3, pp. 243- 245, March 2005

[9] C. Xiao, B. Bing and G. K. Chang, “Delta Compression for Fast Wireless Internet Download,” *IEEE Globecom 2005*, St. Louis, MO, Nov. 28-Dec. 2, 2005

[10] C. Xiao, G.K. Chang, B. Bing, “An SLA-aware Transport Protocol for High Throughput Wide Area Ethernet Services,” *IEEE GLOBECOM 2006*, to appear

[11] C. Xiao, C. Estevez, G.K. Chang, “Performance Evaluation of an SLA-Aware Transport Control Protocol for Ethernet Services,” *OPNETWORK 2006*, Washington DC, Aug. 28-Sep. 1, 2006

[12] C. Xiao, C. Estevez, B. Bing, G. Ellinas, and G.K. Chang, “A Resilient Transport Control Scheme for Metro Ethernet Services Based on Hypothesis Test,” *IEEE ICC 2007*, submitted.

[13] NSF Workshop Report, “Residential broadband revisited: research challenges in residential networks, broadband access, and applications,” January 20, 2004.

[14] Jey K. Jeyapalan, “Municipal optical fiber through existing sewers, storm drains, waterlines, and gas pipes may complete the last Mile,” *International Conference on Pipeline Engineering and Construction*, Baltimore, Maryland, July 13–16, 2003,

[15] S. Floyd, “HighSpeed TCP for large congestion windows,” RFC 3649, Dec. 2003.

[16] M. Allman, V. Paxson, W. Stevens, “TCP congestion control,” IETF RFC2581, Apr. 1999.

[17] K. Fall, S. Floyd, "Simulation-based comparisons of Tahoe, Reno and Sack TCP," *Computer Communication Review*, July 1996.

[18] A. Falk, T. Faber, J. Bannister, A. Chien, R. Grossman, J. Leigh, “Transport protocols for high performance,” *Communications of the ACM*, vol. 46, no. 11, pp. 43-49, Nov. 2002.

[19] T. Kelly, “Scalable TCP: Improving performance in high speed wide area networks,” *ACM SIGCOMM Computer Communication Review*, Vol. 33 , Issue 2 , April 2003, pp. 83 – 91.

[20] D. Katabi, M. Handley, C. Rohrs, "Internet congestion control for future high-bandwidth-delay product environments," in *Proceedings of ACM SIGCOMM'02*, Pittsburgh, PA, Aug. 19-21, 2002.

[21]H. Sivakumar, R. L. Grossman, M. Mazzucco, Y. Pan, Q. Zhang, “Simple Available Bandwidth Utilization Library for High-Speed Wide Area Networks,” *The Journal of Supercomputing*, vol. 34, no 3, pp. 231-242, December 2005.

[22] http://newsinfo.iu.edu/news/page/normal/588.html, Oct. 2006.

[23] Q. Wu, Control of transport dynamics in overlay networks. Ph.D. dissertation, Dept of Computer Science, Louisiana State University, March 2003.

[24] E. He, J. Leigh, O. Yu, T.A. DeFanti, "Reliable blast UDP: predictable high performance bulk data transfer," in *Proc. of IEEE International Conference on Cluster Computing (CLUSTER'02)*, Chicago, Illinois, September 23-26, 2002.

[25] Q. Wu, Nageswara S. V. Rao, "A class of reliable UDP-based transport protocols based on stochastic approximation," in *Proc. of IEEE INFOCOM 2005*, Miami, FL, March 2005.

[26] J. Ziv and A. Lempel, "A universal algorithm for data compression," *IEEE Transactions on Information Theory*, Vol. 23, No. 3, May 1977, pp. 337 – 343.

[27] J. MacDonald. File system support for delta compression. MS Thesis, UC Berkeley, May 2000.

[28] J.J. Hunt, K.P. Vo, W. Tichy, "Delta algorithms: an empirical analysis," *ACM Trans. on Software Engineering and Methodology*, Vol 7, No.2, April 1998, pp. 192-214.

[29] D. Trendafilov, N. Memon, T. Suel, "Zdelta: an efficient delta compression tool," Department of Computer and Information Science, Polytechnic University Technical Report, TR-CIS-2002-02, June, 2002.

[30] D. Korn, J. MacDonald, J. Mogul, K. Vo, "The VCDIFF generic differencing and compression data format," RFC3284, June 2002.

[31] J. Mogul, J. Krishnamurthy, B. Douglis, F. Feldmann, A. Goland, Y. A. Van Hoff, "Delta encoding in HTTP," RFC 3229, January 2002.

[32] M. Ajtai, R. Burns, R. Fagin, D. D. E. Long and L. Stockmeyer, "Compactly encoding unstructured inputs with differential compression," *Journal of the ACM*, Vol. 49, No.3, May 2002, pp. 318-367

[33] C.-J. Chae, E. Wong, R.S. Tucker, “Optical CSMA/CD media access scheme for Ethernet over passive optical network,” *IEEE Phot.Tech. Lett.*, vol. 14, no. 5, May 2002, pp. 711–13.

[34] Michael P. McGarry, Martin Maier, Martin Reisslein, “Ethernet PONs: a survey of dynamic bandwidth allocation (DBA) algorithms,” *IEEE communications magazine*, vol.42, no.8, Aug. 2004, pp.s8-s15

[35] J. Zheng, H. T. Mouftah, “Media access control for Ethernet passive optical networks: an overview,” *IEEE communications magazine*, vol.43, no.2, Feb. 2005, pp.145-150.

[36] Y. Luo, N. Ansari, “Bandwidth allocation for multiservice access on EPONs,” *IEEE communications magazine*, vol.43, no.2, Feb. 2005, pp.s16-s21.

[37] J. Xie, S. Jiang, Y. Jiang, “A dynamic bandwidth allocation scheme for differentiated services in EPONs,” *IEEE Communications Magazine*, Vol. 42, No. 8, August 2004, pp. s32 – 39.

[38] G. Kramer, B. Mukherjee, G. Pesavento, “IPACT: a dynamic protocol for an Ethernet PON (EPON),” *IEEE Communications Magazine*, Vol. 40, No. 2, February 2002, pp. 74 – 80.

[39] P.A. Bonenfant, S.M. Leopold, “Trends in the US communications equipment market: a wall street perspective,” *IEEE Communication Magazine*, Vol. 44, No. 2, pp. 141-147, Feb. 2006.

[40] J. Wang, “Optical Ethernet: making Ethernet carrier class for professional services,” *Proceedings of the IEEE*, Vol. 92, No. 9, pp. 1452-1462, Sep. 2004.

[41] A. Meddeb, “Why Ethernet WAN Transport,” *IEEE Communication Magazine*, Vol. 43, No. 11, pp. 136-141, Nov. 2005.

[42] R. Santitoro, “Metro Ethernet Services – A Technical Overview,” Metro Ethernet Forum, www.metroethernetforum.org/metro-ethernet-services.pdf, Dec. 2003.

[43] H. Chamas, W. Bjorkman, M.A. Ali, “A novel admission control scheme for Ethernet services,” *Proceedings of ICC 2005*, May 2005, pp. 65-69.

[44] P. Risbood, S. Acharya, B. Gupta, “The BEST challenge for next-generation Ethernet services,” *Proceedings of Infocom 2005*, Vol. 2, Mar. 2005, pp. 1049-1059.

[45] H. Chamas, W. Bjorkman, S. Liu, M.A. Ali, “Verizon experience with NG Ethernet services: evolution to a converged layer 1, 2 network,” *IEEE Communication Magazine*, Vol. 43, No. 8, pp. s18-s25, Aug. 2005.

[46] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, New York: John Wiley & Sons, 1991

[47] E. Blanton, M. Allan, K. Fall, L. Wang, “A conservative selective acknowledgment (SACK) –based loss recovery algorithm for TCP,” RFC 3517, IETF, April 2003.

[48] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, “Modeling TCP throughput: a simple model and its empirical validation,” in *the Proceedings of the ACM SIGCOMM*, Sep. 1998

[49] Alberto Leon-Garcia, Indra Widjaja, *Communication Networks*, The McGraw-Hill Companies, Inc., 2004.

[50] R. Santitoro, “Metro Ethernet Services Overview,” www.metroethernetforum.org/ Metro-Ethernet-Services-Overview.ppt, Metro Ethernet Forum, Oct. 2005.

[51] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, “On the Self-Similar Nature of Ethernet Traffic (Extended Version),” *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, Feb. 1994

[52] A. Kesselman, Y. Mansour, “Adaptive AIMD congestion control,” *Proceedings of the twenty-second annual symposium on Principles of distributed computing*, Boston, MA, pp 352 – 359, 2003

[53] Hogg, McKean, and Craig, Introduction to Mathematical Statistics, Sixth Edition, 2004

[54] L. D. Servi, “D/G/1 queues with vacations,” *Operations Research*, Vol. 34, 1986, pp. 619 – 629.

[55] D. Denteneer, J. van Leeuwaarden, and J. Resing, “Bounds for a discrete-time multi-server queue with an application to cable networks,” in *Proc. of the 18th International Telecommunications Conference (ITC)*, Berlin, 2003, pp. 601-612.