

Carrier Ethernet Network Solutions: Transport Protocol and Optical Backplane Design

A Thesis
Presented to
The Academic Faculty

By

Claudio Ignacio Estevez Montero

In Partial Fulfillment
of the Requirements for the Degree

Doctor of Philosophy in Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia

May 2010

Carrier Ethernet Network Solutions: Transport Protocol and Optical Backplane Design

Claudio Ignacio Estevez Montero

Approved by:

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

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

Dr. Henry Owen
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Georgios Ellinas, Co-Advisor
Department of Electrical and Computer
Engineering
University of Cyprus

Dr. Chuanyi Ji
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Calton Pu
College of Computing
Georgia Institute of Technology

Date Approved: November 16, 2009

Acknowledgements

I would like to thank the committee members that compose the panel for my PhD Thesis defense, which are composed of Dr. John Buck, Dr. Chuanyi Ji, Dr. Calton Pu, Dr. Henry Owen, Dr. Georgios Ellinas (Co-Advisor), and Dr. Gee-Kung Chang (Advisor). This distinguished group of scholars provided me with guidance and priceless knowledge.

A special thanks goes to the people of the OPNET Corporation for providing me with one of the most powerful network simulation software. OPNET provided this software at no cost to us. Their generosity did not stop there, they also provided us with professional tutoring to teach us how to use their software and answer any technical questions that we had.

I am in debt to the management of the Packaging Research Center (PRC) for allowing me to use their facilities located in the next-generation substrate lab of the Manufacturing Research Center (MaRC). I would also like to thank the National Science Foundation (NSF) for funding PRC and part of my thesis research.

I cannot express enough gratitude to my advisor Gee-Kung Chang, who accepted me into his research group, who has guided me for the last 4 years, who funded my doctorate studies, who tough me so many things and most importantly for being a good friend.

A lot of my experience I owe to Dr. Georgios Ellinas, my co-advisor, and for that I cannot thank him enough. Dr. Ellinas has greatly expanded my knowledge in computer networks and taught me much about the professional publication and presentation process. He has helped me tremendously and has done so from a distance as he resides in Cyprus.

A special thanks goes to my mentor Dr. Daniel Guidotti. I and Dr. Guidotti spent numerous hours working in the PRC lab. Everything I know about optical interconnect technology I owe to his teachings.

I would like to thank my colleague, mentor and good friend Dr. Chunpeng Xiao who provided me with much guidance throughout my PhD years. His teachings were not limited to computer networks but he gave me advice on professional and personal matters.

Thanks, to the Optical Network Research Group (ONRG) for always been willing to help, not only current members but also members that graduated and still keep in communication. I cannot thank everyone individually, but I would like to mention a few of my closest friends: Zhensheng Jia, Shu-hao Fan, Yin-Jung Chang, Hung-Chang Chien, Lingbin Kong, Oladeji Akanbi, Wei Jian, and Cheng Liu.

I would like to extend a thank you to my colleagues and friends, outside ONRG, who have always been there to help and provide moral support. To mention a few: Aravind Kailas, Alper Akanser, Matthew Trotter, and Aaron Hatch.

I cannot leave out my family, which have supported me and encouraged me to pursue my goals, not just through my doctorate studies but all my life. I would like to give a warm thanks to Dr. L. Antonio Estevez, Mireya Montero, Dr. Marcel Estevez and Nicolas Estevez. I love you all.

Contents

Acknowledgements	iii
List of Tables	viii
List of Figures	ix
List of Abbreviations	xi
Summary	xiv
1. Introduction	1
2. Origin and History of the Problem	5
2.1 Carrier Ethernet Networks	5
2.2 Carrier Ethernet Network Terminology	6
2.3 Ethernet and TCP/IP	7
2.4 Ethernet Services	8
2.5 Bandwidth Limitations in Carrier Ethernet Networks	9
2.5.1 Transport Control Protocol	10
2.5.2 Network-edge Technology	12
3. Review of Previous Work	15
3.1 Ethernet Services Transport Protocol	15
3.2 Side-mounted VCSEL	17

4. Carrier Ethernet Network Solutions	20
4.1 Evolution of Carrier Ethernet Networks	20
4.2 Throughput Bottleneck and Solutions for Carrier Ethernet	21
4.3 Carrier Ethernet Elements	24
4.4 Bandwidth Profile enforced by the User-network Interface	25
5. Ethernet Services Transport Protocol (ESTP)	27
5.1 Overview	27
5.2 Traffic Loss Interval to Multiplicative-decrease Factor Mapping	28
5.3 Congestion Control Obtained by Combining Congestion Feedback with Ethernet Services Information	32
5.4 Analytical Expression for ESTP Throughput	34
5.5 Computational Overhead Raw Comparison between the Proposed Protocol and Traditional TCP	36
5.6 ESTP backward compatibility with Traditional TCP and TCP-friendliness	37
5.6.1 Backward compatibility	37
5.6.2 TCP-friendliness	38
5.7 Bandwidth Profile Translation Protocol	39
5.8 UNI Scheduling Scheme	42
6. Side-mounted VCSEL	45
6.1 Overview	45
6.2 Advantages of Optical Interconnect over CMOS Technology	47
6.2.1 Bandwidth	47

6.2.2 Wire Density	48
6.2.3 Power Efficiency	49
6.3 Architectural Design	50
6.4 Fabrication Process	51
7. Objective and Design of Experiments and Simulations	55
7.1 Overview	55
7.2 Ethernet Services Transport Protocol Simulations	55
7.2.1 Throughput Performance	55
7.2.2 Translation Protocol Performance	56
7.2.3 UNI Scheduling Scheme Performance	57
7.3 Side-mounted VCSEL Experiments	57
7.3.1 Horizontal and Vertical Alignment Sensitivity	59
7.3.2 Coupling Efficiency	60
7.3.3 Crosstalk	61
8. Simulation Results: Ethernet Services Transport Protocol	62
8.1 ESTP Performance Throughput Performance Simulation	62
8.2 Translation Protocol Simulation	65
8.3 UNI Scheduling Scheme Simulation	67
9. Experimental Results: Side-mounted VCSEL	70
9.1 Horizontal Alignment Sensitivity Experiment	70
9.2 Vertical Alignment Sensitivity Experiment	71
9.3 Coupling Efficiency Experiment	73

9.4 Crosstalk Experiment	76
10. Conclusions, Contributions and Future Work	78
10.1 Conclusions	78
10.1.1 Ethernet Services Transport Protocol	78
10.1.2 Side-Mounted VCSEL	79
10.2 Contributions	80
10.2.1 Ethernet Services Transport Protocol	80
10.2.2 Side-mounted VCSEL	80
10.3 Future Work	80
10.3.1 Ethernet Services Transport Protocol	80
10.3.2 Side-mounted VCSEL	81
OPNET Modeler®	82
Next-Generation Substrate Lab	84
Scheduling Schemes	86
C.1 Strict Priority Scheduling Scheme	86
C.2 Weighted Fair Scheduling Scheme	87
Leaky-bucket Algorithm	88
References	90
Vita	96
List of Publications	97

List of Tables

Table 1. Simulation Parameters for Scenario 8.1	63
Table 2. Simulation Parameters for Scenario 8.2.	66
Table 3. Simulation Parameters for Scenario 8.3.	68
Table 4. VCSEL Specifications	75

List of Figures

Figure 1. IBM schematic of side-mounted VCSEL.	18
Figure 2. NEC schematic of side-mounted VCSEL.	19
Figure 3. Evolution of Ethernet into Carrier Ethernet Networks	20
Figure 4. Research areas in which the proposed work focuses.	22
Figure 5. Main elements of Carrier Ethernet Networks	24
Figure 6. Bandwidth Profile enforced by the User-network Interface	26
Figure 7. Value of α is the amount of successful packets transmitted in between two packet losses plus one.	28
Figure 8. Proposed protocol packet loss interval to multiplicative decrease factor mapping.	31
Figure 9. Congestion window behavior of (a) ESTP and (b) of Traditional TCP.	32
Figure 10. (left) graphical display of bandwidth with no dynamic control of CIR_{ESTP} and EIR_{ESTP} . (right) graphical display of dynamic resizing of CIR_{ESTP} and EIR_{ESTP} .	41
Figure 11. Proposed UNI Scheduling Scheme.	43
Figure 12. Conventional architecture for VCSEL-waveguide coupling.	46
Figure 13. Proposed architecture for VCSEL-waveguide coupling.	50
Figure 14. Graphical view of the process required to optically couple the VCSEL to the PIN PD (only VCSEL side is shown).	53
Figure 15. (left) VCSEL aligned with waveguide (Rohm & Haas) after process completion. (right) PIN PD active region aligned with waveguide after process completion.	54
Figure 16. Throughput simulation scenario.	56

Figure 17. Top view of VCSEL placed on sub-anchor board without waveguides.	58
Figure 18. Infrared photograph: light emitted by VCSEL traveling through the waveguide.	59
Figure 19. Background Traffic.	62
Figure 20. Goodput comparison between ESTP, HighSpeed TCP and TCP-Sack.	64
Figure 21. Goodput improvement ratio between ESTP and TCP-Sack.	65
Figure 22. UNI parameter update period vs. the bandwidth utilization efficiency.	67
Figure 23. Throughput behavior of different priority level as RTT is varied.	69
Figure 24. (top) Cross-section of the waveguide (Rohm & Haas) used to gather alignment sensitivity with the horizontal path outlined. (bottom) Horizontal alignment sensitivity results.	71
Figure 25. (top) Cross-section of the waveguide (Rohm & Haas) used to gather alignment sensitivity with the vertical path outlined. (bottom) Vertical alignment sensitivity results.	72
Figure 26. Experimental setup for coupling efficiency experiment.	73
Figure 27. Relation between the VCSEL's injected current and output power.	74
Figure 28. Relation of the photo-current for each channel with varying VCSEL injection current.	74
Figure 29. Coupling Efficiency results for both channels.	75
Figure 30. Infrared camera photographs. (top) Side view. (bottom) View from behind the PD.	76
Figure 31. Crosstalk experimental results.	77
Figure 32. The different working domains of OPNET Modeler®.	83
Figure 33. Strict priority scheduling flow diagram.	86
Figure 34. Weighted fair scheduling flow diagram.	87
Figure 35. State Diagram of Leaky-bucket Algorithm	88
Figure 36. Graphical View of Leaky-bucket Algorithm	89

List of Abbreviations

ACK	Acknowledged Segment
AIMD	Additive Increase and Multiplicative Decrease
ATM	Asynchronous Transfer Mode
BDP	Bandwidth Delay Product
BW	Bandwidth
CBS	Committed Burst Size
CE	Customer Equipment
CEN	Carrier Ethernet Network
CIR	Committed Information Rate
CMOS	Complementary metal-oxide-semiconductor
CoS	Class of Service
cwnd	Congestion Window
DSL	Digital Subscriber Line
E/O	Electrical to Optical Conversion
EBS	Excess Burst Size
EIR	Excess Information Rate
E-Line	Ethernet Line
E-LAN	Ethernet LAN
EPL	Ethernet Private Line
EP-LAN	Ethernet Private LAN
EP-Tree	Ethernet Private Tree
ESTP	Ethernet Services Transport Protocol
EVC	Ethernet Virtual Connection
EVPL	Ethernet Virtual Private Line

EVP-LAN	Ethernet Virtual Private LAN
EVP-Tree	Ethernet Virtual Private Tree
FIFO	First-in first-out
Ge	Germanium
GbE	Gigabit Ethernet
HFC	Hybrid Fiber-Coax
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISI	Inter-Symbol Interference
LAN	Local Area Network
LC	Inductive-Capacitive
MAC	Media Access Control
MAN	Metropolitan Area Network
MEF	Metro Ethernet Forum
MEN	Metro Ethernet Network
NNI	Network-network Interface
O/E	Optical to Electrical Conversion
OSI	Open System Interconnection
PD	Photodiode/Photodetector
PIN	p-Type/Intrinsic/n-Type
QoS	Quality of Service
RFC	Request for Comment
RTT	Round Trip Time
SLA	Service Level Agreement
SPQ	Strict Priority Queuing
TCP	Transmission Control Protocol
ULH	Ultra Long-haul
UNI	User Network Interface
UV	Ultraviolet
VCSEL	Vertical-Cavity Surface-Emitting Laser
VLAN	Virtual Local Area Network

VoIP	Voice over Internet Protocol
VPN	Virtual Private Network
WAN	Wide Area Network
WFQ	Weighted-Fair Queuing
XCP	Explicit Control TCP

Carrier Ethernet Network Solutions: Transport Protocol and Optical Backplane Design

Claudio Ignacio Estevez Montero

Summary

The Metro Ethernet network (MEN) expands the advantages of Ethernet to cover areas wider than LAN. MENs running Ethernet Services as specified by the Metro Ethernet Forum (MEF) are known as Carrier Ethernet Networks (CENs). CENs can cover not only metro areas, but it can expand to cover global areas by connecting multiple MENs. Next-generation CENs are expected to support 100 GbE. With arising technologies for Ultra Long-haul (ULH) networks the bandwidth bottleneck of CENs is shifting to other areas like the transport layer protocol (such as the Transport Control Protocol or TCP) and the chip-to-chip channel capacity found at the network edge, which in general has an electrical backplane. Traditional TCP is well known to have difficulties reaching the full available bandwidth, due to its inefficient AIMD mechanisms under a high-delay-bandwidth-product environment. At the network edge, network equipment with electrical backplanes poses many problems including inductive-capacitive effects that limit its bandwidth. These are the two main issues addressed in this work. To resolve the transport layer issue, this work proposes a transport protocol that fully utilizes the available bandwidth while preserving TCP-friendliness and providing QoS support that is compatible with Ethernet Services. It can guarantee throughputs above the Committed Information Rate (CIR), which is specified in the Service Level Agreement (SLA). To resolve the physical layer limitations, a novel optical coupling technique is examined to encourage the use of optical backplanes for network-edge and core technology. The proposed technique consists of aligning the normal of the laser emission plane, waveguide plane and the normal of the photodetector active region plane with the purpose of reducing optical power loss caused by common methods of light manipulation. By addressing the shortcomings of both Traditional TCP and electrical backplane technology the overall throughput can be significantly increased.

Chapter 1

Introduction

Ethernet has come a long way since its invention in the early 70s. Even though the frame structure, protocols and routing devices of today are very different from those designed and implemented four decades ago the essence of computer interconnection has been preserved. Ethernet started as a way to connect multiple computers to each other, this group of computers is called a local area network (LAN). With further advances the local area network eventually grew to cover much larger areas such as metro areas, called metro area networks (MAN) and beyond, referred to as wide area networks (WAN). Even today Ethernet keeps growing not just in size but in speeds. Ethernet started from just under 3 Mbps transmission rate, which was the first recorded speed, and has evolved to 10 Gbps. Now that we know where Ethernet came from the underlying question is: In which direction is Ethernet headed to?

The growth of Ethernet has made digital sharing simpler, faster and more economical. This has many companies migrating their services from other network types to Ethernet. This transition is not straightforward as Ethernet in its raw form is not carrier-grade. For this reason an organization called the metro Ethernet forum (MEF) has taken the task of organizing, structuring and standardizing Ethernet to make it carrier-grade, and this new version of Ethernet is called Carrier Ethernet. The latest Carrier Ethernet specifications are designed to work at 100 Gbps and cover areas worldwide. With 31 service providers and 76 network equipment manufactures companies certified by the MEF, revenues in the hundreds of dollars per year [47],

and predictions stating billions of dollars in revenues by 2012 [48], it is becoming more and more evident that the next generation of Ethernet is headed toward the Carrier Ethernet Network (CEN).

Next-generation CENs are expected to support 100 GbE. With arising technologies for ultra long-haul (ULH) networks, the bandwidth bottleneck of CENs will not be the core of its framework but rather at the edges of the network. We foresee that two specific areas will be the main limiters of high-bandwidth transmissions: the transport control protocol and the electrical backplanes currently used in network-edge technology. It is well known that the transport control protocol (TCP) has trouble reaching high transmission rates in high bandwidth-delay-product (BDP) networks. With CEN covering global areas (incurring longer delays) and supporting 100 GbE, the performance of TCP will be greatly deteriorated. This is due to the additive-increase multiplicative-decrease (AIMD) congestion-avoidance algorithm that TCP has. This algorithm will increase the congestion window size very conservatively while no loss is detected and will reduce it aggressively if a segment is lost, making it difficult to reach a congestion window size that will fully utilize the available bandwidth. It is clear that modifications to TCP need to be made to overcome this problem. The proposed protocol, which will be discussed in detail in later chapters, is built on top of TCP – maintaining its compatibility and TCP-friendliness – but modified to improve bandwidth utilization and also to provide quality of service (QoS) at the transport layer. The proposed protocol has the capability to gather feedback on the level of congestion the connection is experiencing using only information available at the transport layer. By reading the level of congestion the protocol can adjust the congestion window size accordingly, such that it is not aggressive under low congestion conditions and vice-versa. Besides providing a novel congestion-control mechanism, the proposed protocol is also designed

to provide QoS based on a predetermined service level agreement (SLA). An simple example of how the integration of SLA information can help improve transmission rates is employing the committed information rate (CIR), which is a guaranteed amount of bandwidth provided to the customer. To accommodate this reserved bandwidth in the transport protocol, a minimum congestion window size is determined and maintained during transmission. This will prevent the transmission rate to drop below the CIR. Other QoS attributes are supported by the proposed protocol and will be discussed in detail further ahead.

As mentioned before, the other bandwidth limitation in next generation CENs is the electrical backplane, specifically at the network edge. In CEN terminology the network-edge device is called the user-network interface (UNI). The UNI demarks the point between the customer side and the network side. The UNI is a device that has the capability to multiplex data streams from multiple users and channel them through Ethernet virtual connections (EVCs). Multiple EVCs can be initiated at a single UNI and different services can be provided between different UNIs across the network. To handle all these responsibilities in a 100 GbE network it seems logical that an electrical backplane is no longer a practical or feasible alternative. It has been shown that electrical interconnects decrease bandwidth quadratically with increasing length [13]. To reach throughputs of 100 Gbps it is necessary to use an optical backplane. An optical backplane will require the integration of lasers, optical waveguides, and detectors. The integration of these elements is not trivial and improving the coupling efficiency and reliability is an ongoing research effort. A major difficulty has been the optical alignment of the three components, which is exacerbated by the fact that both a Vertical-Cavity Surface-Emitting Laser (VCSEL) and a standard p-Type/Intrinsic/n-Type (PIN) photo-detector (PD) are vertical-surface emitting and viewing devices, while waveguides are generally placed horizontally. A common

attempt at a solution has been to think in terms of traditional component integration and to assemble micro-lenses and micro-mirrors to collimate and bend the light path. For optical interconnects to become a technology, the integration process must demonstrate optical alignment among hundreds of optical channels in a reasonably straightforward and routine manner. For the technology to become a product, the integration process has to become mass producible and demonstrate a reliable performance advantage at competitive cost and lower power dissipation than is possible with existing products. This work investigates a novel direct integration process for VCSELs and optical waveguides. The approach is to assemble the VCSEL and PIN PD side-mounted on a substrate so that the light emission and viewing directions are both in the plane of the waveguide. The 90° turn is made electrically.

The transport protocol and optical backplane are the main issues discussed in this work, but there are other challenging issues addressed in this work. This document is organized as follows. Chapter 2 describes the main problems that are addressed in this work. Chapter 3 summarizes some of the previous work that has been done in these fields. Chapter 4 introduces the carrier Ethernet networks solutions. Chapter 5 discusses the proposed transport protocol, called Ethernet Services transport protocol (ESTP). Chapter 6 describes the architecture of the side-mounted VCSEL, the proposed solution for the backplane design at the network edge. Chapter 7 describes the simulations and experiments performed, including their setups and parameters used. Chapter 8 shows the results of the ESTP, while Chapter 9 shows the results of the side-mounted VCSEL work. Chapter 10 discusses the conclusions, contributions and future work that is derived from this work.

Chapter 2

Origin and History of the Problem

2.1 Carrier Ethernet Networks

The Ethernet network has come a long way from a simple multipoint data exchange network with collision detection over a coaxial cable bus. In the 1970s, the Ethernet was the architecture of an experimental network by the Xerox Corporation that reached a 3-Mbps transmission rate [12]. Because of its simplicity and bandwidth capabilities, it soon became the most popular technology for local area network (LAN) architectures, overshadowing competing alternatives like token ring and ARCNET. The demand for interconnectivity kept expanding to the extent of covering an entire metropolitan area. This Ethernet-based computer network was called Metro Ethernet Network (MEN). As the Ethernet evolved, it underwent changes in architecture and protocol to better suit the higher volume of end systems and higher data rates. One of the main catalysts for the growth of Ethernet-based networks was the need for large businesses to connect their offices to their Intranet with an inexpensive and high-bandwidth technology. What made MENs inexpensive was mainly that it used layer 2 (L2) switches for all its internal structures. The switches have a very simple and low-cost design, and are easily configured. The main setback with the L2 switches was that they were unable to isolate traffic, making privacy protection an impossible feat and therefore unsuited for service provider (SP) applications. The development of a new technique, called virtual LAN (VLAN), enabled transparent tunneling of traffic. With the capability to provide services over MENs, a new

carrier-grade network was born, called the carrier Ethernet network (CEN). CEN is a MEN network that can provide ubiquitous, standardized, carrier-class services and is defined by five attributes: standardized services, scalability, reliability, quality of service (QoS), and service management. These carrier-class services are called Ethernet services and their specifications are determined by the Metro Ethernet Forum (MEF). The MEF was created in 2001 and its first specifications were not released until 2004; this year demarks the birth of the standardized and certifiable class of CENs. A portion of the proposed work focuses on Ethernet services, but before describing what Ethernet services are, it is essential to understand the terminology used in CENs.

2.2 Carrier Ethernet Network Terminology

The MEF has defined various terms to make CENs documentation organized and unambiguous. This section describes the terminology, starting from the outside of the network moving toward the core. The farthest CEN component is the customer equipment (CE). The CE can be a workstation, server, data storage device, multiplexing device, or any other type of device that the customer wishes to connect to its intranet or internet via the CEN. The CE is connected to the CEN through the user-network interface (UNI). The UNI is the physical interface or port that marks the demarcation point between the customer and the service provider. The UNI is owned and managed by the service provider. UNIs are currently designed to support 10 Mbps, 100 Mbps, 1 Gbps, or 10 Gbps. The MEF is planning to add 100 Gbps support to future specifications. At the other end of the UNI is the MEN, which transports data from one end system to another. On some occasions, the data flows from one MEN to another MEN operated by a different service provider. When this occurs the data must go through a network-

network interface (NNI). This is an active project of the MEF and the details of the specification are still under development.

Ethernet services are provided through an Ethernet virtual connection (EVC). The EVC is a service container or tunnel that connects two or more customer sites. It does this by associating UNIs to a particular EVC. Data from an EVC cannot cross under any circumstances to another EVC. There are three types of EVCs: point-to-point, multipoint-to-multipoint, and point-to-multipoint. A single UNI is allowed to multiplex different EVCs, but without any data exchange between EVCs. With the understanding of these terms, Ethernet services can then be explained.

2.3 Ethernet and TCP/IP

Much of the work described in this document focuses on the transport layer, yet we proposed to solve issues in carrier Ethernet networks so it is important to understand the link between Ethernet and TCP/IP. Ethernet is a data link layer protocol, more specifically a MAC layer protocol. Each system in the world has a unique physical address (also called MAC address). When a system connects to a local area network, it is assigned an IP address, this IP address is mapped to the physical address in the local area network, but the physical address is only known by the switching device to which the end system is connected to. The core network only knows where all the IP addresses are located. It would be impractical to keep track of the location of the physical addresses in the core of the network because many of the systems (also can be referred as nodes in a topological sense) are constantly changing location, which is why the network layer protocols are necessary. One way to visualize this is to think of the network-layer protocols as location-based routing algorithms and MAC layer protocols as system-based routing algorithms.

The need for a connection-oriented protocol, such as TCP, is indispensable in the network. In a digital information domain composed of binary code sequences, the integrity of the information is critical. If even one bit was altered the binary sequence would be corrupted and unusable. For this reason when data is sent over a shared channel it is indispensable to preserve the integrity and order of each packet. Because of the large number of applications and their specific needs, it would be unpractical to embed all the different functionalities of the transport layer protocols into a single network layer format. It is much simpler to have an application decide what it needs, choose a transport layer service and encapsulate a specific segment format into a network layer datagram.

In this work it is suggested that TCP should be replaced by ESTP, but the network and MAC layers are left unaltered. Our concern is mainly solving the drawbacks of carrier Ethernet networks (not Ethernet as a MAC layer protocol), which at the present time we believe are found in the physical and transport layers. The implications and assumptions of the term *carrier Ethernet network* are described in much more detail in the following chapters, and are mainly associated with the quality of service that these networks provide.

2.4 Ethernet Services

Ethernet services are categorized by their tunneling architecture. To satisfy the needs of the service providers' applications, three distinct types of services were defined: E-Line, E-LAN, and E-Tree. The E-Line is the architecture type in which a point-to-point EVC is established between exactly two UNIs. E-Lines can be port-based or VLAN-based. Port-based means that each EVC goes to a dedicated UNI and neither UNI can support any other EVCs. In VLAN-based connections, multiple EVCs are multiplexed by a single UNI. A port-based E-Line is called an Ethernet private line (EPL). EPL offers network-edge privacy at the hardware level

because each EVC has a dedicated UNI and does not share a medium, at the network edge, with other EVCs. EPLs can replace TDM private lines. A VLAN-based E-Line is called an Ethernet virtual private line (EVPL). It can support multiplexed EVCs in a single UNI. The network-edge privacy is done at the software level, because multiple EVCs are sharing the same UNI, but the data is kept private by the UNI's algorithms. It is less costly than EPL and it is easily scalable, combined with its simplicity, which make it one of the most popular services offered by CENs. EVPLs can replace Frame Relay and ATM services. One of the CENs solutions proposed in this document focuses on the EVPL architecture, mainly because it has QoS capabilities, but more details will be provided later. Another service type is the E-LAN, composed of the Ethernet private LAN (EP-LAN), which is port-based, and Ethernet virtual private LAN (EVP-LAN), which is VLAN-based. The last service type is the E-Tree, similar to E-Line and E-LAN; it also has a port-based and VLAN-based service, which are Ethernet private tree (EP-Tree) and Ethernet virtual private tree (EVP-Tree), respectively. Because these are not part of the main focus of the proposed work, their details will be omitted from this section. Now that CEN and its services have been described, an explanation of CEN's bandwidth limitations will follow.

2.5 Bandwidth Limitations in Carrier Ethernet Networks

With each generation of technologies, the bandwidth bottleneck shifts from one Open Systems Interconnection (OSI) layer to another. Throughout the years protocols have been optimized, channel capacities have been increased, and compression levels have been improved. With the expected launch of 100 GbE in CEN, two main bandwidth bottlenecks exist. These are the transport control protocol and the electrical backplane of the network-edge technology. More details on both of these aspects are described below.

2.5.1 Transport Control Protocol

The Transport Control Protocol (TCP) started as a request for comment (RFC) memorandum in 1974, with several additions in subsequent years. TCP provides congestion avoidance, delivery guarantee, and data sequence organization, making the Internet Protocol (IP) best-effort network a reliable one. The trade-off for reliability is delay. TCP is not well-suited for time-sensitive applications, such as voice over IP (VoIP). With a new generation of networks arising, the link capacity is increased and the networks extend to cover greater distances (i.e., increasing delay). These large bandwidth-delay product (BDP) networks will pose a problem to TCP. The BDP is a measure of how much data is in-flight at one point in time. The in-flight data in TCP is controlled by the congestion window (cwnd) size. The cwnd size is determined by the additive-increase multiplicative-decrease (AIMD) algorithm of TCP. This algorithm decreases the cwnd aggressively in the event of a packet loss and increases the cwnd very conservatively, making it difficult to reach the full available bandwidth. It is necessary to make modifications to the standard type to be able to fully utilize the provided bandwidth in newer versions of the transport protocol. TCP is composed of four intertwined congestion-control algorithms: slow-start, congestion avoidance, fast retransmit, and fast recovery [7]. TCP went through several versions like Tahoe, Reno, New Reno, and Sack. A simulation-based comparison of the performance of these different versions is compared in earlier work [8]. TCP Tahoe, the most native of these four versions, had the slow-start, congestion avoidance and fast retransmission algorithms. In TCP Reno, fast recovery was integrated to the TCP Tahoe congestion-control algorithms. TCP New Reno is very similar to Reno, with the exception that the retransmit wait time, after multiple consecutive packet losses, was eliminated. TCP Sack is built on top of TCP New Reno with an added feature called selective acknowledgement, hence the name Sack. In

previous implementations of TCP, the receiver acknowledges the last successful segment received. However, other non-contiguous segments could have been received but not notified to the sender. If segments time out, this forces the sender to retransmit all segments after the last successful segment. With selective acknowledgement the sender acknowledges multiple non-contiguous segments such that the receiver only has to retransmit the lost segments. TCP Sack became the root from which many other protocols branched out, including HighSpeed TCP [9], Scalable TCP [10], Explicit Control TCP (XCP) [11], and SLA-aware Transport Protocol [1]. The proposed protocol is built on top of the SLA-aware Transport Protocol, which is a shifted-lower-bound version of TCP Sack. Like the proposed work, the SLA-aware Transport Protocol assumes a reserved bandwidth so it shifts the lower bound of TCP Sack to match the data rate of the reserved bandwidth. HighSpeed TCP tweaks the AIMD parameters of TCP to perform better in high-bandwidth-delay networks; the parameters are obtained empirically. Scalable TCP runs a multiplicative-increase multiplicative-decrease algorithm to compute the cwnd size. The main problem with this technique is that it is very aggressive in both directions (increase and decrease), and for networks with long delays this protocol could be unstable. In XCP a congestion header is added to the packet format. Routers communicate with the transport protocol through this header field to obtain congestion feedback. The problem with this protocol is that this capability is not available in TCP so current networks are not designed to have this feature; and replacing or reprogramming all the routers in the network to support this feature might not be a feasible solution. Another problem is that routers are designed to route by reading layer 2 or 3 information, and for the router to write congestion feedback in the layer 4 packet format requires further decapsulation of the packet and adds more delay, which defeats the purpose.

2.5.2 Network-edge Technology

The network-edge technology of CENs is currently designed to support bitrates of 10 Mbps, 100 Mbps, 1 Gbps, and 10 Gbps; future specifications will include support for 100 Gbps. The current components have an electrical backplane. It has been shown that electrical interconnects decrease bandwidth quadratically with increasing length and decrease bandwidth linearly with decreasing cross-sectional area, due to the skin effect of the inductive-capacitive (LC) limited lines [13]. An optical backplane will alleviate intrinsic problems related to the aspect ratio of conductive lines, as well as extrinsic problems common to electrical interconnects, such as propagation loss, pre-emphasis requirements, active equalization, power consumption, signal dispersion, clock distribution, system synchronization, shielding, capacitive isolation, wave reflection, crosstalk, wave reflection, and pin inductance. These shortcomings will impact the system in different ways: the signal power is increased to overcome the propagation loss, which can reach 50 dB/m on FR-4. This power increase, besides lowering the efficiency, generates heat, which requires higher cooling demand. Capacitive effects on the signal plane cause signal dispersion, which increase the inter-symbol interference (ISI). The lack of shielding can result in crosstalk between two channels, but the inclusion of shielding increases the bulk density. These electrical interconnect difficulties make high-bandwidth transmissions impractical. To reach throughputs of 100 Gbps it is necessary to use an optical backplane with multiple parallel channels. By using optical switching mechanisms, conversions between the electrical and optical domains can be avoided to further decrease delay. An optical backplane will require the integration of lasers, optical waveguides, and detectors. The integration of these elements is not trivial and improving the coupling efficiency and reliability is an ongoing research effort. A major difficulty has been the optical alignment of the three components, which

is exacerbated by the fact that both a VCSEL and a standard PIN PD are vertical-surface emitting and viewing devices, while waveguides are generally placed horizontally. A common attempt at a solution has been to think in terms of traditional component integration and to assemble micro-lenses and micro-mirrors to collimate and bend the light path. The use of 45° micro-mirrors has been done repeatedly [16][17][18][19][21][23]; other techniques include the use of curved mirrors [20]. For optical interconnects to become a technology, the integration process must demonstrate optical alignment among hundreds of optical channels in a reasonably straightforward and routine manner. For the technology to become a product, the integration process has to become mass producible and demonstrate a reliable performance advantage at competitive cost and lower the power dissipation as much as possible with existing products.

This work investigates a novel direct integration process for VCSELs and optical waveguides. The approach is to assemble the VCSEL and PIN PD side-mounted on a substrate so that the light emission and viewing directions are both in the plane of the waveguide. The 90° turn is made electrically rather than optically, where losses are negligible. Recently, side-mounted coupling has also been investigated at IBM in 2006-7 [14][22] and NEC in 2007 [15]. One approach used in this work to create the waveguide for the VCSEL-side-mounted coupling is to place a liquid, photo-definable monomer between the two O/E components and form a monomer-component interface. Then, polymerization of the monomer film between the VCSEL and PIN PD is initiated with a stripe of UV light defined by a glass mask. The VCSEL, PD, and waveguide are thereby aligned in one exposure. Similarly, multiple parallel channels can be aligned optically and simultaneously in exposure, which is amiable for mass production. The finished result is that the light is directly injected into the waveguide and travels in the same plane of the waveguide, so no additional micro-components are needed in the light path, which

inevitably result in power loss. Another approach that could be implemented to couple the side-mounted VCSEL into the waveguide is to use a solid monomer film prior to exposure. It can be prepared by evaporating the solvent in a soft bake cycle after spinning the sample or by laminating a pre-made dry film monomer. The latter approach eliminates any accumulation of monomer on any kind of protuberating elements on the substrate during spinning, therefore creating a flat and even surface prior to exposure.

Chapter 3

Review of Previous Work

3.1 Ethernet Services Transport Protocol

TCP has evolved from its original version (introduced in 1981 [7]) to faster versions [9][10][11]. But even the fastest versions still pose latency problems, especially in high delay-bandwidth product networks. The previous work done in this field can be categorized in 3 types of protocols: 1) The first type is where the protocol is a additive-increase multiplicative-decrease (AIMD) variant of TCP, 2) has a different congestion control algorithm than AIMD, but packet format is still compatible with that of TCP, and 3) completely different protocol and it is not backward compatible with TCP. Most of the research in this field falls in the first and second category, the third is less studied do to its incompatibility. A protocol in the third category would need a forklift type upgrade, which is very costly and difficult.

An example of a protocol that falls in the first category that is better suited for high bandwidth-delay product networks is HighSpeed TCP, an adjusted version of Traditional TCP in which the AIMD values have been linearly scaled to increase the congestion window (*cwnd*) faster and decrease it less aggressively [9]. The main drawback is that these values are fixed (hardcoded) and if the network environment changes the protocol cannot adjust for this. Furthermore, if the network experiences high packet loss the protocol behaves exactly the same as Traditional TCP. This means improvement is only possible if the network itself is already healthy.

An example from the second category is a protocol called Scalable TCP, which has a multiplicative-increase algorithm [10], opposed to the more common additive-increase algorithm. The advantage that it offers is that it can reach the maximum available bandwidth much quicker than the additive-increase method, but this makes the algorithm less stable as it is aggressive in both directions. It also becomes bursty, and bursts are constrained by the UNI, as it will only allow up to an excess burst size (EBS) to go through, making this protocol unsuitable for CEN.

A protocol from the third category would be the Explicit Control TCP (XCP) which adds a congestion field in the packet format [11]. This novel idea produced good theoretical results, but in practice it meant that routers had to be reconfigured in order to be able to write congestion feedback data into the packets. Also, this made the packet incompatible with Traditional TCP and it would delay packet routing since each packet had to be decapsulated (in some cases twice for L2 routers) to reach the packet header. Many other protocols have been proposed, but most are AIMD variants aimed for specific kinds of networks [57][58][59]. The protocol proposed in this paper has a dynamic congestion-avoidance algorithm that only utilizes information available at layer 4 and the packet format is the same as Traditional TCP, hence making it compatible with the original version. Having the same packet format makes the protocol easily deployable. Furthermore, using only information from layer 4 makes it transparent to the CEN, which is one of the requirements imposed by the Metro Ethernet Forum (MEF) [75].

The proposed protocol falls in the first category, a variation of the AIMD parameters, but rather than have them fixed, they are dynamic. In control systems, this can be viewed as open loop or closed loop system, where our case would be a closed loop system because it gathers feedback, and it uses this feedback to change the variable, the variable is also affected by its

environment. The two inputs to the system are packet loss (environment) and last packet loss occurrence (feedback). Using these two variables the congestion can be computed.

3.2 Side-mounted VCSEL

To understand the reason for choosing the proposed architecture it is important to understand the physical limitations of metal mediums and optical mediums. With the decrease in size of transistors, switching is now achieved at greater speeds and lower power consumption. Metal mediums (commonly copper) are becoming the limiting factor of achievable bandwidth. The distance length and cross-section of the transmission line is one of the dominating reasons for this bandwidth constraint. By increasing the distance by a factor of L , it reduces the bandwidth by a factor of L^2 [13]. For this reason optical interconnects are appealing for chip-to-chip high data-rate performance applications. Metal mediums do have at least one advantage over optical mediums; and that is that the line can bend sharply (e.g. 90° turn) while paying a very small power penalty. Because board-level transmission lines are capacitance-limited there will be a loss due to the length, but the power loss due by the bend is very small. This occurs because electrical current behaves much like a fluid that is being pumped through a pipeline.

Optical mediums offer much greater bandwidth capabilities among many other advantages. It reduces or eliminates problems like crosstalk, electromagnetic absorption, wave reflection, and impedance matching. One problem that optical mediums have is that they have the difficult task of guiding light. Light does not behave like electrical current and its isotropic nature makes it difficult to conceal. Optical fibers and polymer waveguides have been successful at guiding light with low or acceptable power loss, but at the small scale of microelectronics it has been difficult to create optical devices such as mirrors and lenses; and even when successful,

the integration process has been a difficult task. Mirrors created by a 45° cut of the polymer waveguide are the simplest to build, but these come at the cost of a rather large power penalty. In applications using VCSELs, which are already low power devices, these mirrors have been the common solution to guide the light into the waveguide [25][26][27][28][29][30]. Other solutions for 90° bend also exist [31][32]. The two major characteristics of the mediums that are contrasted in this work are the losses caused by length versus the losses caused by guiding. Choosing an architecture that benefits from the metal medium's advantages and the optical medium's advantages will yield a better channel.

Side-mounted-VCSELs are not a new concept. IBM did work in 2006 [14] (see Figure 1) in which a VCSEL was placed in a flexible transmission line substrate and the substrate was gently bent 90° over some distance and connected to a vertically placed board. The vertical board had a clamping frame in which the VCSEL was mounted. The VCSEL was aimed at an optical waveguide integrated into the main horizontal substrate.

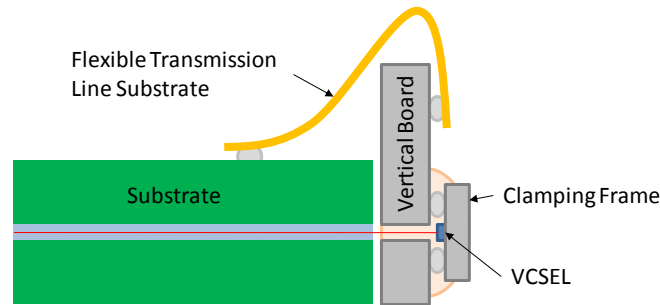


Figure 1. IBM schematic of side-mounted VCSEL.

Even though this setup was able to generate an open eye-diagram at 12.5 Gbps the construction is complex and can be further simplified. Also the electrical lines in the flexible substrate and vertical board could be eliminated with a simpler architecture.

In 2007, NEC also did work on side-mounted VCSELs [15] (see Figure 2) coupled to optical fibers instead of polymer waveguides. The reason for choosing to side-mount the VCSEL in this work was for the convenience of a side pluggable device. In this paper the main focus is on having the VCSELs side-mounted on a simple, small and inexpensive sub-anchor board transmitting light through a one-piece polymer waveguide with no mirrors, lenses or other optical aids.

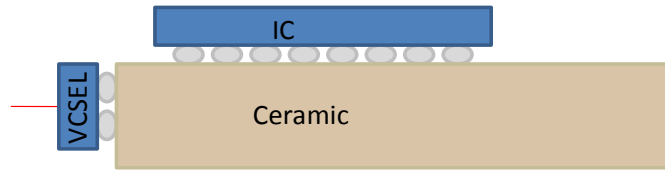


Figure 2. NEC schematic of side-mounted VCSEL.

Carrier Ethernet Network Solutions

4.1 Evolution of Carrier Ethernet Networks

Carrier Ethernet Networks have evolved immensely since the creation of Ethernet. Approximately every decade Ethernet technology increases in speed by 10 fold. In the last decade it has increase 100 fold. Ethernet technology shows no sign of deceleration. The next generation of Ethernet will be Carrier Ethernet, which is expected to have throughputs of 100 Gbps. This evolution projection is shown in Figure 3.

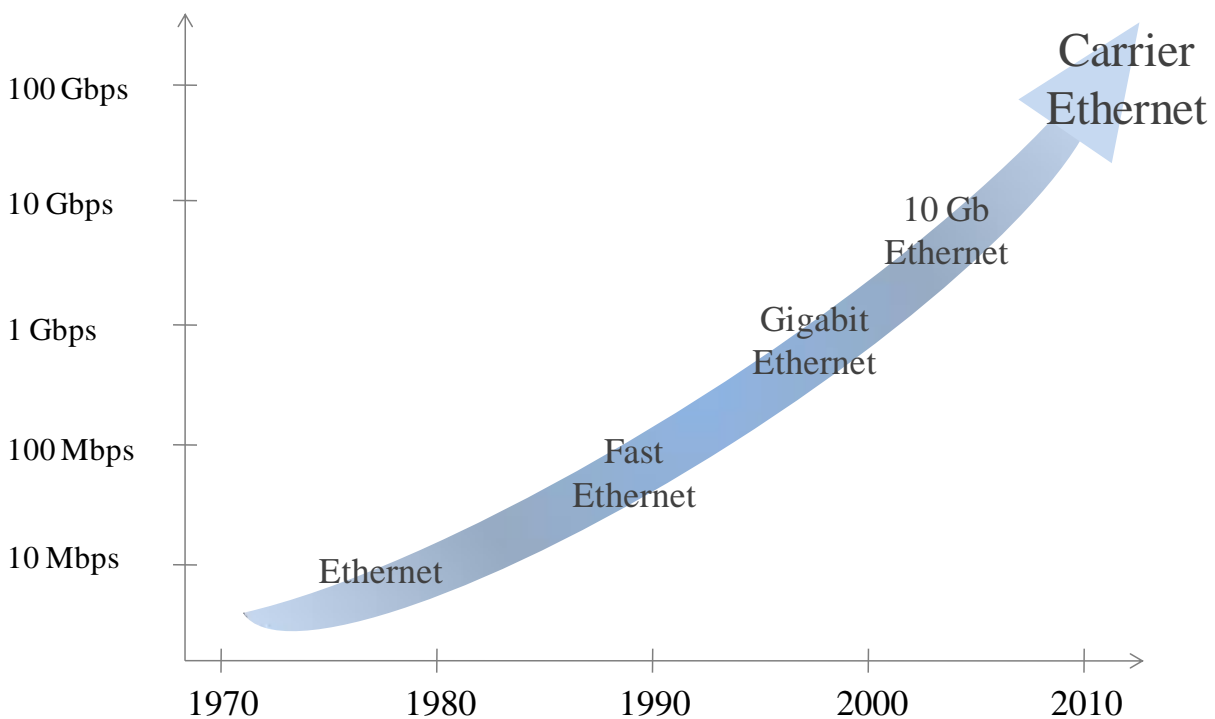


Figure 3. Evolution of Ethernet into Carrier Ethernet Networks

Because the evolution of Ethernet has been so successful it has caught the attention of many businesses, and slowly more and more companies are migrating their network structure to Carrier Ethernet.

4.2 Throughput Bottleneck and Solutions for Carrier Ethernet

The objective of this work is to provide bandwidth solutions to the Carrier Ethernet framework. Carrier Ethernet is rapidly becoming a ubiquitous and standardized service. Aiming at the top market sectors such as media, healthcare, finance, education, and government; and its services are becoming a necessity rather than a commodity. The Metro Ethernet Forum (MEF) is a global industry alliance that is setting the standards of Carrier Ethernet. Since 2004 the MEF has been releasing technical specifications that define the architecture and services of the carrier-grade Ethernet. Future MEF specifications are expected to have support for 100 GbE. To meet these bandwidth requirements, the next-generation Metro Ethernet network (MEN) will undergo hardware, and the less obvious, protocol changes. This work concentrates on two research areas, found in Layers 1 and 4 of the Open Systems Interconnection Basic Reference (OSI) model (see Figure 4), that improve and aid the bandwidth requirements: A transport protocol designed for Ethernet services and an optical interconnect component coupling technique. The reason for focusing on these areas is because they are known for having bandwidth limitations. At the transport layer, traditional TCP performs poorly in long-distance high-bandwidth environments because the bandwidth depends on transmission delay. At the physical layer, electrical interconnects pose bandwidth problems at high frequencies. Both problems are described in more detail ahead.

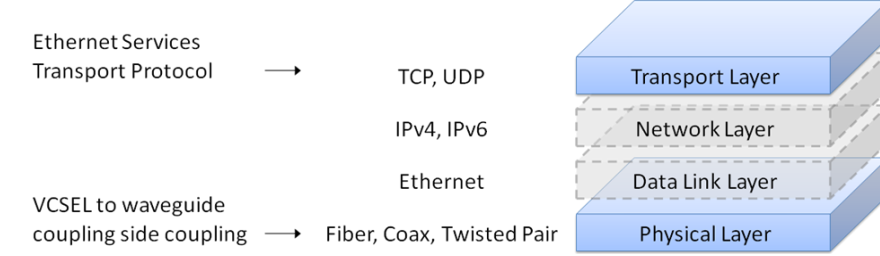


Figure 4. Research areas in which the proposed work focuses.

It is well known that traditional TCP poses bandwidth limitations for long-distance transmissions. TCP penalizes the congestion window aggressively with every packet loss incident and it increases it conservatively slowly (to avoid congestion), making it difficult to reach the maximum allowable bandwidth. With the growth of the carrier-grade Ethernet it is becoming essential to have a transport protocol that can take advantage of the full available bandwidth while simultaneously supporting the quality-of-service (QoS) mechanisms offered by Metro Ethernet services. For the past three decades, much work has been performed to make up for the shortcomings of traditional TCP. Most of this work has been based on tweaking the additive-increase multiplicative-decrease (AIMD) parameters. The Ethernet services transport protocol (ESTP) takes a different approach by making the AIMD parameters vary dynamically depending on the QoS bandwidth profile parameters specified by Ethernet services and the congestion feedback information that is available at the transport layer. The congestion feedback helps the transport protocol make decisions influenced by the network conditions, therefore increasing the utilization efficiency and the overall transmission throughput. It is demonstrated that a significant improvement is achieved over traditional TCP, and the increase can be achieved while maintaining fairness under different scheduling schemes.

Carrier Ethernet networks are designed to support bit rates of 10 Mbps, 100 Mbps, 1 Gbps, and 10 Gbps (future specification will support 100 Gbps). To be able to reach

transmission speeds of 100 Gbps it is necessary to use an optical backplane. It has been shown that electrical interconnects decrease bandwidth quadratically with increasing length and decrease bandwidth linearly with decreasing cross-sectional area due to inductive-capacitive (LC) lines [13]. An optical backplane will alleviate problems caused by electrical interconnects, such as clock distribution, system synchronization, crosstalk, voltage isolation, wave reflection, impedance matching, and pin inductance, which make high-bandwidth transmissions very difficult. The optical backplane will require the use of lasers, optical pipes, and detectors. The integration of these elements is not trivial, and improving the coupling efficiency is an ongoing research topic. A major difficulty has been the optical alignment of the three components, which is exacerbated by the fact that both a VCSEL and a standard PIN PD are surface emitting and viewing devices, while waveguides are generally placed horizontally. A common attempt at a solution has been to think in terms of traditional component integration and to assemble micro-lenses and micro-mirrors to collimate and bend the light path. For optical interconnects to become a technology, the integration process must demonstrate optical alignment among hundreds of optical channels in a reasonably straightforward and routine manner. For the technology to become a product, the integration process has to become mass producible and demonstrate a reliable performance advantage at competitive cost and lower power dissipation than is possible with existing products. This work experiments with a novel direct integration process for VCSELs. The approach is to assemble the VCSEL and PIN PD side-mounted on a substrate so that the light emission and viewing directions are both in the plane of the waveguide. The 90° turn is made electrically. A liquid, photo-definable monomer is placed between the two O/E components and forms a monomer-component interface. Polymerization of the monomer film between VCSEL and PIN PD is initiated with a stripe of UV light defined by a glass mask.

The finished result is that the light is directly injected into the waveguide and travels in the same plane of the waveguide, so no additional optical methods, which cause loss, are required to direct the light.

4.3 Carrier Ethernet Elements

Carrier Ethernet Networks are composed of three main elements: Customer Equipment (CE), User-network Interface (UNI), and Ethernet Virtual Connections (EVC), as shown in Figure 5. These elements are the minimum necessary to have an end-to-end Carrier Ethernet Network. There are also Network-network Interfaces (NNI), which connect Metro Ethernet networks with other Metro Ethernet networks. A brief description of these elements follows.

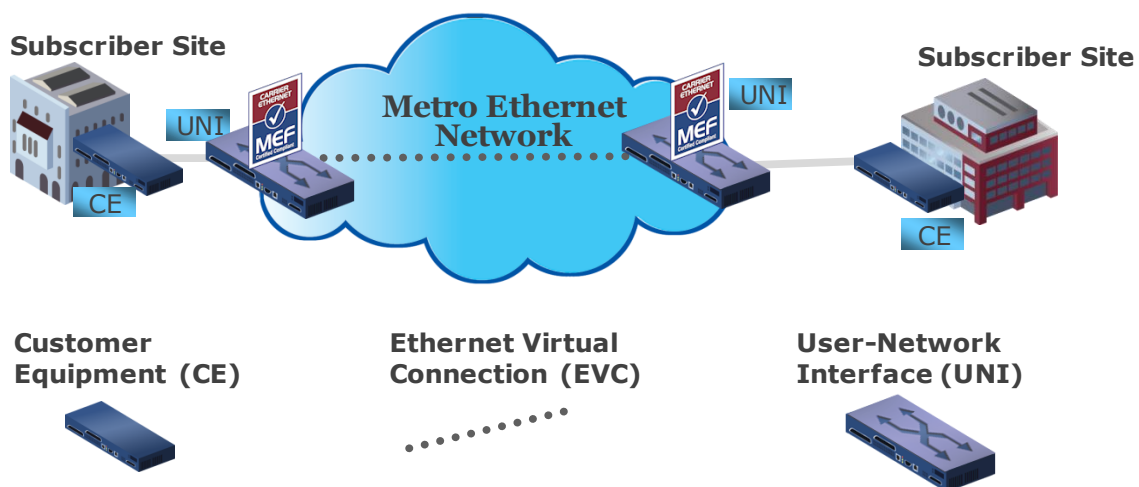


Figure 5. Main elements of Carrier Ethernet Networks

Customer Equipment: These are devices that are owned and managed by the customer and act as a gateway to the Carrier Ethernet network. They are always connected to the UNI.

User-network Interface: This is located at the edge of the Carrier Ethernet network, it is the ingress point of the network. This device has many responsibilities, including: open and terminate EVCs, police traffic from subscribers, maintain the traffic from different subscribers isolated from each other, perform QoS services, etc. There exist three types of UNI. Type I is a

simple device that performs the above functions. Type II performs all the tasks that type I can perform but it can additionally supply information to the customer equipment upon request. Type III can perform all the tasks type II can perform but it can additionally request information to the customer equipment, so the requests can be bidirectional. Because in this work it is necessary to have this bidirectional communication link, every instance in which UNI is mentioned in this work it is referred to UNI type III.

Ethernet Virtual Connection: It connects two UNIs via a virtual link. This link prevents information from one subscriber to reach another subscriber. One UNI can open multiple EVCs (if the subscriber has this service enabled).

4.4 Bandwidth Profile enforced by the User-network Interface

When traffic reaches the ingress point of Carrier Ethernet networks it means it is entering the UNI. The UNI upon receiving this traffic it will enforce a traffic shaping but with specific boundaries that are determined between the subscriber and the CEN provider; and these parameters are recorded in the service level agreement. Two of the most important traffic shaping parameters are the committed information rate (CIR) and the excess information rate (EIR). The CIR is the rate that is guaranteed to the subscriber. This amount of bandwidth should be available to the subscriber at all times. The excess information rate (EIR) is highest rate the subscriber is allowed to reach, any traffic flowing faster than this rate will be discarded. The most common technique to control the flow of traffic through the network is the use of leaky bucket algorithms. Because we have two rates that we like to manage a dual-leaky bucket algorithm is more appropriate. This process is shown in Figure 6.

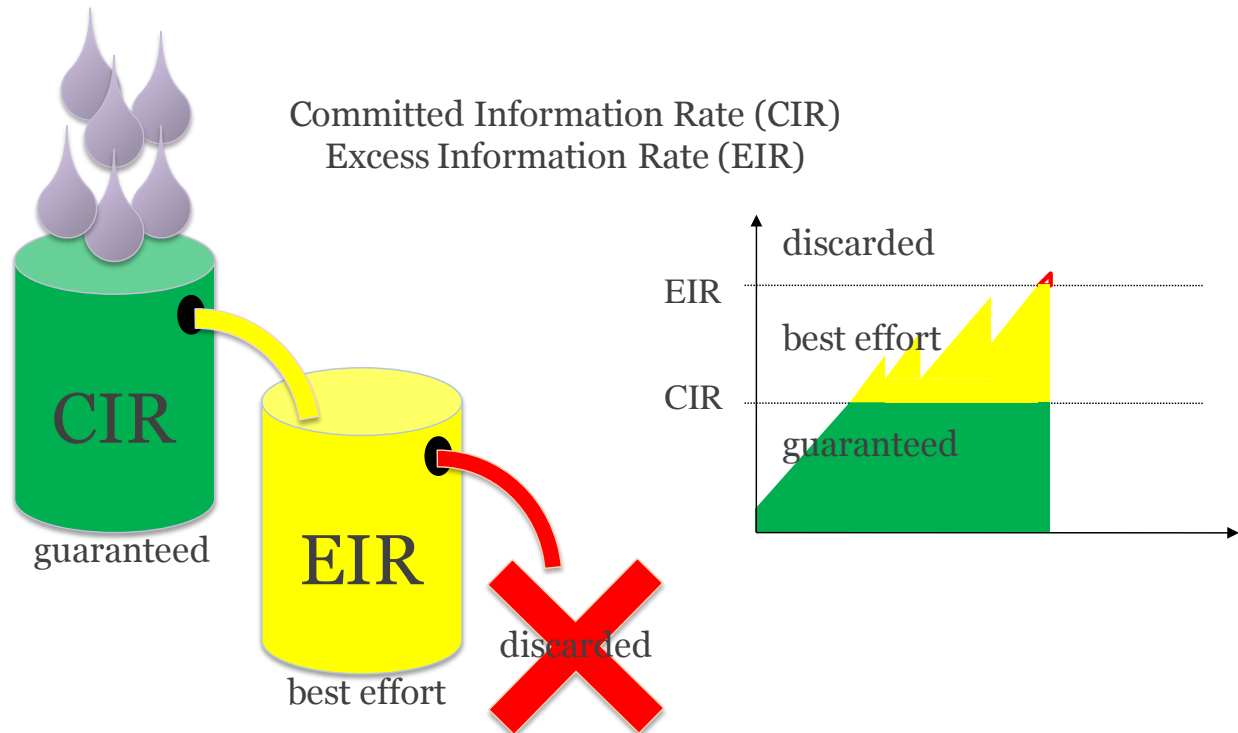


Figure 6. Bandwidth Profile enforced by the User-network Interface

When the traffic entered the UNI, it will go into the first leaky bucket algorithm. This algorithm will assign tokens to each packet that arrives, the amount of tokens are finite and controls the bust size. The tokens are generated at the CIR rate, so all the traffic that has a token is allowed to enter the network, but any traffic that does not have a token is moved to a second leaky bucket. This second leaky bucket generates tokens at a rate of EIR. All the packets that have tokens will be allowed to enter the network if there is no other CIR traffic from another subscriber waiting to enter the network (i.e. this traffic is treated in a best-effort manner). The traffic that does not have a token will be discarded.

This is just one of the tasks that the UNI performs and it is in no way limited to this process, but the details of the other tasks do not affect or concern the study performed in this work.

Chapter 5

Ethernet Services Transport Protocol (ESTP)

5.1 Overview

TCP was designed for best-effort packet-switched networks without previous knowledge about the 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; therefore, it is not possible to distinguish triple-duplicate ACKs – when three packets contain the same ACK value as the packet with the last ACK value of the sequence – caused by congestion from triple-duplicate-ACKs caused by random loss in a best-effort network. In Ethernet Services, the QoS parameters specified by the SLA offer valuable information. One of the goals of this work is to incorporate this information into the transport protocol to have better control of the data flow over a network that employs Ethernet Services, and as such the proposed protocol is referred to as the Ethernet-Services Transport Protocol (ESTP) in this work.

The guaranteed loss rate for CIR traffic can be used to estimate the congestion level of the network. However, the state of the network is an inherent property of the network, which is not directly known by the TCP sender. In this section, a traffic loss profile model is proposed that is used to capture the random nature of the network state. For each packet loss indicated by triple-duplicate-ACKs, the protocol will estimate the congestion level and will adjust the

congestion window size. In this work we consider packet loss information only, without any queuing model for the network or any additional functions incorporated in the router.

5.2 Traffic Loss Interval to Multiplicative-decrease Factor Mapping

The principle of this method is based on the estimation of the congestion level of the network. The level of congestion is related to the amount of successful packets delivered between two packet losses. The instantaneous amount of packets delivered between two packet losses is defined here with the parameter α . The value of α is mapped into an exponential profile to determine a less strenuous multiplicative-decrease factor defined as $map(\alpha)$. The purpose of choosing an exponential profile is to match the exponential probability distribution exhibited for the interval between two packet losses. By having the congestion window dependent on the network congestion level, the size of the congestion window can be controlled more efficiently.

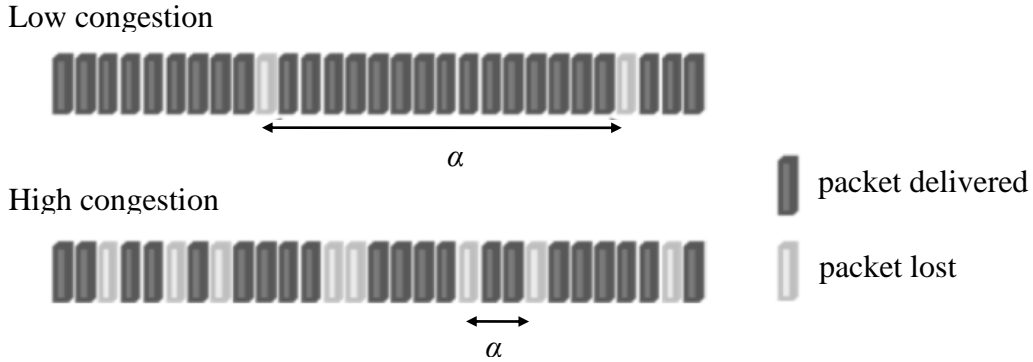


Figure 7. Value of α is the amount of successful packets transmitted in between two packet losses plus one.

The exact definition of α is the amount of successfully delivered packets found between two packets losses plus one. This can be easily seen by considering the metric distance to be one packet. Then, the distance between two lost packets is the amount of successful packets found in-between these two losses plus one (the lost packet). The value of α will range from 1 to ∞ , which implies that the extreme cases are two consecutive losses and a lossless transmission. The value

of α is mapped to a value of $map(\alpha)$, which ranges from 2 to 1. The reason for this range is because, in the event of a packet loss, when we divide the current congestion window size by the mapping value we do not want the new congestion window size to increase, hence the lower limit of this range is 1; to understand the reason for choosing an upper limit of 2, a little familiarity with Traditional TCP is required. In the event of a packet loss, Traditional TCP divides the current congestion window size by 2 to compute the new size. This means that dividing by any number greater than 2 could potentially have a performance lower than Traditional TCP, so an upper bound of 2 is chosen for this mapping range. This means that the performance of the new protocol is lower bounded by that of Traditional TCP.

By chosen this range this means that the proposed protocol can decrease the congestion window size by a factor no greater than two (inclusive), and no smaller than one (exclusive). Since a factor of 1 is selected for $\alpha = \infty$ (lossless transmission) it is obvious that the protocol will never choose the exact factor of $map(\alpha) = 1$. This mapping scheme is desirable because if α takes a small value, it is assumed to be the result of a highly congested network (Figure 7) and therefore the congestion window size is reduced more aggressively. Inversely, if the value of α is large, it is assumed that the network is not experiencing high levels of congestion and the congestion window is only slightly reduced in size. To achieve this mapping and at the same time match the traffic loss probability distribution, an exponential function is chosen. The boundary conditions are: $map(1) = 2$ and $map(\infty) = 1$.

To obtain the mapping function of ESTP it is necessary to familiarize with the probability density functions (pdf) of the period of time between two packet losses. The probability of a packet loss has a Bernoulli distribution. If we assume that a packet loss is a successful event, then the number of successful events in a fixed period of time has a Poisson distribution, i.e. the

packet loss rate can be modeled as a Poisson distribution. The time between two successful events can then be modeled as an exponential distributed process. The metric unit does not need to be time to follow the Poisson distribution; it can also be distance, area, volume or packets. Time is the most common unit, such that the number of successful events per unit time is usually called arrivals.

The pdf of an exponential distribution is:

$$f(\alpha; \tau_\alpha) = \begin{cases} \frac{1}{\tau_\alpha} e^{-\frac{\alpha}{\tau_\alpha}} & \alpha \geq 0 \\ 0 & \alpha < 0 \end{cases} \quad (5.1)$$

where, $\tau_\alpha = E[\alpha]$.

It is desirable to match the mapping function to the pdf of the distance between two events, since $\exp(\infty) = 0$ then one has to be added to the exponential term. This yields $map(\alpha) = \exp(-\alpha/\tau) + 1$. To satisfy $map(1) = 2$ we need to shift the expression by 1, which yields $map(\alpha) = \exp[-(\alpha-1)/\tau] + 1$.

The mapping function (see Figure 8) is then derived to be:

$$map(\alpha) = e^{-\frac{\alpha-1}{\tau_\alpha}} + 1 \quad (5.2)$$

The *cwnd* expression is then:

$$cwnd_{n+1} = \frac{cwnd_n}{\left(e^{-\frac{\alpha-1}{\tau}} + 1 \right)} \quad (5.3)$$

This excludes the effects of $cwnd_{MIN}$, which is explained in chapter 5.

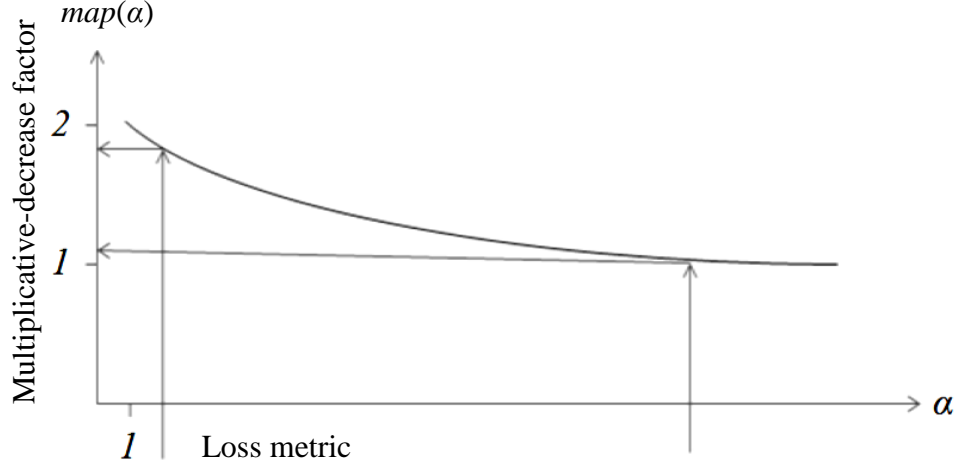
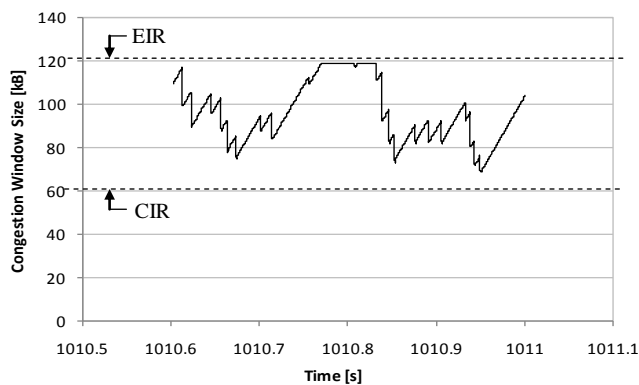


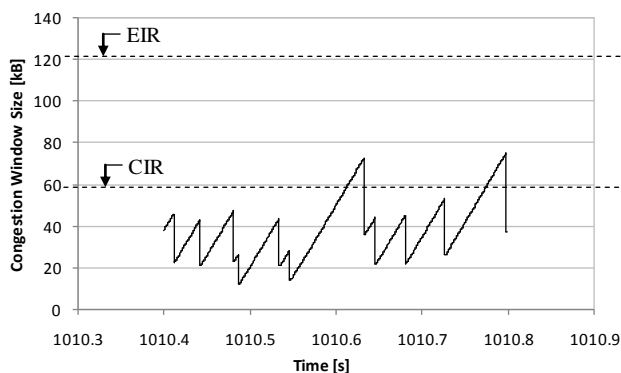
Figure 8. Proposed protocol packet loss interval to multiplicative decrease factor mapping.

If the committed packet loss rate is called β , then τ is chosen as β^{-1} , which is the average value of α . This scheme is intended to behave in the following manner: If $\alpha \ll \tau$, a relatively aggressive reduction to the congestion window is enforced. This will cause the congestion window to reduce its size drastically, therefore reducing the throughput and the possibility of causing congestion. As a result, the value of α will tend to increase. If $\alpha \gg \tau$, a lenient congestion-window reduction occurs and the throughput is mildly affected. As the throughput increases so will the possibility of congesting the system. As a result, the value of α will tend to decrease. The overall effect is that α will vary around the value of τ for moderately congested systems. This matching is the mechanism that provides the loss guarantee specified by the SLA. Ideally, the value of $E[\alpha]$ is closely matched to the committed packet loss rate. This behavior can be observed in Figure 9(a) (CIR and EIR are explained in the next sub-section), and the contrasting behavior of Traditional TCP can be observed in Figure 9(b). Notice that for the extreme case of two consecutive packet losses ($\alpha = 1$), the proposed protocol behaves like the SLA-aware protocol [1], which means that the proposed protocol is lower-bounded by the SLA-

aware throughput performance. If the SLA information is not taken into account, such that $cwnd_{MIN} = 0$, then for $\alpha = 1$ the proposed protocol behaves like Traditional TCP.



(a)



(b)

Figure 9. Congestion window behavior of (a) ESTP and (b) of Traditional TCP.

5.3 Congestion Control Obtained by Combining Congestion Feedback with Ethernet

Services Information

By incorporating Ethernet Services information to the mechanism just described (effective use of available band-width by mapping packet loss interval to a less strenuous multiplicative factor) the proposed protocol can further show improvement in terms of

throughput performance, since these techniques are independent of each other. Combining the congestion control mechanisms discussed in the previous subsection with the SLA-aware mechanisms, the following multiplicative decrease expression can be derived for congestion control:

$$\begin{aligned}
 cwnd_{n+1} &= \frac{cwnd_n - cwnd_{MIN}}{map(\alpha)} + cwnd_{MIN} \\
 cwnd_{n+1} &= (cwnd_n - cwnd_{MIN}) \left(e^{-\frac{\alpha-1}{\tau}} + 1 \right)^{-1} + cwnd_{MIN}
 \end{aligned} \tag{5.4}$$

This technique takes full advantage of the bandwidth provided to the subscriber by maintaining the throughput above the CIR (see Figure 9(a)). The throughput can be further improved by utilizing the EIR information. $cwnd_{MAX}$ can be obtained similarly to $cwnd_{MIN}$ but using EIR rather than CIR: $cwnd_{MAX} = RTT \cdot EIR$. With this value, the upper bound of the congestion window can be controlled. An upper bound is set because once the $cwnd$ exceeds $cwnd_{MAX}$, the throughput will exceed EIR and the traffic policing enforced at the lower layers will discard packets exceeding this rate. Once that happens, the protocol at layer 4 will initiate its congestion control mechanism and decrease the congestion window. This is unnecessary since there is no congestion in the network. By maintaining the $cwnd$ at its maximum value, the throughput will not be reduced by the protocol's congestion control and the subscriber will get maximum throughput until a random loss (e.g. link loss) or congestion loss occurs.

The additive increase properties of AIMD of the proposed protocol will then be expressed as:

$$cwnd_{n+1} = \min(cwnd_{MAX}, cwnd_n + 1 / cwnd_n) \tag{5.5}$$

Traditional TCP is then a special case when $cwnd_{MAX} = \infty$.

The traffic-loss profile matching and the $cwnd$ boundary mechanism implemented by the additive-increase- multiplicative-decrease (AIMD) techniques distinguishes the proposed protocol from other transport protocols and make its implementation desirable for Ethernet Services. Simulation results show the successful operation of congestion feedback control by using SLA information. This is demonstrated by the congestion window size control (see Figure 9(b)).

5.4 Analytical Expression for ESTP Throughput

To derive the analytical expression of ESTP's throughput, it is necessary to have the general throughput expression in terms of the additive increase (γ) and multiplicative decrease (β). This is derived in [57] and found to be:

$$T_{\gamma,\beta}(p, RTT, T_o, b) \approx \frac{1}{RTT \sqrt{\frac{2b(1-\beta)}{\gamma(1+\beta)}} p + T_o \min \left(1, 3 \sqrt{\frac{(1-\beta^2)b}{2\gamma}} p \right) p(1+32p^2)} \quad (5.6)$$

The right term of the denominator is due to packet losses detected by timeout, rather than triple-duplicate-ACKs. Since we are maintaining a congestion window large enough to maintain a CIR rate, it is a fair assumption that most, if not all, packet losses will be detected by triple-duplicate-ACKs and neglecting this right term of the denominator would be a more accurate model. The only circumstance in which this term would have more of an impact would be if T_o is chosen to be in the millisecond range, in which case this term would be $\log_{10}(1/p)$ orders of

magnitude smaller than that on the left. If $p = 0.001$, the right term is 3 orders of magnitude smaller and can be neglected. For both cases we have:

$$T_{\gamma,\beta}(p, RTT, b) = \frac{MSS}{RTT \sqrt{\frac{2b(1-\beta)}{\gamma(1+\beta)}} p} \quad (5.7)$$

which multiplied by the maximum segment size (MSS), which is specified in bits, yields the expression in bits per second (bps), rather than packets per second.

To find the throughput considering the committed information rate (CIR) it is necessary to define new terms. The packet loss rate when the throughput is lower than the CIR is p_1 , and when the throughput is higher than CIR the packet loss rate is p_2 . The throughput can be written as:

$$T_{ESTP} = CIR \cdot (1 - p_1) + T_{BE}(p_2) \quad (5.8)$$

$$T_{ESTP} = CIR \cdot (1 - p_1) + \frac{MSS}{RTT \sqrt{\frac{2b(1-\beta)}{\gamma(1+\beta)}} p_2} \quad (5.9)$$

Since a transmission rate of CIR is reserved for the user, the packet loss rate p_1 can only be caused by intrinsic loss. For rates higher than CIR the data rate is not guaranteed and treated as best effort traffic, so p_2 is caused by intrinsic and extrinsic (e.g. congestion) losses.

In ESTP traffic the value of β is not fixed and depends on the metric distance between two packet losses, which we defined as α .

$$\beta(\alpha) = \frac{1}{map(\alpha)} = \left(e^{\frac{\alpha-1}{\tau}} + 1 \right)^{-1} \quad (5.10)$$

Therefore, the throughput T_{ESTP} also depends on α .

$$T_{ESTP}(\alpha) = CIR \cdot (1 - p_1) + \frac{MSS}{RTT \sqrt{\frac{2b(1 - \beta(\alpha))}{\gamma(1 + \beta(\alpha))}} p_2} \quad (5.11)$$

Because α is a random variable with exponential distribution to find the expected value of T_{ESTP} we need to convolute the deterministic function of T_{ESTP} in terms of α with the pdf of α .

$$T_{ESTP} = E[T_{ESTP}(\alpha)] = \int_0^{\infty} T_{ESTP}(\alpha) \cdot f(\alpha; \tau_{\alpha}) d\alpha \quad (5.12)$$

Since the throughput of ESTP is upper bounded by EIR the complete expression is:

$$T_{ESTP} = \min \left(EIR, \int_0^{\infty} T_{ESTP}(\alpha) \cdot f(\alpha; \tau_{\alpha}) d\alpha \right) \quad (5.13)$$

5.5 Computational Overhead Raw Comparison between the Proposed Protocol and Traditional TCP

Computational overhead is always a concern when protocols gather feedback from the system while utilizing it to improve the throughput. Doing a comprehensive analysis of the computational overhead is beyond the scope of this work; nonetheless, it is desirable to have a measure of the additional overhead employed by the proposed protocol when compared to the Traditional TCP. ESTP includes new variables to the protocol such as $map(\alpha)$, α , τ , $cwnd_{MAX}$, $cwnd_{MIN}$, CIR and EIR .

In the event that a packet is delivered successfully, Traditional TCP reads the *cwnd* value, it inverts it, and it adds the results to itself and overwrites it into *cwnd*. ESTP adds an extra operation were it compares the *cwnd* size with *cwnd_{MAX}* before overwriting the new *cwnd* variable. Assuming each operation takes the same amount of time, which is a lenient assumption, ESTP adds an extra 25% computational overhead. In the event of a packet loss Traditional TCP reads *cwnd*, divides it by 2 and overwrites the *cwnd* variable. ESTP computes α , which takes 2 operations, then computes $\text{map}(\alpha)$, which takes 4 operations, subtracts *cwnd_{MIN}* to *cwnd*, divides by $\text{map}(\alpha)$ and adds *cwnd_{MIN}* to total, which have 2 new operations over TCP. All together it adds 8 extra steps, increasing the computational by nearly 300% in the event of a packet loss. There are other processes involved in ESTP, but these can be performed during idle times. To compute the overall additional computational overhead we can multiply the additional computational overhead of the specific event by the frequency of the event. For the packet loss computational overhead to have a considerable effect on the overall estimation it would need to have at least 1/10 of the frequency of a successfully delivered packet. In reality it is much lower, which means that the overall additional computational overhead of ESTP over TCP is of approximately 25%. To put this in perspective, if the average processing time per event is 10ns, this average will be increased to 13ns.

5.6 ESTP backward compatibility with Traditional TCP and TCP-friendliness

5.6.1 Backward compatibility

The two main methods of designing a new protocol are to start from a base protocol and alter its characteristics or to start from no specific point and designing everything in a new way, which might not resemble any previous work. Both methods have advantages and disadvantages. When designing a transport layer protocol that is connection-oriented it is very important,

perhaps even required, to base it on Traditional TCP. 90% of internet traffic is TCP and creating a protocol that is incompatible with TCP will require a forklift type upgrade, meaning all end systems will require an upgrade before they are operational. For this reason ESTP was based on TCP, preserving the packet format. This makes ESTP backward compatible with TCP. This implies that an end system running ESTP can send a file to an end system running TCP and is able to utilize the advantages of ESTP. In the reserve scenario, in which an end system is running TCP is sending a file to an end system running ESTP, the transmission speed will be that of TCP. The system acting as a server is the system that can utilize ESTP's high throughput speeds. This is an advantage in the practical sense because in general servers act in one way or another as data storage units, may this be email, http files, file storage, etc., and end systems usually request data from the servers. What this means is that if only the servers had an upgrade from Traditional TCP to ESTP this would increase the overall throughput of the network significantly. This compatibility also allows a much smoother transition, by gradually upgrading the systems to the new transport layer protocol, rather than upgrading the entire network at once.

5.6.2 TCP-friendliness

Another advantage of having ESTP based on Traditional TCP is that it will inherit the TCP-friendliness properties of Traditional TCP. A TCP-friendly protocol is one that will adjust its transmission rate based on the congestion level of the system. Algorithms such as UDP base their transmission rates on the needs of the application, rather than on the available bandwidth. This affects neighboring TCP connections because these have congestion control algorithms and are more "conscientious" to its neighbors and if many transmissions are initiated at one time all TCP connections will reduce their throughputs if congestion occurs, but protocol lacking congestion detection cannot adjust to this situation. ESTP has a novel congestion control

algorithm that is TCP-friendly as it will adjust its transmission rate causing less impact on TCP transmission, yet it performs much more efficiently under low-congestion conditions.

5.7 Bandwidth Profile Translation Protocol

The bandwidth profile, as defined by the MEF, is provided in a per-UNI, per-EVC or per-CoS basis. The work done in [1] assumes that the CIR and EIR values are readily available, but in practice the MEF does not provide bandwidth profile parameters for that resolution level, i.e., per-TCP-connection. Doing so would be impractical and would consume valuable processing time, as there could be thousands of transport layer connections running simultaneously. To alleviate this load, a simple translation protocol can be placed on the LAN components that are connected to the UNI, such that the UNI can dynamically inform the end systems of what parameters should be employed to utilize the available bandwidth more efficiently.

The translation protocol should perform the following steps. Initially, it needs to detect the number of transport layer connections that are running at a specific point in time. This also includes UDP besides ESTP connections (it is assumed that all TCP connections are replaced by ESTP). The following assumptions are made to derive the translation equation: 1) The only types of transport layer protocols used are UDP and ESTP. 2) UDP sessions have a fixed data rate. 3) ESTP sessions have a variable data rate. 4) UDP data rate is much less than the CIR of the UNI, such that thousands of UDP sessions will not exhaust the CIR. The first assumption could be extended to other transport layer protocols if the data rate requirements were known. The second assumption could also be extended to a variable UDP data rate, but it would still need a lower bound, as this protocol is used for applications such as VoIP, in which a low data rate could greatly degrade the quality of the output signal.

Finding the number of opened connections is a trivial problem that can be done in several ways, so it is a design preference. If it is desired to avoid using the UNI to compute this, it can be done by simply programming the ESTP to increase the number of local open connections by 1 when a child process is invoked and to decrease it by 1 when a child process is closed. A child process is opened by the main process of ESTP (as is by TCP). Each child process manages the parameters of a single ESTP connection, which makes a trivial task to compute the total number of active connections running in a particular end system. Even if the connection terminates unexpectedly the child process must go through the closing state to terminate the connection properly. This local value, which contains the amount of open connections, can be sent periodically to the UNI or a proxy node, such as the CE. If it sends it to the CE, the CE can add the values of all the end systems before sending to the UNI. The UNI has the bandwidth profile information, so once all the UDP and ESTP sessions are accounted for the UNI quickly performs the simple computation shown below.

$$CIR_{ESTP} = \frac{CIR_{UNI} - N_{UDP} BW_{UDP}}{N_{ESTP}} \quad (5.14)$$

where,

CIR_{UNI} – EIR per-UNI as established in the SLA

CIR_{ESTP} – translated EIR value for ESTP

N_{UDP} – number of UDP sessions

BW_{UDP} – fixed data rate for single UDP session

N_{ESTP} – number of ESTP sessions

Similarly,

$$EIR_{ESTP} = \frac{EIR_{UNI} - N_{UDP} BW_{UDP}}{N_{ESTP}} \quad (5.15)$$

where,

EIR_{UNI} – EIR per-UNI as established in the SLA

EIR_{ESTP} – translated EIR value for ESTP

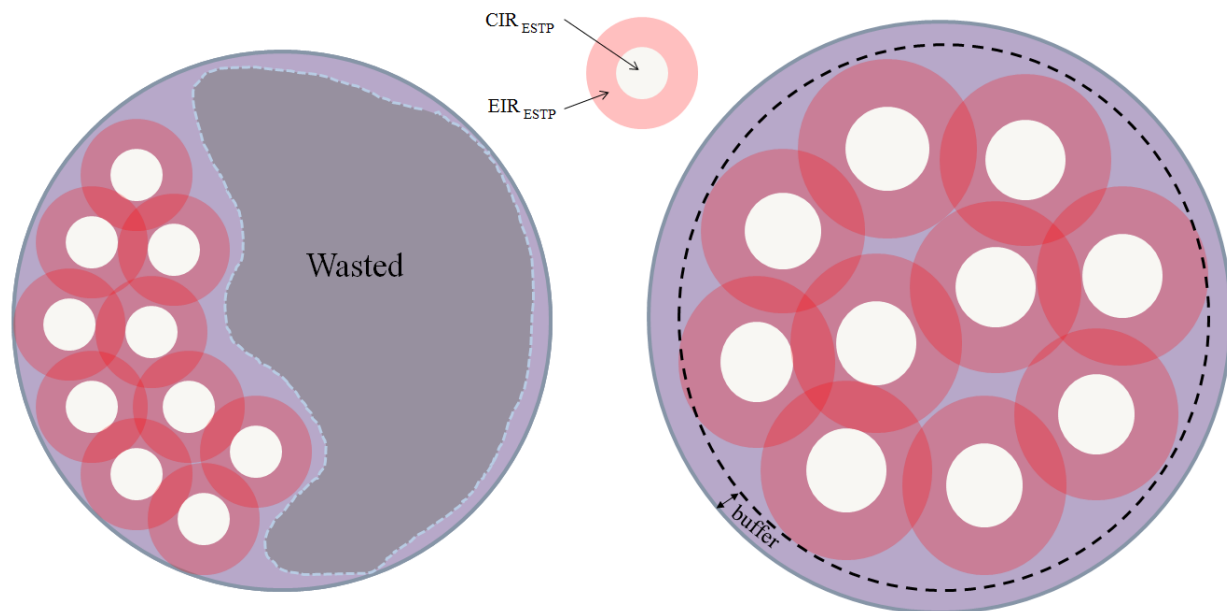


Figure 10. (left) graphical display of bandwidth with no dynamic control of CIR_{ESTP} and EIR_{ESTP} .
(right) graphical display of dynamic resizing of CIR_{ESTP} and EIR_{ESTP} .

These CIR_{ESTP} and EIR_{ESTP} parameters are the foundation of the protocol. Without this feedback control the sizes would need to be hardcoded and depending on the amount of open connections at one time the bandwidth could be underutilized, as shown on the left side of Figure 10. If the CIR_{ESTP} and EIR_{ESTP} were dynamically controlled by the translation protocol then the sizes of the individual connections would be such that they efficiently fill the available bandwidth, as shown on the right side of Figure 10. The buffer show is to alleviate abrupt changes in the number of active connections. This is described in more detail in Chapter 7.

Once the values for CIR_{ESTP} and EIR_{ESTP} are computed, the UNI sends them to the CE, which takes care of distributing this information to the end systems. Notice that the number of connections will change before the new CIR and EIR values are computed and returned, which

will cause the UNI to be underutilized or overutilized for small periods of time. This is discussed in more detail in the results section. It should also be noted that this expression is for per-UNI bandwidth profile, but the expression for the per-EVC and per-CoS cases can be easily derived by identifying the traffic that enters each traffic policer and using an expression similar to the expression above.

5.8 UNI Scheduling Scheme

There are several ways to perform the scheduling scheme at the UNI, but we propose a simple scheme that will add a relatively small computational overhead, while following all the specifications of the MEF bandwidth profile.

The proposed scheduling scheme is as follows: The traffic enters the UNI and it is inserted into a leaky bucket with rate CIR and burst size CBS. Each priority level will be assigned its own value for CIR and CBS. The traffic that meets the rate limit will be forwarded to the weighted fair queuing (WFQ) scheduler; any excess will be forwarded to a second leaky bucket with rate EIR. The traffic that is within the EIR will be sent to a FIFO scheduler, while the excess data will be discarded. The output of the WFQ scheduler will then be sent to the highest priority port of a strict priority queuing (SPQ) scheduler, while the output of the FIFO scheduler is sent to the second level of the SPQ. Figure 11 shows a graphical representation of this scheme.

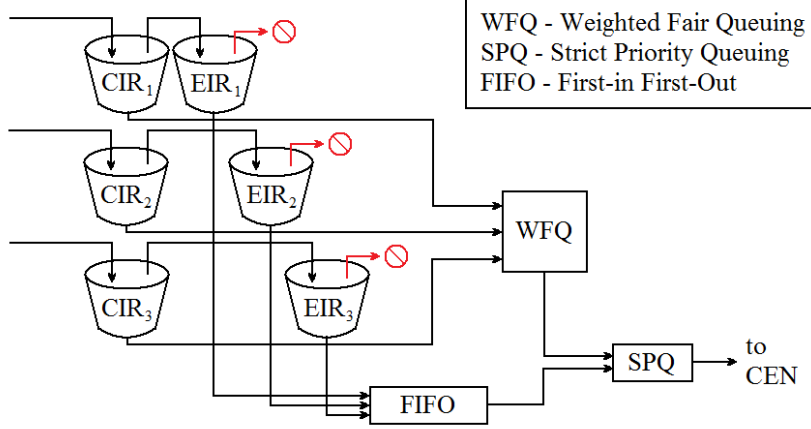


Figure 11. Proposed UNI Scheduling Scheme.

It could be argued that instead of a FIFO scheduler, another WFQ scheduler could be used to maintain fairness even on the best-effort traffic. This is not necessary to satisfy the requirements and would cause further computational overhead. The WFQ scheduler that is already in place is necessary because fairness is required and it is known that WFQ will guarantee a traffic flow that is proportional to the weight assignment even to the lowest priority level. The weights can be calculated as follows.

$$\Phi_k = \frac{EIR_k}{\sum_i EIR_i} \quad (5.16)$$

It is best to calculate the weights using the EIR values, rather than the CIR, because in a low packet loss network the individual throughputs are expected to be very close to their respective EIR. If

$$\frac{CIR_x}{EIR_x} \neq \frac{CIR_y}{EIR_y} ; x \neq y \quad (5.17)$$

then the weights are mismatched and one or more traffic flows could experience packet loss due to the surpassing of EIR. ESTP is designed to increase the *cwnd* up to *cwnd_{MAX}*, which depends of EIR. At the SPQ, the guarantee data will be serviced as soon as it arrives to the scheduler, and the SPQ will service the best-effort traffic only when the committed data queue is empty.

Chapter 6

Side-mounted VCSEL

6.1 Overview

Optical interconnects are recently attracting more attention in the areas of high performance computing and data transport technology. An optical interconnect that is designed to serve these purposes has three components: 1) a laser, generally a VCSEL, 2) a photo-detector, and 3) an optical waveguide pipe connecting the two. When assembled, these three components constitute an optical channel. In the optical channel, the optical waveguide light pipe can be constructed in two ways: 1) optical fibers, which are passive, require component integration, and in general are used for long distances (meters or kilometers); or 2) separately fabricated polymer waveguides, which are also passive and also require component integration, but in general are used for short distance transmission (few centimeters). Photo-definable optical polymers offer advantages that can be very useful for direct optical integration and direct optical alignment. By direct optical integration it is meant that the waveguide light pipe forms an interface with the laser at one end and an interface with the detector at the other. By direct optical alignment it is meant that the waveguide core is defined by exposing the monomer between laser and detector to a stripe of light that initializes polymerization.

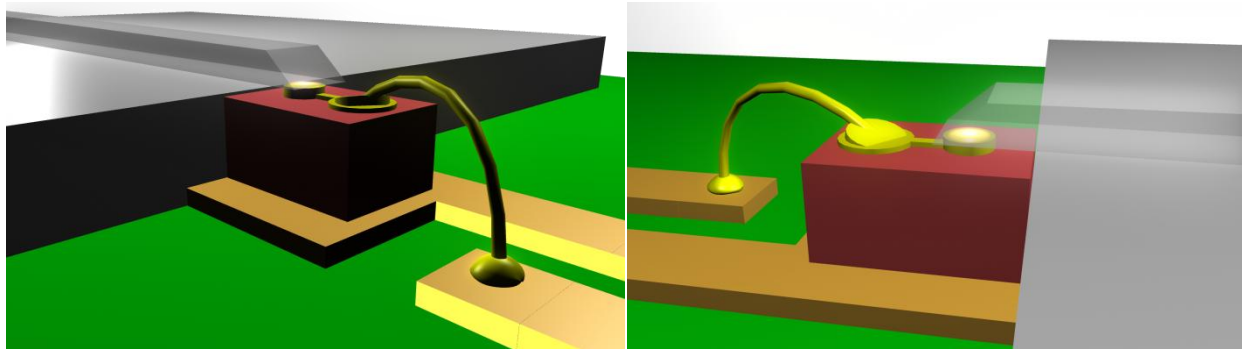


Figure 12. Conventional architecture for VCSEL-waveguide coupling.

The continuing challenge over the past ten years has been to integrate these three components: laser, waveguide, and photo-detector. VCSEL technology has been an attractive solution in optical communication for the low-power consumption property that this component provides. One major drawback of coupling VCSELs to waveguides is that VCSELs emit light from a cavity located on their top surface, hence the name, while the waveguide lies horizontally on the substrate; coupling the light into the waveguide will require a 90° turn of the laser beam. Coupling this emitted light into a polymer waveguide is a perplexing challenge. A major difficulty has been the optical alignment of the three components which is exacerbated by the fact that both a VCSEL and a standard PIN PD are surface emitting and viewing devices while waveguides are generally placed parallel to the substrate. A common attempt at a solution has been to think in terms of traditional component integration and to construct and/or assemble micro-lenses and micro-mirrors to collimate and bend the light path. Many approaches have been taken, the most popular one being using 45° mirrors (see Figure 12). While this method can produce good results, its process can be costly and/or cause great optical loss. The component-to-waveguide coupling is an ongoing research topic. Because the current architecture has low efficiency, it discourages the industry from employing optical interconnect technology. For optical interconnects to be a practical manufacturable technology, the integration process must

demonstrate optical alignment among hundreds of optical channels in a reasonably straight forward and routine manner. For the technology to become a product the integration process has to become mass producible and demonstrate a reliable performance advantage at more competitive costs and greater power efficiency.

6.2 Advantages of Optical Interconnect over CMOS Technology

It is important to mention the reasons that optical interconnection is preferred over CMOS technology (electrical interconnections). The three main reasons that are presented and discussed (but not limited to these) are: bandwidth, wiring density and power efficiency [61]. These reasons are correlated, but we will discuss them separately.

6.2.1 Bandwidth

The bandwidth of CMOS technology (specifically copper) is intrinsically and extrinsically limited. Some of the intrinsic reasons that cause this bandwidth limitation include: inductance, capacitance, skin effect, electromagnetic-radiation absorbance; and one of the extrinsic reasons is dielectric susceptibility. Because of these limitations, the longer the distance that the signal must travel through this medium, the lower is the bandwidth that it is capable of carrying. The bandwidth is inversely proportional to distance squared (as mentioned in section 3.2), and even after a few centimeters we drop to the megahertz range. D.A.B. Miller [10] quantifies this drop with the following expression: $B \cong 10^{16} \frac{A}{l^2} \text{ bits/s}$. To get a better feeling of what this implies a few reasonable parameters are substituted into this equation to give a grasp of the effects of increasing the distance. Taking a cross-sectional area of 10 microns by 10 microns and a length of 5 centimeters we get a maximum bandwidth of 400 Mbps. For the same cross-sectional area but with a length of 2.5 centimeters we obtain 1.6 Gbps bandwidth. This example shows clearly the effects that copper transmission lines have over the bandwidth.

In contrast with CMOS technology we have optical interconnect technology, where the throughput is limited mainly by how fast the laser source can switch, and not by the bandwidth. The bandwidth of the optical medium is much larger than that of the copper medium. There are factors that could degrade the performance of optical channels, such as impurities found in the medium or optical alignment issues, which could affect the bandwidth. So there is a fundamental tradeoff between complexity and bandwidth.

6.2.2 Wire Density

As mentioned earlier, these advantages are correlated. One method to improve the bandwidth in copper transmission lines is to design the wiring thicker. This will reduce the resistance of the medium. This solution has a costly tradeoff; by increasing the thickness of the copper transmission lines the device has a higher concentration of wires which create more weight, more crosstalk, and will require more real estate. Weight and space are issues concerning the marketability of the product; in general, it is desirable to maintain the product lightweight and small. One example of applications where the weight is an important issue (at large scale) is in the aviation industry [62], where planes want to upgrade the copper wiring to fiber optic cables; not just because they need the bandwidth, but because it will reduce the weight of the vehicle significantly. The new weight cannot be estimated by simply taking the current amount of copper wiring weight and multiplying by the glass to copper weight ratio because one fiber optic cable can transport the same amount of information than many copper cables through multiplexing techniques. On a smaller scale the weight might not be such a critical issue, but crosstalk will have a bigger influence. Signals sent over adjacent parallel copper transmission lines could interfere with each other if the signal strength was increased to overcome the distance loss, so either solution is degrading system performance. The crosstalk between polymer waveguides is

very small, as seen in section 9.4, at this separation the amount of radiation that leaks from one channel to another is almost 3 orders of magnitude smaller than the amount of radiation emitted by the output of the same channel. This allows for a high wire density, mainly limited by the degree of waveguide miniaturization that can be achieved during the fabrication process and the width of the E/O components.

6.2.3 Power Efficiency

The power efficiency is an important issue, for various reasons. One reason is trivial, it is desirable to administer smartly the power budget so that there is enough power at the receiver end to get good signal-to-noise ratio. Power efficiency is one of the main factors that determine the performance of the system. In CMOS technology, the power source is usually increased if the desired throughput is higher. This is done to overcome the intrinsic losses of the lines that arise as the frequency is increased. Also, besides the power used to counteract the intrinsic losses there will be other power sources that will consume more power if the throughput is increased. The system may require equalizers or other components that improve the signal recovery (such as noise reduction) that will also consume part of the power budget. Another reason is that power efficient systems can have a low-power source, which generates less heat. Heat can be detrimental to the system performance, and cooling systems get more expensive as the temperature emitted by the electrical devices increases. A very simple example is a CMOS chip been cooled by a fan; if the power is raised to overcome the intrinsic losses of copper transmission lines to enable higher throughputs then instead of a fan, perhaps it will be necessary to install a water cooling system. This upgrade can be very expensive since there are many considerations that need to be taken in account.

6.3 Architectural Design

A feasible alternative to using 45° mirrors, or other light bending techniques, is to perform this 90° bend in the electrical domain rather than in the optical domain. Bending in the electrical domain will cause a negligible loss of signal in the electrical domain while improving greatly the coupling in the optical domain, which improves significantly the overall efficiency of the end-to-end transmission system. The proposed architecture is shown in Figure 13.

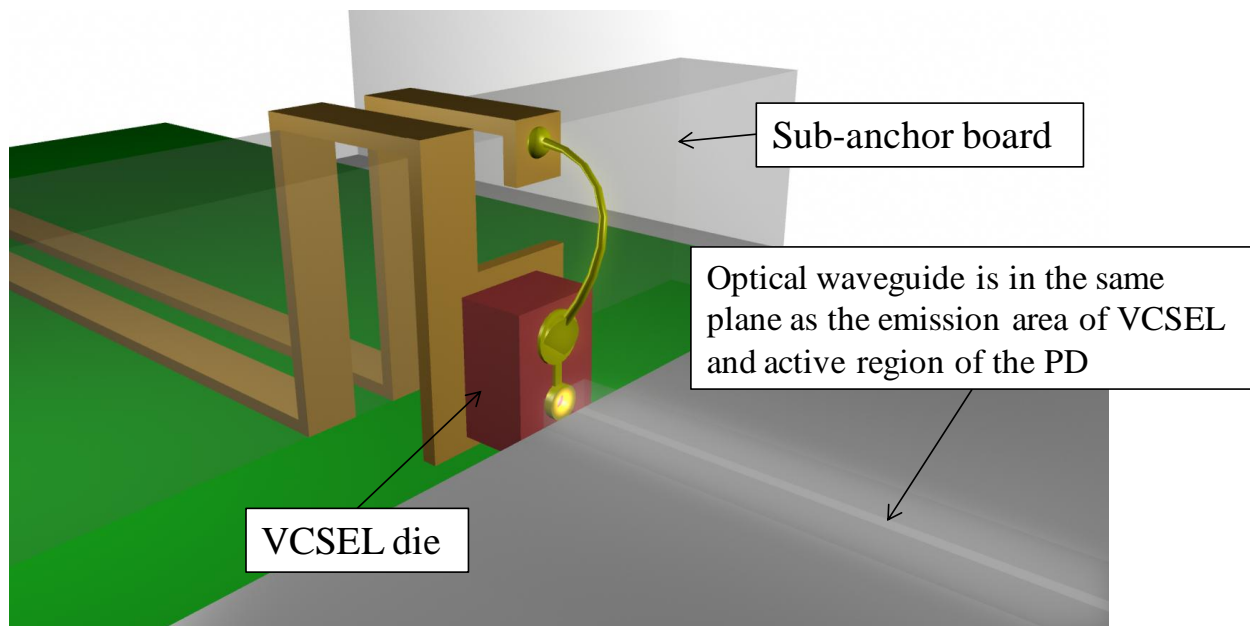


Figure 13. Proposed architecture for VCSEL-waveguide coupling.

This architecture has low power loss (i.e., greater power efficiency) and may be an attractive alternative for companies interested in low-power high-bandwidth optical interconnect technology. The process is described in more detail in the next section, but basically the approach is to assemble the VCSEL and PIN PD side-mounted so that the light emission and viewing directions are both in the plane of the substrate. To achieve this these active components are first mounted on a sub-anchor board, making the side assembly less complex, then the sub-anchor board is rotated so that the 90° turn is made electrically. Same procedure is used at the

photo-detector end. A liquid, photo-definable monomer is placed between the two O/E components and forms a monomer-component interface. Polymerization of the monomer film between VCSEL and PIN PD is initiated with a stripe of UV light defined by a glass mask. A proximity UV stepper is used for mask alignment between the two. In addition to placing liquid monomer between the two O/E components, we experiment with solvent containing monomers and solvent free monomers; each requiring a different integration sequence and each showing advantages and challenges. One advantage of a liquid photo-definable optical monomer is that it can easily make an interface with lasers and photo-detectors, and the interface becomes permanent after polymerization. This interface greatly improves the coupling efficiency, and therefore the power efficiency. An alternative to liquid monomers is the use of a dry film optical polymer laminate, which is beginning to become available (e.g. Rohm & Haas), is that the VCSEL and PD can be assembled with their active areas in near contact with the laminated thin film after the lamination process, then a light stripe connecting the laser and detector can again be used to initiate polymerization to define the waveguide core. There are advantages and disadvantages to using thin film polymers or liquid monomers; these are discussed in the next section.

6.4 Fabrication Process

The exact fabrication process varies according to the different materials used, especially with the specific monomer/polymer type. The process described below is the integration process used in our experiment. In high speed applications at 10 Gbps and higher, the process would be modified by the use of through vias and flip-chip O/E component attachment instead of using the expedient of wire bonding, but the principle approach would remain the same. The following are

the principal steps we used for side-mounting the VCSELs. The sub-anchor here consists of a cut section of glass microscope slide with gold metal-coating.

The major sequence steps in our process for a solvent-less monomer, in this case OrmocerTM are briefly as follow (Figure 14 has diagrams of the different steps):

Step 1. The sub-anchor is placed vertically to easily mount the VCSEL's back-side contact using a silver epoxy. The VCSEL emission area is placed at the height of the waveguides, which is known. The PIN PD is done in a similar manner.

Step 2. The VCSELs and PDs are wirebonded to their respective sub-anchor boards. Once this part is completed each sub-anchor board is tilted to face the waveguide substrate to achieve the 90° turn and fixed to the substrate on which waveguides are to be fabricated and aligned with the respective components.

Step 3. The liquid photo-definable optical monomer is placed on the substrate by spin coating. This creates an interface between the VCSEL emission area and the PIN PD active region. Application of the optical monomer by a die extrusion process is to be preferred instead of the spin coating expedient.

Step 4. The glass mask is placed such that the emission area of the VCSEL is aligned with the active region of the PIN PD. The monomer is exposed to UV light beginning the polymerization process. After the polymerization is completed the excess monomer is developed leaving behind the polymer waveguides. The Mercury I-line proximity exposure tool we use has a depth of field of about 3 mm and is reasonably uniform over a 6-inch x 6-inch area.

Depending on the type of liquid photo-definable optical monomer used the process could be varied. For example, when using a solvent-based monomer solution such as optical material

from Rohm & Haas steps 1 and 2 remain the same but the subsequent steps are altered accordingly.

Alternate step 3. The passive waveguide array is fabricated separately. The sub-anchor board is placed and the emission area of the VCSEL is aligned with the waveguide. Similarly, the PIN PD active region is aligned with the waveguide at opposite end. Step 4 is not required.

The case of dry-film optical core, when it becomes available, will be processed in a way that is similar to standard dry film photo-resist laminates. The tools required for this process are standard in the board processing industry. It can be seen that the process is very simple and relatively inexpensive. Because of its simplicity it is an attractive solution for fabrication of fiber-to-the-home components and other mass produced networking products.

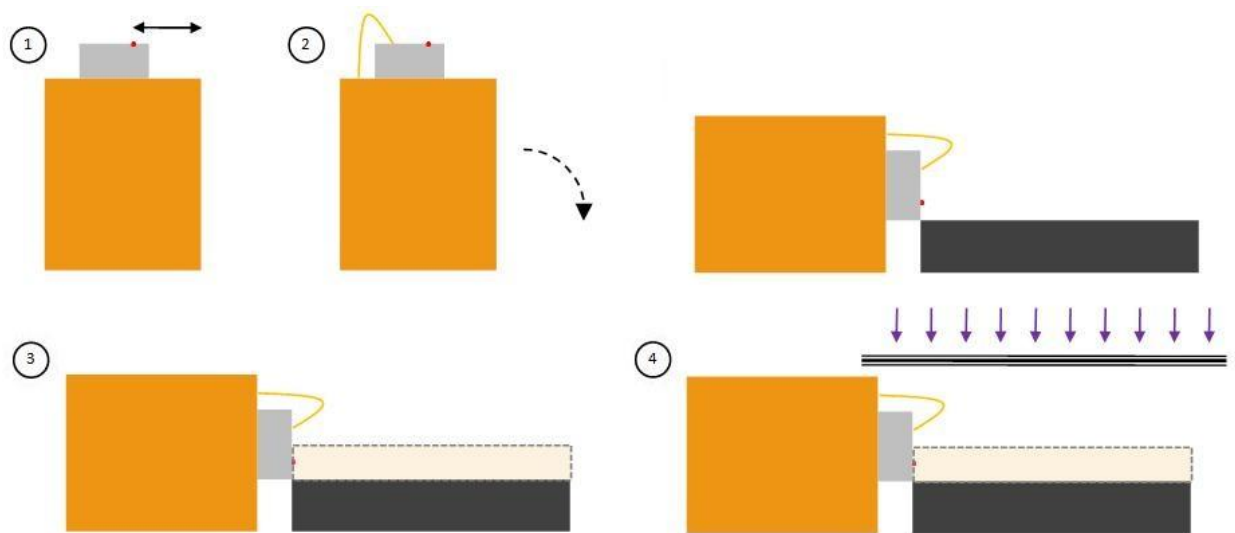


Figure 14. Graphical view of the process required to optically couple the VCSEL to the PIN PD (only VCSEL side is shown).

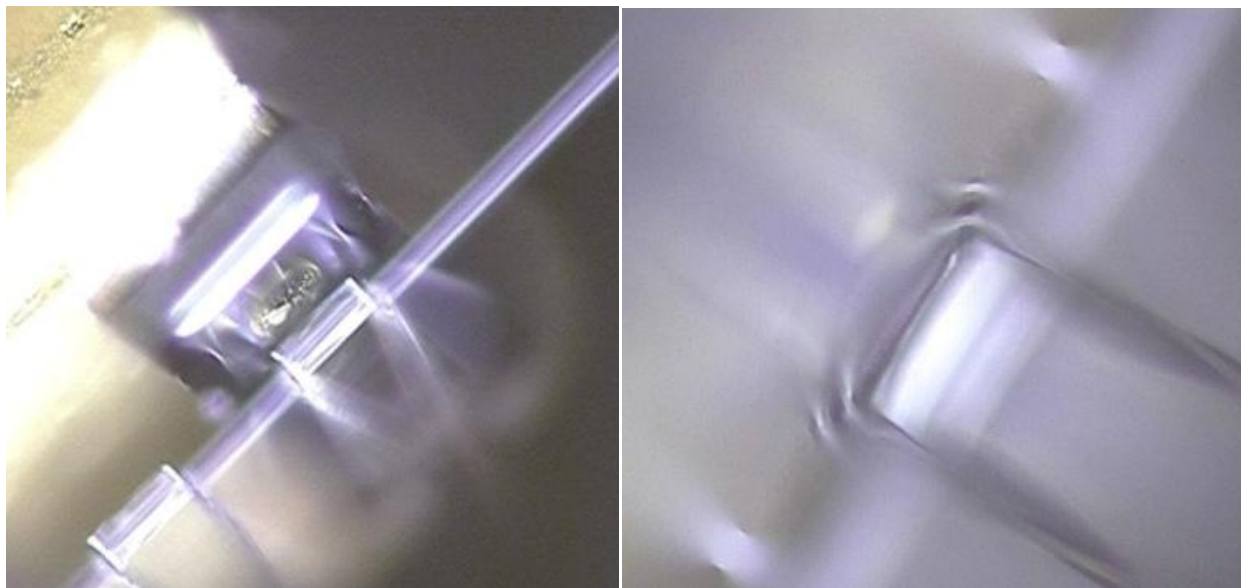


Figure 15. (left) VCSEL aligned with waveguide (Rohm & Haas) after process completion.
(right) PIN PD active region aligned with waveguide after process completion.

Figure 15 shows microscope photographs of both E/O components aligned with the waveguides after process completion. This was a preliminary test without wirebonds. If the VCSEL was powered the light emitted would be directly injected into the waveguide with no aid from mirrors or lenses that would cause additional power loss. Similarly, at the other end, the active region of the PIN PD would collect the output light without a mirror or any other optical components.

The experiments performed using this process are described in Chapter 7. The results are shown in Chapter 9.

Objective and Design of Experiments and Simulations

7.1 Overview

To test the performance of the solutions proposed in this work a series of simulations and experiments are necessary. The first set of simulations shows the improvement of performance of the Ethernet services transport protocol over Traditional TCP in various scenarios. The later portion of this chapter describes the experiments performed to test the capabilities of the side-mounted VCSEL architecture.

7.2 Ethernet Services Transport Protocol Simulations

A series of simulations were performed to test the performance of ESTP. Each simulation has a specific purpose. The purpose and description of each simulation follows.

7.2.1 Throughput Performance

A throughput performance is necessary to understand the ratio of improvement that ESTP is capable of achieving under common network conditions. The simulation setup consists of two end-user pairs competing for bandwidth over a single line (see Figure 16). One pair is transmitting the studied traffic, while the other is generating background traffic. The studied traffic consists of a 1-GB file transmission. Background traffic consists of exponentially distributed file sizes with a mean of 10 MB, with an inter-request time exponentially distributed with a mean of 1 second. The shared line has a 1-Gbps link capacity and a 0.001 packet loss rate. The protocols that are compared in this work are HighSpeed TCP, TCP-Sack, and ESTP.

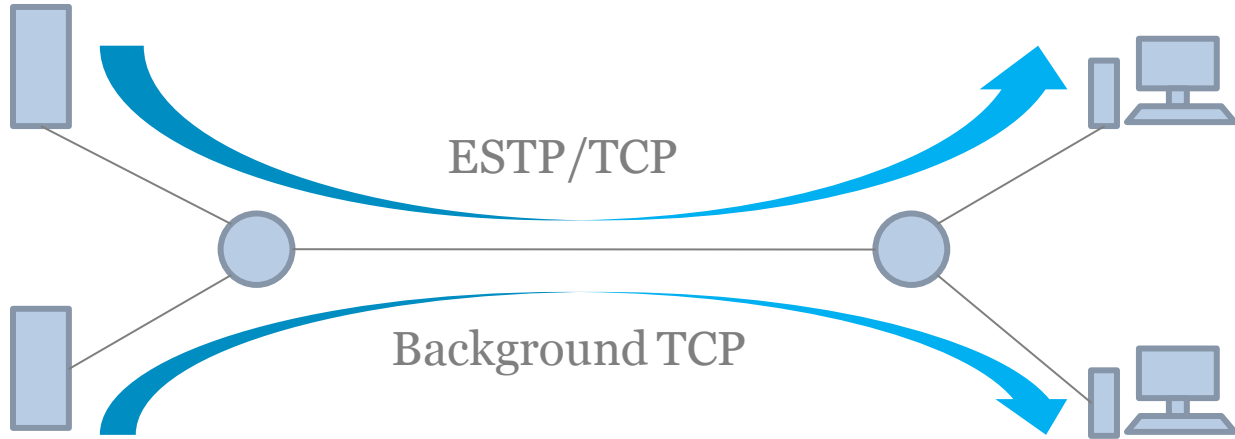


Figure 16. Throughput simulation scenario.

The background traffic is seeded with the same value for all simulations so each burst of background traffic will begin at the exact same time for all protocol cases. Once the burst of background traffic begins transmission, it will compete for bandwidth and will diverge slightly for each case.

7.2.2 Translation Protocol Performance

As mentioned in Chapter 5, ESTP requires a translation protocol to better utilize the available bandwidth. A simulation scenario was created to show how this interaction, between ESTP and the translation protocol, affects the performance of the network, specifically the bandwidth usage efficiency. This simulation scenario is similar to that of Figure 16, except there is no background traffic, just two end systems connected by UNIs. In this case, instead of having one single large file the server will send many small files. The purpose is to emulate a situation where we have ~100,000 sessions per millisecond running at 100 Gbps. To approximate this situation without generating a lengthy simulation a less strenuous scenario having ~100 sessions per millisecond running at 100 Mbps was chosen.

The way the simulation is performed the workstation informs the UNI how many active connections it has at one time. The UNI recalculates the CIR_{ESTP} and EIR_{ESTP} parameters and

sends the updated parameters. Because this translation process is not instantaneous, but rather it takes a small delay to refresh the parameters, the number of active connections could have changed by the time the latest updated parameters arrive to the workstation. For this reason the bandwidth efficiency is a function of this update period. Since new connections will use the latest CIR_{ESTP} and EIR_{ESTP} parameters available there is a possibility that the channel is overutilized if too many connections are initiated at one time and the CIR_{ESTP} and EIR_{ESTP} parameters have not been reduced accordingly. A buffer zone can be added to alleviate the effects of these abrupt changes in the number of active connections. An alternative way to do this is to have the workstation request permission to the UNI to initiate a connection, this will avoid the use of a buffer but it will cause an initial delay while the workstation waits for approval, but this method was not simulated rather just mentioned for the reader's benefit. The above case was simulated for different size buffers, the greater the buffer the less chance of overutilizing the available bandwidth, but there is less total available bandwidth.

7.2.3 UNI Scheduling Scheme Performance

Common scheduling schemes are not fully compatible with carrier Ethernet network specifications, as specified by the MEF. So a new scheduling scheme, using conventional well-known techniques, was designed to work in compliance with MEF rules. This scheduling scheme is described in Chapter 5 and the results are shown in Chapter 8.

7.3 Side-mounted VCSEL Experiments

The transport protocol by itself cannot provide high-bandwidth transmissions if the intermediate devices are not able to handle high data rates. To achieve data rates of 100 Gbps, devices (such as routers and UNIs) must have an optical backplane. For this reason we have gathered results on an innovative optical interconnect technique of VCSEL-to-waveguide

coupling. To gather the results a sample was fabricated, as shown in Figure 17, the VCSEL is mounted on the sub-anchor board that has already being rotated 90° but no monomer has been placed yet.



Figure 17. Top view of VCSEL placed on sub-anchor board without waveguides.

The next step is to place the monomer on top of the substrate covering the VCSEL and photodetector creating an interface between the liquid and the component, this interface is the key to low-power loss since the light is injected directly into the waveguide going through just one surface change (materials with different refractive indices), hence reducing the amount of reflections. The sample is covered with a glass mask, which is aligned with the VCSEL and PD and exposed to UV light to cure the monomer and create the polymer waveguide. Because the VCSEL is mounted sideways the beam is in the same plane of the waveguide and no mirrors are necessary. A photograph taken with an infrared camera shows the light traveling through the waveguide (Figure 18).

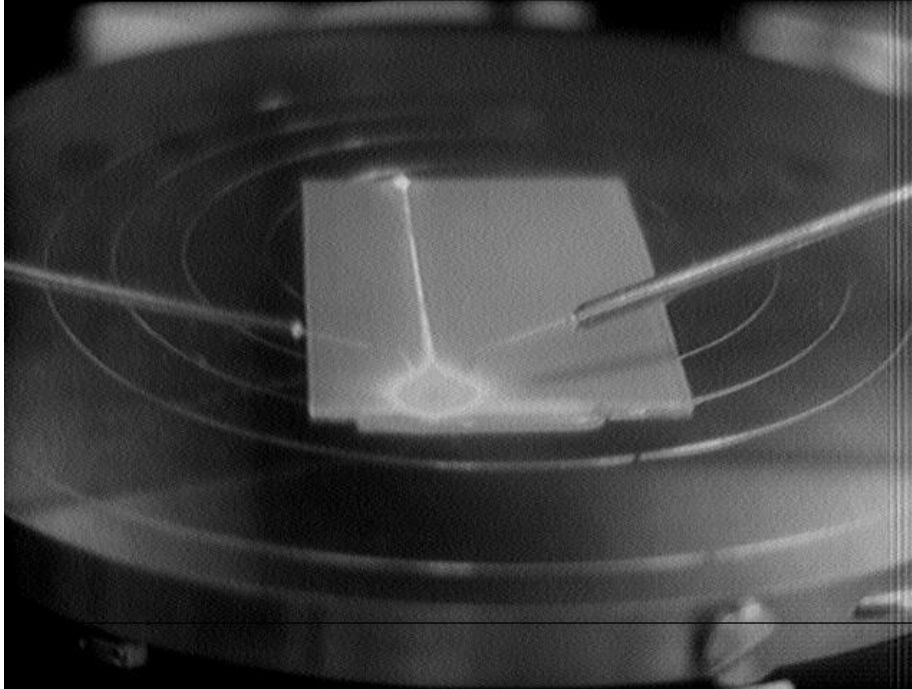


Figure 18. Infrared photograph: light emitted by VCSEL traveling through the waveguide.

Several optical devices were integrated using this process and experiments performed to test the characteristics of the proposed architecture. These experiments are aimed to test the feasibility of using the proposed architecture in mass producible high-data-rate equipment.

The experiments include:

1. Horizontal alignment sensitivity experiment,
2. Vertical alignment sensitivity experiment,
3. Coupling efficiency experiment, and
4. Crosstalk experiment.

The individual reasons for having these tests are explain in the following sections.

7.3.1 Horizontal and Vertical Alignment Sensitivity

Because this architecture requires alignment of different components is it important to study the effects that misalignment could have on the design. An experiment was conducted to

gather this data. For both horizontal and vertical cases the experiment setup is the same, only that the direction in which the data is taken is different.

Because the laser and the waveguide need to change position with respect to each other a sample with the interfaces connected to each other is not possible, so a thin gap of air is necessary to disjoint these sections. In our experiment the laser position was fixed and only the waveguides were moved in a 3-axis stage. Since we are interested more in the power loss (i.e. power difference between input and output), the z-axis was just placed such that good percentage of the light was transmitted, but this variable is not of interest. The center was found by simply finding the point of maximum power transmission. Once centered, different measurements were taken with a resolution of 0.5 microns. The results are shown in Chapter 9.

7.3.2 Coupling Efficiency

This is an indispensable experiment as it determines how the new architecture performs. Most publications in this field compute the coupling efficiency of the VCSEL side only, obtaining results as high as 70-80% coupling efficiency [63][64]. This range is considered high coupling efficiency. When performed end-to-end it is very difficult to obtain double digit coupling efficiencies, especially with low-cost methods.

The process described at the beginning of this section (7.3) is used to make the sample for this experiment. The sample is made end-to-end, meaning it has a VCSEL, waveguide and PIN PD. The waveguide is 10.8 cm long, since we used a polymer called LightLink™ from Rohm and Haas with a 0.015 dB/cm power absorption [61], the theoretical power absorption loss of this sample is 4%, when the waveguide imperfections are added the total power loss increases, but all these factors are included in our optical coupling efficiency measurement. Since we know the output power of the VCSEL as a function of injected current and the responsivity of the PIN

PD, we can compute the output power received by the PIN PD and hence compute the coupling efficiency as a function of injected current. Since all the elements necessary to gather the coupling efficiency data are available, all that is left is to setup the experiment. This setup is relative simple. The VCSEL is connected to a power supply and the PIN PD to a power meter. The injected currents and voltages are recorded separately, similarly with the measured current and voltages. Later the respective powers can be computed and hence the coupling efficiencies.

7.3.3 Crosstalk

This experiment is necessary to see what level of interference exists between adjacent channels. This experiment was performed with waveguides that had 250 microns separation, standard separation for VCSEL arrays. The setup is similar to that of section 7.3.2, but instead of measuring the same output channel we measure the output of the adjacent channel. Crosstalk is defined in this work as the ratio of output power of the adjacent channel and the input power of the reference channel. The results are shown in Chapter 9.

Simulation Results: Ethernet Services Transport Protocol

8.1 ESTP Performance Throughput Performance Simulation

The simulation setup consists of two end-user pairs competing for bandwidth over a single line. One pair is transmitting the studied traffic while the other is generating background traffic. The studied traffic consists of a 1-GB file transmission. Background traffic (Figure 19) consists of exponentially distributed file sizes with a mean of 10MB, with an inter-request time exponentially distributed with a mean of 1 second. The shared line has a 1-Gbps link capacity and a 0.001 packet loss rate. The protocols that are compared in this work are HighSpeed TCP, TCP-Sack, and the proposed protocol – Ethernet-Services Transport Protocol (ESTP). The parameters used are described in Table 1.

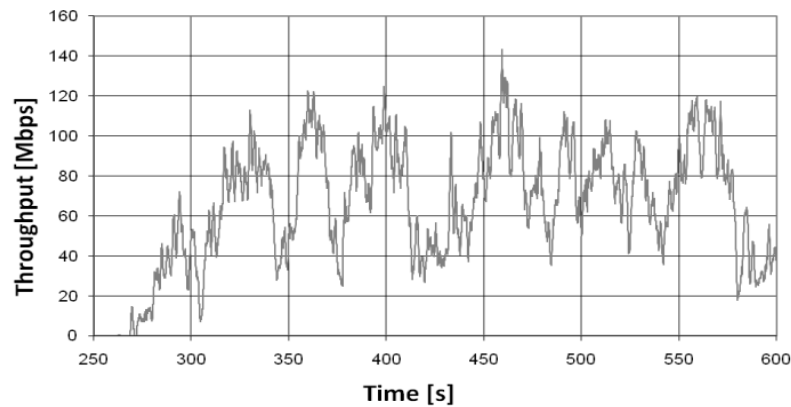


Figure 19. Background Traffic.

The background traffic is seeded with the same value for all simulations, so each burst of background traffic will begin at the exact same time for all protocol cases. Once the burst of background traffic begins transmission, it will compete for bandwidth and it will diverge slightly for each case.

Table 1. Simulation Parameters for Scenario 8.1

CIR	200 Mbps
CBS	250 kB
EIR	400 Mbps
EBS	500 kB
Link traffic loss	0.1%
Round-trip propagation delay	[1, 10] ms
Traffic shaping buffer size	64 kB
Receiver buffer size	128 kB
Maximum segment size	1460 B
Max ACKed packets	2

Figure 20 shows the goodput of the compared protocols for an RTT range of 1 to 10 ms. Note that the goodput and throughput differ slightly. Throughput is a measure of all data flow over a specified link, as opposed to goodput that does not take into account the retransmitted packets, the ACK packets, the TCP packet header or the IP packet header. It only considers the data load of the packet. The goodput is thus calculated by dividing the transmitted file size by the total transmission time.

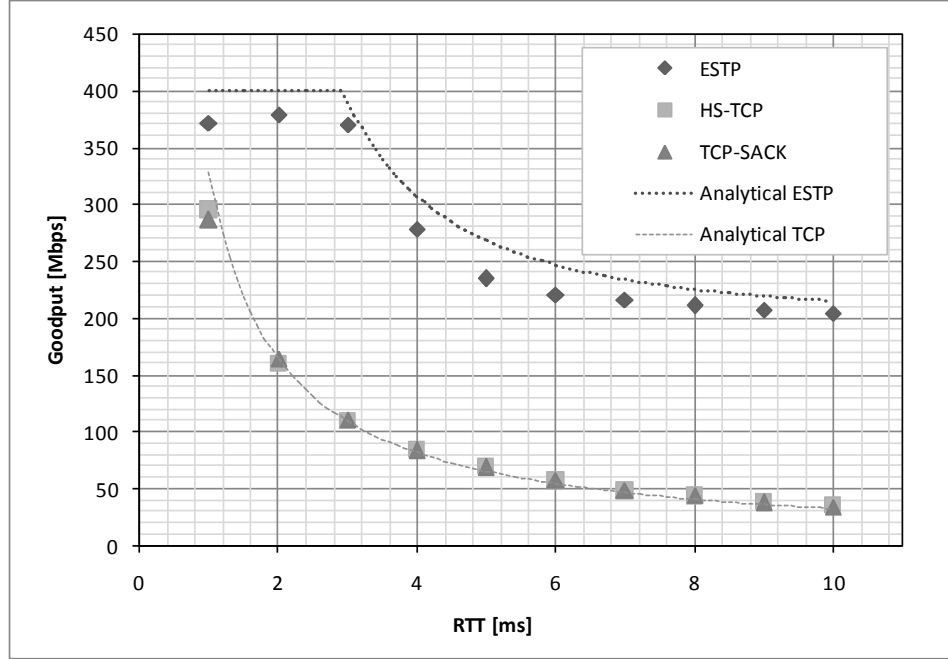


Figure 20. Goodput comparison between ESTP, HighSpeed TCP and TCP-Sack.

The goodput of the proposed protocol is approximately 6 times greater than that of HS-TCP or TCP-Sack for an RTT of 10 ms. The throughput improvement of the proposed protocol is a result of knowing the layer 3 restrictions established in the SLA and using this information to make smarter use of the congestion window. If the CIR and EIR are increased, the improvement ratio relative to these protocols will also increase. It should be noted that HS-TCP performs similarly to TCP-Sack because it is not designed for environments with loss rates higher than 10^{-3} . The behavior observed for the 1 ms and 2 ms RTTs is attributed to the EIR restriction. In this simulation the EIR was set to 400 Mbps. Therefore, the average throughput cannot exceed this parameter for a particular time window, limited by the EBS. For the cases where RTT is greater than 5 ms the goodput improvement ratio behaves linearly with an improvement increase of approximately 50% for every additional millisecond RTT increase (see Figure 21). Both transmission goodputs are decreasing with increasing RTT, but the TCP-Sack goodput is

decreasing faster than ESTP's goodput. This constant increase in improvement-ratio occurs because the proposed protocol is lower bounded by the CIR. The greater the CIR chosen, the greater the improvement ratio resulting from the proposed protocol over TCP-Sack.

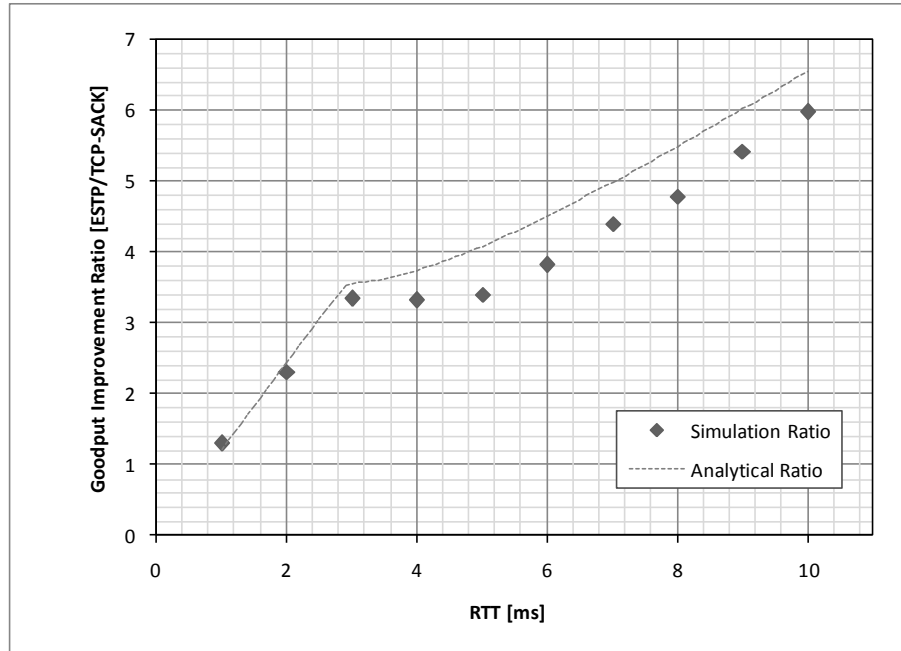


Figure 21. Goodput improvement ratio between ESTP and TCP-Sack.

8.2 Translation Protocol Simulation

OPNET Modeler® 14.5 was also used to simulate the proposed translation and scheduling techniques and obtain the performance results. The first simulation scenario consists of a server transmitting small files through the CEN passing through the UNI which is enforcing a bandwidth profile and at the other end is received by a workstation. The small files have different sizes generated randomly with exponential distribution with an average of 1kB. The inter-request time is also randomly generated with exponential distribution with an average of 10 μ s. The simulation parameters are given in Table 2. The goal of the simulation is to observe

how the translation protocol performs as a function of the update period. The update period is the time interval in which the UNI refreshes the CIR and EIR parameters given the number of open connections. Having an EIR of 100Mbps and ~100 sessions per millisecond, is a rough approximation of having an EIR of 100Gbps and ~ 100,000 sessions per millisecond. The burst sizes were also scaled down and a buffer parameter was also introduced. The purpose of this parameter is to have a gap between the available bandwidth and the total EIR. Once the sessions have been assigned new values of CIR and EIR, ESTP will reach the EIR if the packet loss is small. If a new ESTP session is initiated, it will have the same parameters as the other session, but the UNI may not have yet updated the new parameters to reflect this change, hence causing the data rate to overflow EIR and packet losses to occur. The bandwidth utilization efficiency is computed by doing a time average of the throughput divided by EIR.

Table 2. Simulation Parameters for Scenario 8.2.

CIR	50 Mbps
CBS	5 kB
EIR	100 Mbps
EBS	10 kB
Link packet loss	0.00001%
Round-trip propagation delay	10 ms
Traffic shaping buffer size	64 kB
Receiver buffer size	128 kB
Maximum segment size	1460 B
Max ACKed packets	2

The results of this simulation are shown in Figure 22. For short update periods the bandwidth was 90-95% utilized, and for update periods closer to 1 ms the efficiency dropped about 10-15%. It is desirable to update the parameters with the lowest frequency possible to reduce computational overhead while having the greatest bandwidth utilization efficiency. If processing

is not an issue, such that update periods of 1 μ s or lower can be easily supported, a small buffer can be introduced to avoid overflowing the UNI. The greater the buffer gap, the lower the probability of overflow, at the cost of reducing the maximum allowable throughput. The buffer gap is computed in the following way:

$$buffer = 1 - \frac{BW_{ESTP}}{EIR} \quad (8.1)$$

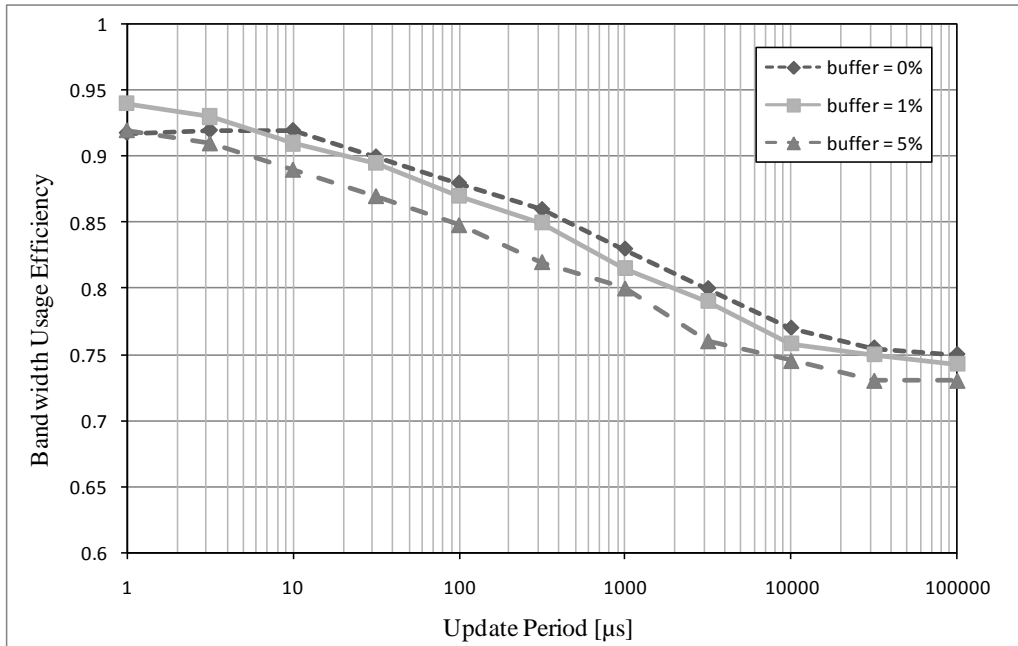


Figure 22. UNI parameter update period vs. the bandwidth utilization efficiency.

8.3 UNI Scheduling Scheme Simulation

The third simulation scenario consists of three servers and three workstations, each pair having a different priority level. No workstation can communicate with any other server but its own (this is the purpose of having EVCs), but all share the same UNI. The simulation parameters are shown in Table 3.

Table 3. Simulation Parameters for Scenario 8.3.

CIR ₁	200 Mbps
CBS ₁	20 kB
EIR ₁	400 Mbps
EBS ₁	40 kB
CIR ₂	100 Mbps
CBS ₂	10 kB
EIR ₂	200 Mbps
EBS ₂	20 kB
CIR ₃	50 Mbps
CBS ₃	5 kB
EIR ₃	100 Mbps
EBS ₃	10 kB
Link packet loss	0.00001%
Round-trip propagation delay	[1-10] ms
Traffic shaping buffer size	64 kB
Receiver buffer size	128 kB
Maximum segment size	1460 B
Max ACKed packets	2

In this scenario, we vary the round trip time (RTT) of the link and observe the throughput changes. The service rate of the UNI was set to 700Mbps. The corresponding weights as given by

$$\Phi_k = EIR_k / \sum_i EIR_i$$

were calculated, obtaining $\Phi_1 = 0.571$, $\Phi_2 = 0.286$ and $\Phi_3 = 0.143$.

Figure 23 shows the results of the second simulation scenario. In that figure we observe that the throughput of the priority 3 data falls within the range of 50 Mbps and 100 Mbps which are the CIR and EIR rates for this level. For the RTTs between 1 and 3 ms the data flow is constrained to the EIR. For RTTs greater than 4 ms, the distance is far enough that ESTP cannot reach the EIR. For all RTTs the throughput is above the CIR satisfying the requirements of the SLA.

The curves for priority 2 and 3 data also satisfy the requirements imposed by the SLA. Since all traffic fell within the respective set of SLA requirements for their priority level this shows

that by using the weighting factors of Φ_1 , Φ_2 , and Φ_3 it is possible to successfully provide fairness within the design rates of CIR and EIR. Because all traffic flows exceed their respective CIR, this implies that the CIR leaky-bucket and the FIFO scheduler are working as expected. Since the throughput of priority 1 traffic is twice as much as that of priority 2 and four times as much as that of priority 3, it means that the weighted fair scheduling scheme is also working correctly, and finally since the guaranteed traffic is met and it can be observed that there is also best-effort traffic (any traffic that exceeds CIR) it implies that the strict priority scheduling scheme is also working properly. These three characteristics prove that the scheduling scheme works as designed and therefore can provide bandwidth QoS control guarantees as specified by the MEF.

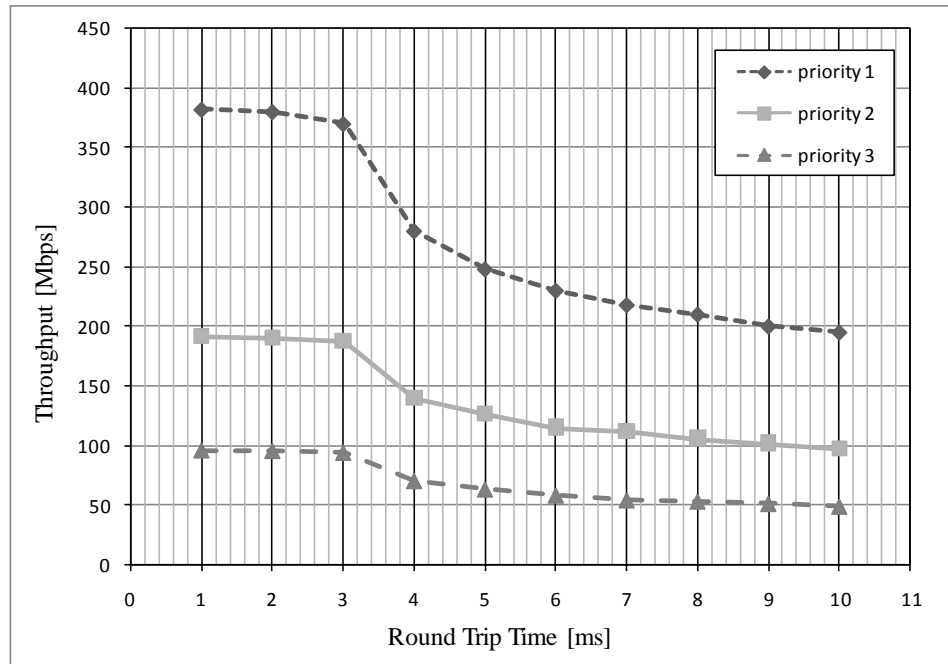


Figure 23. Throughput behavior of different priority level as RTT is varied.

Experimental Results: Side-mounted VCSEL

9.1 Horizontal Alignment Sensitivity Experiment

In this experiment polymer waveguides were made on a silicon wafer substrate. The waveguide consisted of a core surrounded by the undercladding and uppercladding. The silicon wafer containing the waveguide was cleaved at both ends to obtain a clean flat surface. After cleaving it was placed on a static fixture. The VCSEL was mounted on a gold plated sub-anchor board. To vary the horizontal position it was placed on a three-axis crossed-roller bearing translation stage with a micrometer minimal reading of 1 μm . On the opposite side of the waveguide a 5 mm Germanium photodetector (with a 0.25 V reverse bias) was placed on a static fixture to collect all light transmitted through the waveguide. The translation stage was centered and then moved approximately 50 μm off-center, then data was gathered in 1 μm intervals. The results of this test are shown in Figure 24. The top figure shows the approximate scanned path.

From Figure 24 it can be seen that the half-energy width is approximately 50 μm wide. This width indicates that the alignment sensitivity is rather lenient. The total width of the core is approximately 60-70 μm . It should be emphasized that no other optical techniques or devices were used besides those described in the experimental setup.

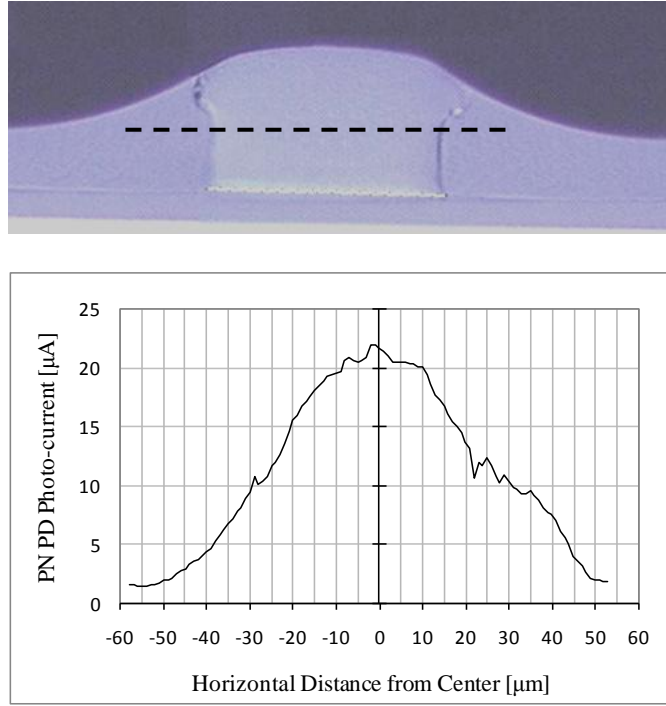


Figure 24. (top) Cross-section of the waveguide (Rohm & Haas) used to gather alignment sensitivity with the horizontal path outlined. (bottom) Horizontal alignment sensitivity results.

9.2 Vertical Alignment Sensitivity Experiment

The experimental setup of this test is the same as the horizontal alignment sensitivity experiment setup. The polymer waveguides are placed on a static fixture. The sub-anchor board with the VCSEL is placed on a three-axis crossed-roller bearing translation stage and the Germanium photodetector (also with a 0.25 V reverse bias) was placed on a static fixture. The translation stage was centered and then moved approximately 40 μm of center, and then data was gathered in 1 μm intervals. The results of this test are shown in Figure 25.

Even though the height of the core is approximately 50 μm , the half-power height is of 45 μm . The reason for having an abrupt drop (at the bottom), opposed to the results from the horizontal case, is that the undercladding is very thin and lies over the silicon wafer which has a

highly reflective surface. The top portion also exhibits an abrupt change, but occurs for a different reason. In this case the core is bounded by air rather than the uppercladding, forming a higher index of refraction difference and therefore creating a greater total internal reflection angle, hence at the air-cladding interface light is totally reflected and the cladding polymer acts as a core material. The residual power that appears on the left part of the graph is due to light propagating through the lower cladding layer. This test also shows that the configuration is very forgiving to misalignment.

Combining the results of both alignment sensitivity tests it can be said that if the VCSEL is misaligned by $20\text{ }\mu\text{m}$ in any direction, that is to say within a radius of $20\text{ }\mu\text{m}$, the power penalty that the system will suffer will not exceed 3dB.

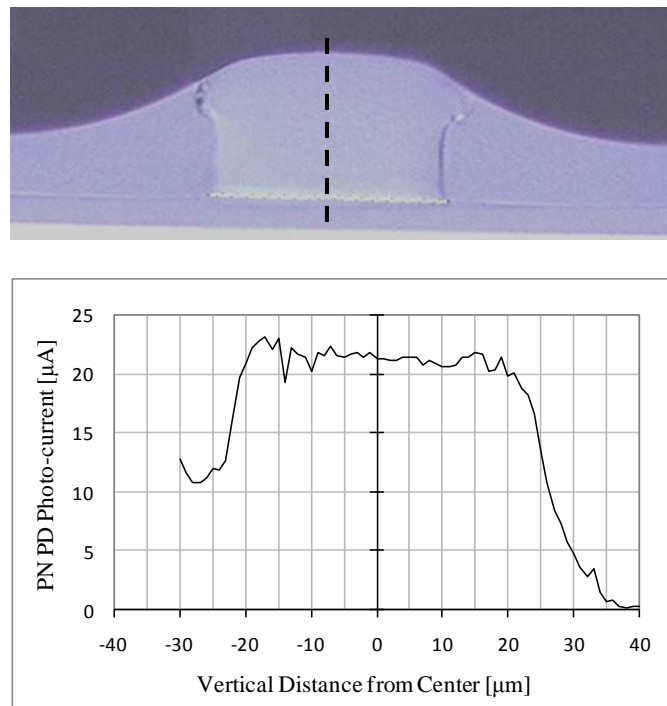


Figure 25. (top) Cross-section of the waveguide (Rohm & Haas) used to gather alignment sensitivity with the vertical path outlined. (bottom) Vertical alignment sensitivity results.

9.3 Coupling Efficiency Experiment

For the experimental setup a sample was build using the process described in the process section of this paper. The sample had two VCSELs mounted on the sub-anchor board. The VCSELs were powered and the current generated by the PIN PD was recorded. The waveguides are 10.8 cm long. The injected current into the VCSEL was varied from 0.5 – 3 mA in intervals of 0.25 mA. Figure 26 shows the experimental setup. On the left side is the VCSEL and on the right side is the PIN PD.

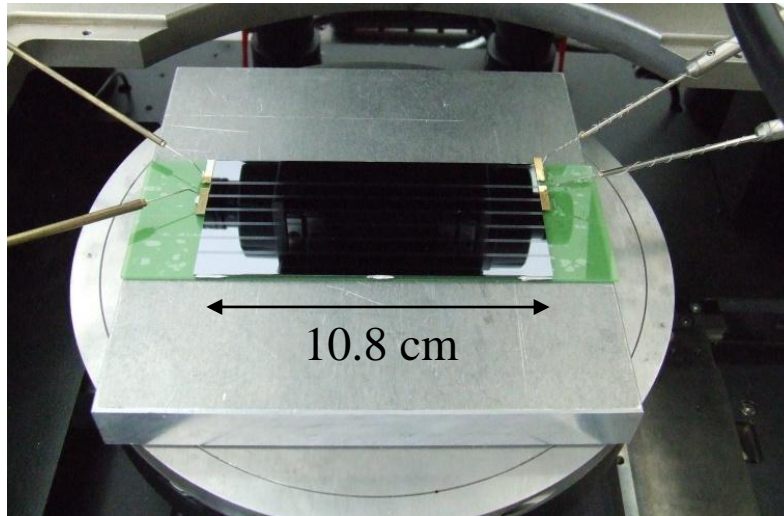


Figure 26. Experimental setup for coupling efficiency experiment.

The first step was to measure the output power of the VCSEL bare die. A current versus output power was obtained and the results are shown in Figure 27. While each VCSEL die behaves differently and there is a slight variation in the slope efficiency for each die, a randomly chosen VCSEL from the same batch as the test dies (obtained from Lumei Optoelectronics) was chosen as the calibration die for this experiment and is used to calculate the coupling efficiency. The calibration VCSEL was placed 5 mm from a 1 cm diameter Silicon PN photodiode on which all emitted light is incident.

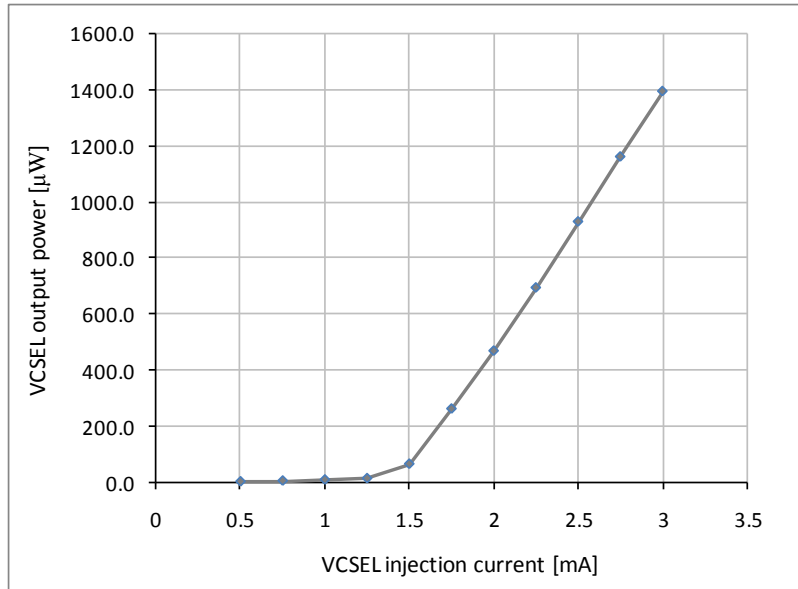


Figure 27. Relation between the VCSEL's injected current and output power.

The VCSELs from both channels were powered. The photocurrent generated by the PIN PD for each channel was plotted against the VCSEL injected current. The results of both channels are shown in Figure 28. It can be seen that both channels behave relatively the same. This indicates that the process is fairly stable and capable of producing consistent results.

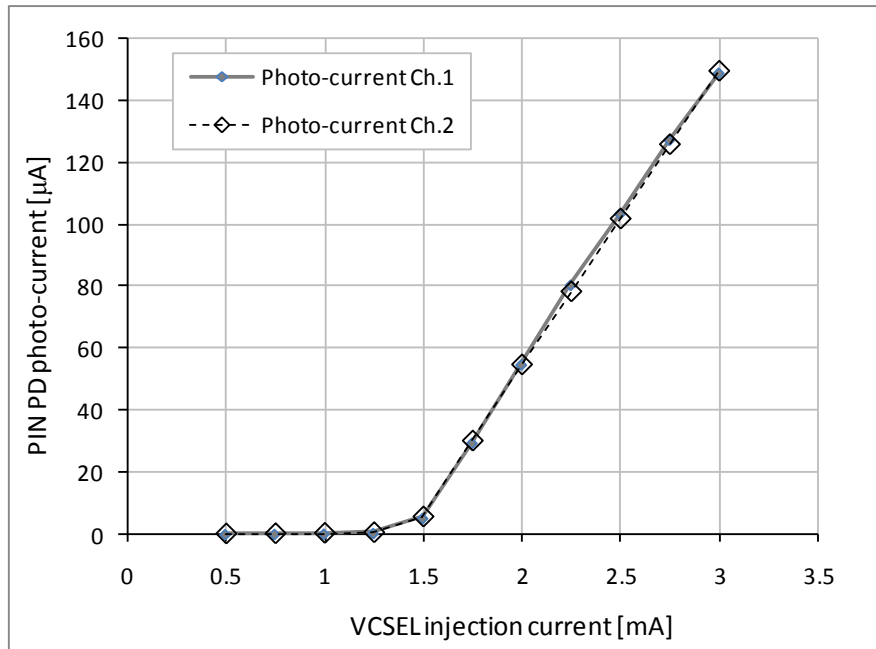


Figure 28. Relation of the photo-current for each channel with varying VCSEL injection current.

To calculate the coupling efficiency we use the PIN PD responsivity from Table 4 to calculate the input optical power to the detector and find the ratio to the measured output VCSEL power. For injection currents above the VCSELs threshold current we obtain coupling efficiencies between 18% and 20% for 10.8 cm long polymer waveguides fabricated from Rohm & Haas optical polymer. The coupling efficiencies are shown in Figure 29. Considering the simplicity and inexpensive process to build these optical channels these relatively high coupling efficiencies are very encouraging.

Table 4. VCSEL Specifications	
Parameter	Value
VCSEL Threshold Current	1.4 mA
VCSEL Slope Efficiency	0.9 W/A
PIN PD Responsivity	0.58 A/W
PIN PD Dark Current	30 nA

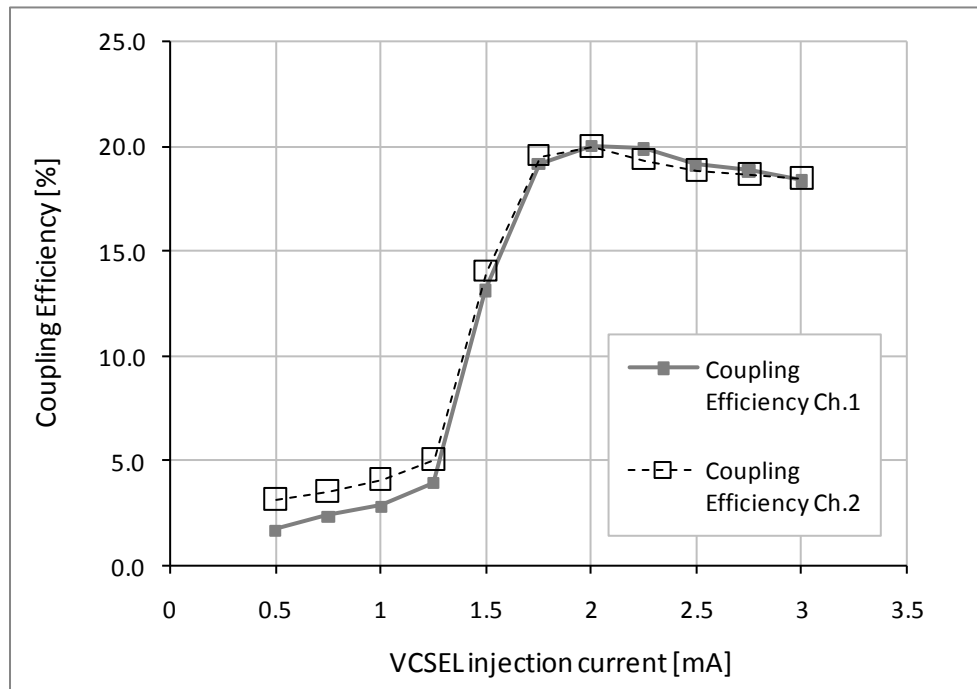


Figure 29. Coupling Efficiency results for both channels.

To obtain a better understanding of the reason for achieving a 20% coupling efficiency some photographs were taken with an infrared camera with a Silicon Vidicon active element. The photographs reveal that there is very little scattering of light by the polymer waveguide, an example is shown in Figure 30. To obtain these results the VCSEL injection current was brought up to 4 mA. At currents lower than 2.5 mA the waveguides that had light injected into them were nearly undistinguishable from those that carried no light.

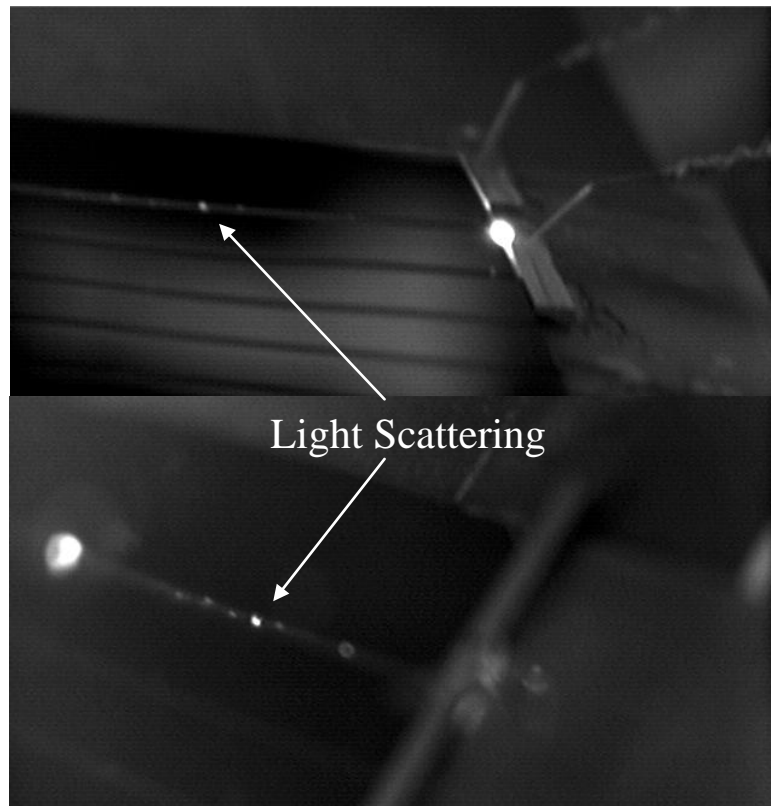


Figure 30. Infrared camera photographs. (top) Side view. (bottom) View from behind the PD.

9.4 Crosstalk Experiment

An important parameter that describes how much interference exists between adjacent channels in the system is crosstalk. To determine the crosstalk a VCSEL from one channel is powered and the PIN PD of an adjacent channel is reverse biased. The VCSEL injection current is varied in the same way as the coupling efficiency experiment, i.e. from 0.5 – 3 mA.

The results of the crosstalk are plotted in Figure 31. For values greater than the VCSEL threshold current the crosstalk is approximately -27 dB, which means that only 1/500th part of the signal gets fed into the adjacent channel. The higher crosstalk below threshold is due to fluorescence emission which is thought to have a greater divergence and can couple to the adjacent waveguide. The parallel waveguide array in this experiment consists of 50- μm wide waveguides on a 250 μm pitch.

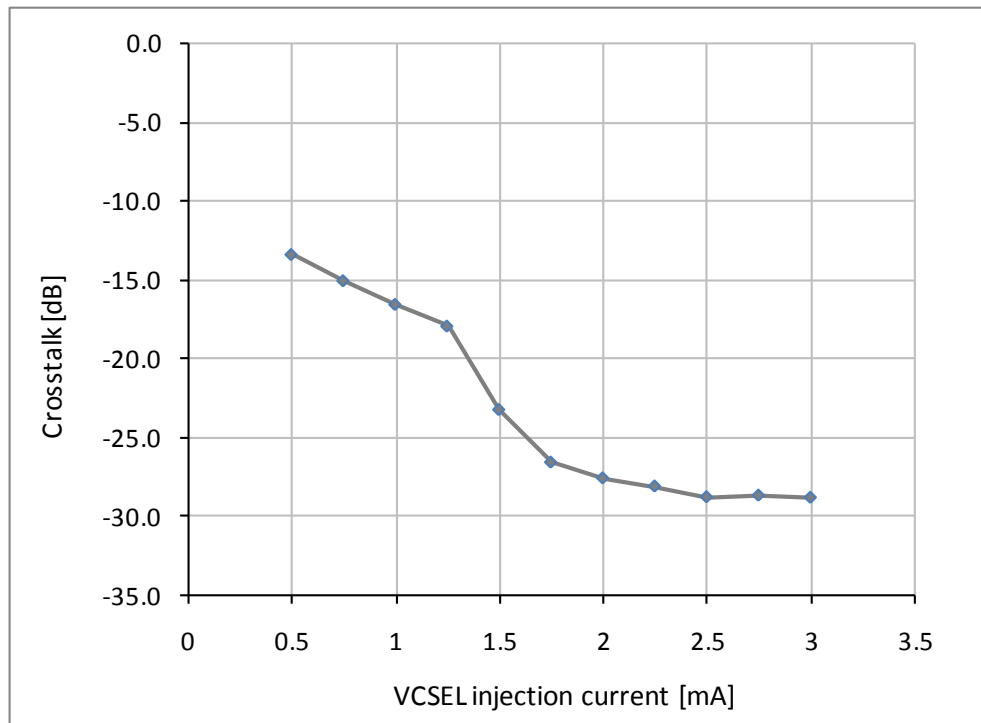


Figure 31. Crosstalk experimental results.

Conclusions, Contributions and Future Work

10.1 Conclusions

10.1.1 Ethernet Services Transport Protocol

In this work a novel congestion control scheme is proposed for Ethernet Services running on carrier-grade metro Ethernet networks. The proposed transport scheme combines Ethernet Services information and effective utilization of the available bandwidth. As a result, the throughput is significantly increased. It achieves this by mapping the packet loss interval to a less strenuous multiplicative factor and efficiently using the SLA information, such as EIR and CIR. Incorporating this information into the transport protocol helps control the traffic flow more efficiently. The congestion window is controlled adaptively with the current congestion status, which is estimated by the amount of packets successfully delivered in-between two packet losses. Congestion control aggressiveness is proportional to the congestion level. By incorporating the Ethernet Services information and effective use of available bandwidth, the proposed protocol can achieve a 600% increase in goodput for a round-trip time of 10 ms over TCP-Sack with a traffic shaping excess information rate (EIR) of 400 Mbps, and a committed information rate (CIR) of 200 Mbps. Simulation results are reinforced by the analytical results, which show a 650% increase for the same parameters.

Additionally, two QoS control mechanisms were proposed: a translation protocol that provides ESTP with bandwidth profile information, and a simple scheduling scheme. By

providing dynamic adjustment of the CIR and EIR to ESTP, bandwidth utilization efficiencies of over 90% were achieved for update periods of 1-10 μ s. For longer update periods, in the 1-10 ms range, the bandwidth utilization efficiency was between 75% and 80%. These efficiencies could not have been achieved if the aforementioned parameters were fixed. Moreover, if small update periods are easily supported by the system, a buffer gap may be added to further increase the bandwidth utilization efficiency. It was also shown that a simple scheduling scheme can be used to effectively provide all the bandwidth profile requirements as specified by the MEF, while maintaining fairness. Alternate scheduling schemes could be used, but significant improvement over this simple scheme should be demonstrated to justify the added complexity costs. All of these mechanisms are enforced at the network edge by the UNI, and are oriented to improve the performance of Carrier Ethernet Networks.

10.1.2 Side-Mounted VCSEL

Efficient coupling scheme of VCSELs and PIN PD through waveguides of optical interconnects systems is an on-going and challenging research topic. Previous technical approaches favor flat-mounted VCSELs and the use of mirrors to guide the light into the waveguide and into the PIN PD. The use of mirrors causes a relatively large power penalty and hampers the scalability of integration. By using the proposed integration process, we greatly reduced the losses caused by the use of mirrors or other optical coupling techniques with a similar purpose. The process is very simple and the required components are inexpensive. Coupling efficiencies of approximate 20% are achieved in this work over a waveguide of 10.8 cm length. The process exhibited a lenient alignment constraint. A misalignment of 20 μ m in any direction from the center of the waveguide was found to cause a loss of less than 3 dB. The similarity in multi-channel system performance shows that the process is stable and reproducible.

The manufacturing process can be inexpensive and with the performance attributes shown before, this architecture is a promising solution for mass produced networking equipment such as information access networking products for home and office environments.

10.2 Contributions

10.2.1 Ethernet Services Transport Protocol

In Carrier Ethernet network, data has to travel vast distances. This distance will degrade the performance of Traditional TCP. By incorporating ESTP customers of carrier-grade service providers will be able to enjoy the committed information rate they are purchasing regardless of the location of the end systems. The novel congestion avoidance algorithm, will allow the transmission throughput to maintain a higher speed level than Traditional TCP without compromising the congestion level of network.

10.2.2 Side-mounted VCSEL

This simple, yet innovative alternative to high frequency transmissions will allow companies to design low-cost high-speed networking equipment that can outperform equipment that is based on a more complex architectural VCSEL-waveguide coupling design. This will bring down the cost of network-edge routing equipment eliminating this bottleneck and therefore increasing the amount of traffic that can be potentially served to the customer.

10.3 Future Work

10.3.1 Ethernet Services Transport Protocol

This work has shown the improvement of ESTP over Traditional TCP by means of simulations and analytical derivations, but it is still pending a benchmark test.

10.3.2 Side-mounted VCSEL

It has been proven by experimentation that this method is a viable alternative, each individual experiment was a proof of concept. What is left is to perform a demo in which two chips are connected by side-mounted VCSELs coupled to polymer waveguides and the device still performs as if it had the electrical paths.

Appendix A

OPNET Modeler®

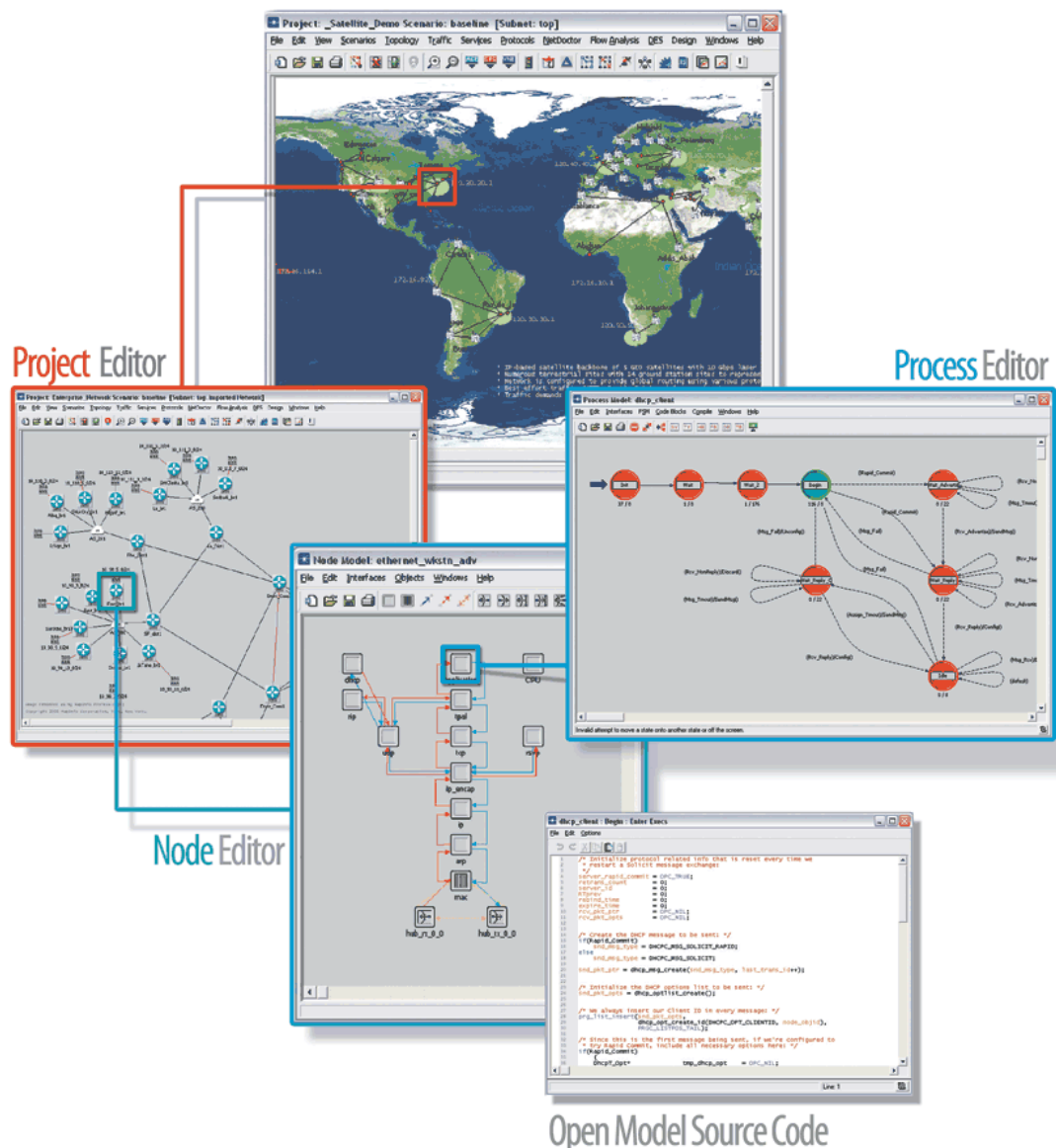
OPNET Modeler® is a powerful network simulation software, created for research and development (R&D). OPNET Modeler® accelerates the R&D process for analyzing and designing communication networks, devices, protocols, and applications. Users can analyze simulated networks to compare the impact of different technology designs on end-to-end behavior before deploying these ideas on the actual network, saving time and money in tests and giving the user a good idea of the way in which the network will react to these changes.

OPNET Modeler® has a hierarchical modeling environment. It is organized in three tiers with an intuitive graphical interface. These three domains are: Network, node and process (see Figure 32). There is another domain that could be considered part of the process domain, which is the source code.

The network domain contains nodes, links and subnets. Nodes can represent network devices or a group of devices. Examples of nodes are: a single workstation, a cluster of servers, and a router. Links can be point-to-point or can emulate a bus. A Subnet is a bundle of nodes. It behaves exactly as if the objects inside it were placed instead of the subnet. Subnets keep these groups of nodes organized and can be easily duplicated.

The node domain is where the basic building blocks, also called modules, are placed. The modules are mainly composed of processors, queues, and transceivers. Processors are fully programmable by means of their process model. Queues are used to buffer and manage data packets. Transceivers are the interfaces used to move data to or from the node domain.

The process domain is composed of states. There are two kinds of states, forced and unforced. When a forced state is entered it runs all the blocks of code inside it and it exits the state. An unforced state will perform the enter code and will remain idle until the proper interrupt triggers the exit sequence. Each state contains blocks of C code, which can be considered a lower level tier.



© OPNET Technologies, Inc. 2009. All rights reserved. Used with permission.

Figure 32. The different working domains of OPNET Modeler®.

Appendix B

Next-Generation Substrate Lab

The next-generation substrate lab is a cleanroom, which means its environment has a low level of environmental pollutants such as dust, airborne microbes, aerosol particles and chemical vapors. According to the US FED STD 209E cleanroom standard the next-generation substrate lab is a class 1000, however this standard was discontinued in 2001. The equivalent in the ISO 14644-1 standard, which is widely used worldwide, is class ISO 6. Class 1000 or ISO 6 has a controlled level of contamination that is specified by the number of particles per cubic meter at a specified particle size. The next-generation substrate lab has 35,000 particles per cubic meter of 0.5 μm (or larger) diameter. For the same diameter size particle, ambient air has 35,000,000 particles per cubic meter, i.e. 1000 times more polluted.

The air entering the cleanroom from outside is filtered to exclude dust, and the air inside is constantly circulated through high efficiency particulate air (HEPA) filters to remove internally generated contaminants. Staff enters and leaves through airlocks and wears protective clothing, which includes: hats, face masks, gloves, protective eyewear, boots and coveralls. The equipment inside the cleanroom is designed to generate minimal air contamination. Common materials such as paper, pencils, and fabrics made from natural fibers are often excluded; however, alternatives are available. The cleanroom is not sterile (i.e., free of uncontrolled microbes) and more attention is given to airborne particles. Particle levels are usually tested using a particle counter.

The cleanroom has a laminar flow of air that not only keeps the environment low on particles, but it creates a positive pressure so that if there are any leaks, air leaks out of the chamber instead of unfiltered air coming in.

Scheduling Schemes

C.1 Strict Priority Scheduling Scheme

Strict Priority Scheduling: This scheduling mode will always service traffic from the highest available priority queue that has an active service request.

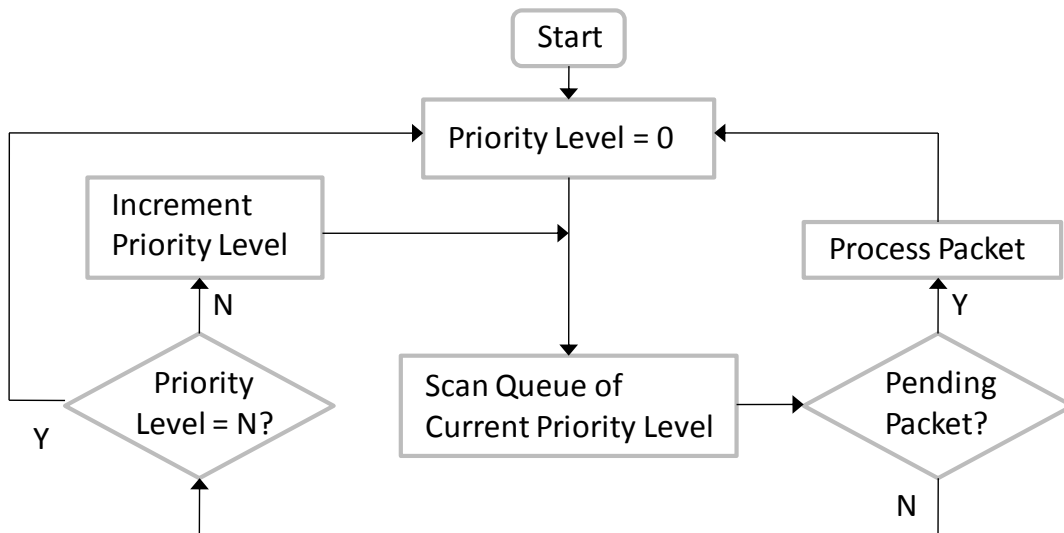


Figure 33. Strict priority scheduling flow diagram.

C.2 Weighted Fair Scheduling Scheme

Weighted Fair Scheduling: This scheduling mode will continue to service traffic from the current class of service queue until it has exhausted its credit allocation or no active requests remain; then, the weighted fair scheduler begins to service the next class of service.

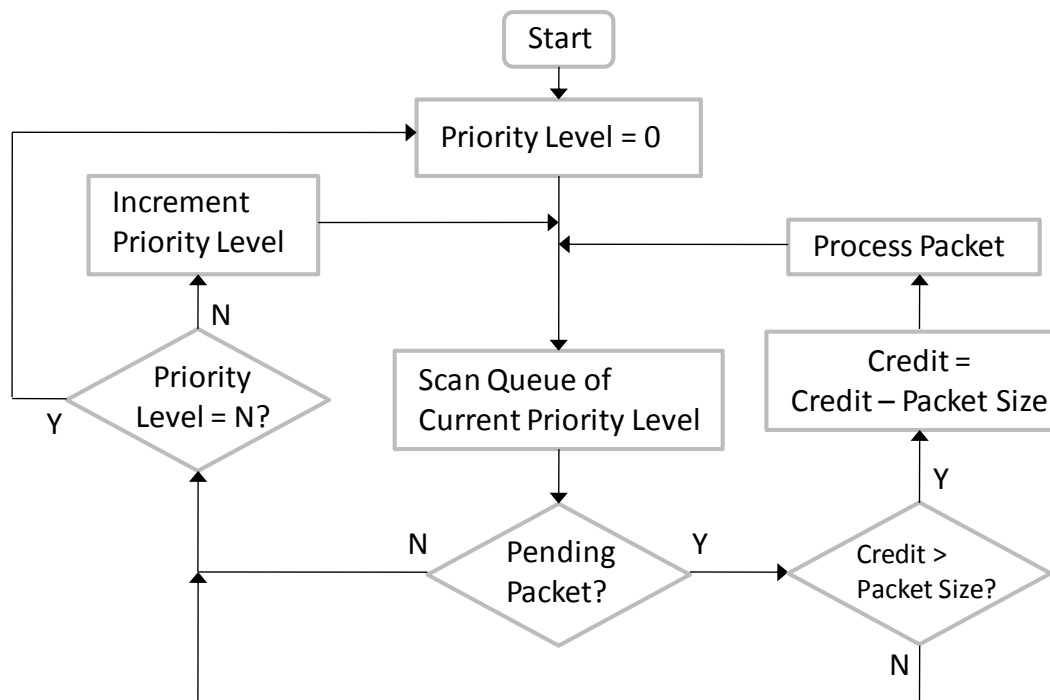


Figure 34. Weighted fair scheduling flow diagram.

Appendix D

Leaky-bucket Algorithm

The leaky-bucket algorithm is a way to manage the data rates that flow at some ingress point. This section should help understand better the functionality of this algorithm.

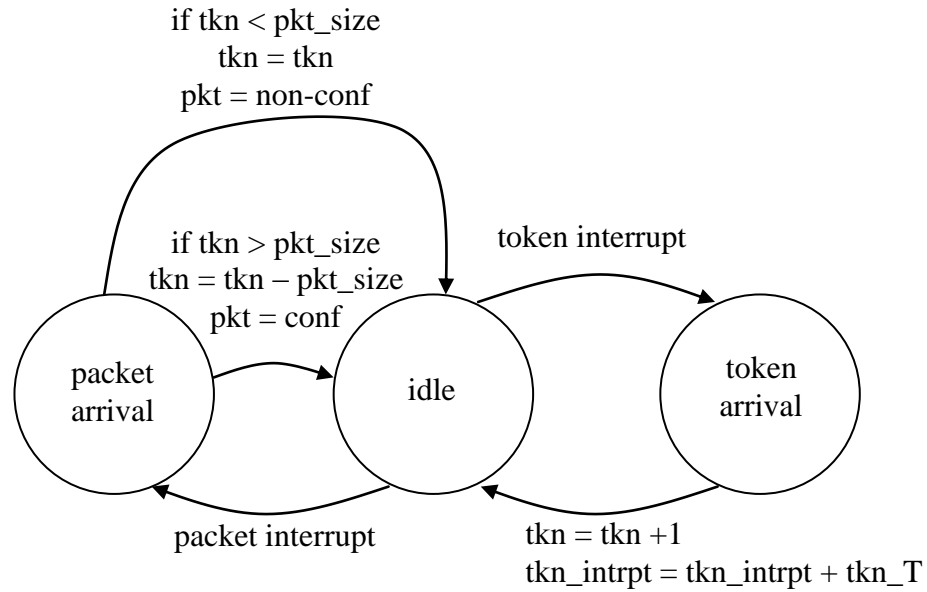


Figure 35. State Diagram of Leaky-bucket Algorithm

The state diagram is shown in Figure 35. The normal state of the leaky-bucket algorithm is the idle state. If a packet arrives an interrupt message is received and the algorithm goes into the packet arrival state. In this state the algorithm checks if the packet size (in bytes) is smaller than the amount of token available. Each token is equivalent to one byte. If the packet size is smaller, then the packet is considered conforming, and in general this implies it is sent to the network. Also, the number of tokens available is updated, which is the amount of tokens

available minus the packet size. In the event that the packet size is greater than the amount of tokens available then the packet is marked as non-conforming. This usually means the packet is discarded, treated as lower priority, or queued. A graphical representation of the algorithm can be seen in Figure 36.

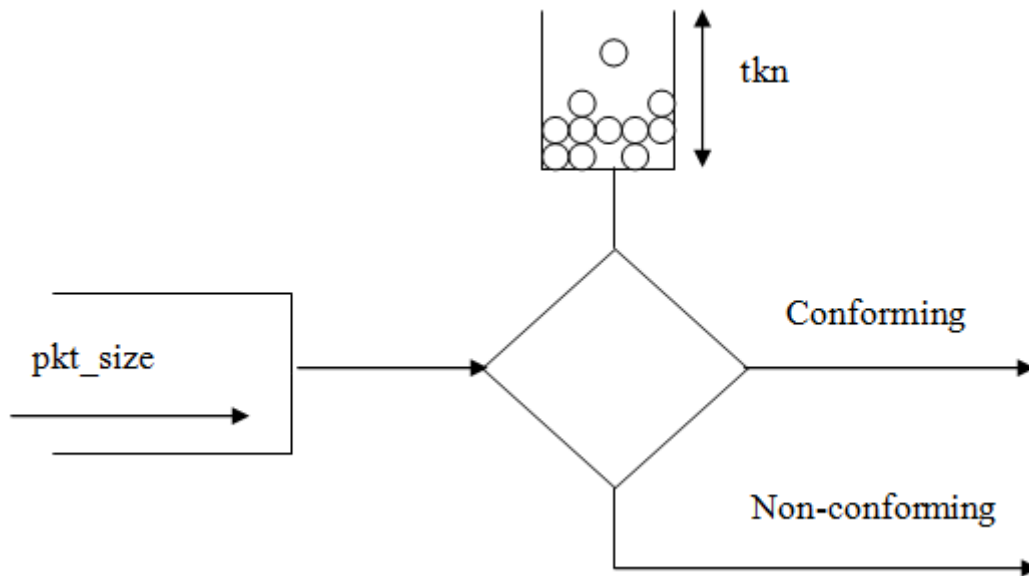


Figure 36. Graphical View of Leaky-bucket Algorithm

References

- [1] C. Estevez, G. Ellinas, G.-K. Chang, "Broadband Data Transport Protocol Designed for Ethernet Services in Metro Ethernet Networks," IEEE Globecom 2008, New Orleans, LA, November 2008.
- [2] C. Estevez, C. Xiao, G.-K. Chang, "Simulation Study of TCP Acceleration Mechanisms for Broadband Access Networks," OPNETwork 2006, Washington, DC, August 2006.
- [3] C. Estevez, D. Guidotti, G.-K. Chang, "A Novel Lightwave Device Integration and Coupling Process for Optical Interconnects," Electronic Components and Technology Conference, San Diego, CA, May 2009.
- [4] C. Estevez, G.-K. Chang, G. Ellinas, "Broadband Data Transport Protocol for Metro Ethernet Services," IEEE SouthEastCon 2009, Atlanta, GA, March 2009.
- [5] C. Xiao, G.K. Chang, B. Bing, "An SLA-aware Transport Protocol for High Throughput Wide Area Ethernet Services," IEEE GLOBECOM 2006, San Francisco, CA, November 2006.
- [6] E. Gubbins, "Carrier Ethernet's Growth Curve Continues," Telephony Online Magazine, Jan 2008.
- [7] M. Allman, V. Paxson, W. Stevens, "TCP congestion control," IEFT RFC 2581, April 1999.
- [8] K. Fall, S. Floyd, "Simulation-based comparisons of Tahoe, Reno and Sack TCP," Computer Communication Review, July 1996.
- [9] S. Floyd, "HighSpeed TCP for Large Congestion Windows," IEFT RFC 3649, December 2003.
- [10] T. Kelly, "Scalable TCP: Improving Performance in Highspeed Wide Area Networks," Computer Communication Review, Vol. 33, No. 2, April 2003, pp. 83-91.
- [11] D. Katabi, M. Handley, C. Rohrs, "Internet Congestion Control for Future High-bandwidth-delay Product Environments," Proceedings of ACM SIGCOMM '02, Pittsburg, PA. August 2002.

- [12] Cisco Systems Inc., "Internetworking Technologies Handbook," Cisco Press, 4th Ed., Ch. 7, September 2003.
- [13] D.A.B. Miller, "Rationale and Challenges for Optical Interconnects to Electronic Chips," Proceedings of the IEEE, Vol. 88, No. 6, June 2000, pp. 728-749.
- [14] C. Berger, B.J. Offrein, M. Schmatz, "Challenges for the Introduction of Board-Level Optical Interconnect Technology into Product Development Roadmaps," Proceedings of the SPIE - The International Society for Optical Engineering, Vol. 6124, No. 1, February 2006, pp. 61240J-1-12.
- [15] M. Oda, J. Sakai, H. Takahashi, H. Kouta, "Chip-to-Chip Optical Interconnection for Next-generation High-performance Systems," LEOS 2007. 20th Annual Meeting of the IEEE Lasers and Electro-Optics Society, 2007, pp. 638-639.
- [16] A.L. Glebov, M.G. Lee, K. Yokouchi, "Integration Technologies for Pluggable Backplane Optical Interconnect Systems," Optical Engineering, Vol. 46, No. 1, January 2007, pp. 15403-1-10.
- [17] L. Schares, et al. "Terabus: Terabit/Second-class Card-level Optical Interconnect Technologies," IEEE Journal of Selected Topics in Quantum Electronics, Vol. 12, No. 5, September 2006, pp. 1032-44.
- [18] S. Hiramatsu, M. Kinoshita, "Three-dimensional Waveguide Arrays for Coupling Between Fiber-optic Connectors and Surface-mounted Optoelectronic Devices," Journal of Lightwave Technology, Vol. 23, No. 9, September 2005, pp. 2733-9.
- [19] K. Nieweglowski, K.-J. Wolter, "Optical Analysis of Short-distance Optical Interconnect on the PCB-level," 2006 First Electronic System integration Technology Conference, IEEE Cat. No. 06EX1494, 2006, p. 6.
- [20] S. H. Hwang, M. H. Cho, S.-K. Kang, H. S. Cho, T.-W. Lee, H.-H. Park, "Optical Interconnection Platform Composed of Fiber-embedded Board, 90°-bent Fiber Block, and 10-Gb/s Optical Module," Journal of Lightwave Technology, Vol. 26, No. 11, June 2008, pp. 1479-85.
- [21] S. Hiramatsu, T. Mikawa, "Optical Design of Active Interposer for High-speed Chip Level Optical Interconnects," Journal of Lightwave Technology, Vol. 24, No. 2, February 2006, pp. 927-34.
- [22] R. Dangel, C. Berger, R. Beyeler, L. Dellmann, F. Horst, T. Lamprecht, N. Meier, B.J. Offrein, "Prospects of a Polymer-waveguide-based Board-level Optical Interconnect

- Technology,” Proceedings 11th IEEE Workshop on Signal Propagation on Interconnects, May 2007, pp. 131-4.
- [23] H. S. Cho, K.-M. Chu, S. Kang, S. H. Hwang, B. S. Rho, W. H. Kim, J.-S. Kim, J.-J. Kim, H.-H. Park, “Compact Packaging of Optical and Electronic Components for On-board Optical Interconnects,” IEEE Transactions on Advanced Packaging, Vol. 28, No. 1, February 2005, pp. 114-120.
 - [24] Y. Ishii, S. Koike, Y. Arai, Y. Ando, “SMT-compatible large-tolerance "OptoBump" interface for interchip optical interconnections,” IEEE Transactions on Advanced Packaging, v 26, n 2, May 2003, p 122-7.
 - [25] Han Seo Cho, Kun-Mo Chu, Saekyoung Kang, Sung Hwan Hwang, Byung Sup Rho, Weon Hyo Kim, Joon-Sung Kim, Jang-Joo Kim, Hyo-Hoon Park, “Compact packaging of optical and electronic components for on-board optical interconnects,” IEEE Transactions on Advanced Packaging, v 28, n 1, Feb. 2005, p 114-20.
 - [26] K. Nieweglowski, K.-J. Wolter, “Optical analysis of short-distance optical interconnect on the PCB-level,” 2006 First Electronic System integration Technology Conference, 2006, p 6.
 - [27] Byung Sup Rho, Saekyoung Kang, Han Seo Cho, Hyo-Hoon Park, Sang-Won Ha, Byoung-Ho Rhee, “PCB-compatible optical interconnection using 45°-ended connection rods and via-holed waveguides,” Journal of Lightwave Technology, v 22, n 9, Sept. 2004, p 2128-34.
 - [28] S. Hiramatsu, M. Kinoshita, “Three-dimensional waveguide arrays for coupling between fiber-optic connectors and surface-mounted optoelectronic devices,” Journal of Lightwave Technology, v 23, n 9, Sept. 2005, p 2733-9.
 - [29] A.L. Glebov, M.G. Lee, K. Yokouchi, “Integration technologies for pluggable backplane optical interconnect systems,” Optical Engineering, v 46, n 1, Jan. 2007, p 15403-1-10.
 - [30] L. Schares, et al. “Terabus: terabit/second-class card-level optical interconnect technologies,” IEEE Journal of Selected Topics in Quantum Electronics, v 12, n 5, Sept.-Oct. 2006, p 1032-44.
 - [31] M. Morimoto, K. Suematsu, R. Sugizaki, K. Takahashi, and H. Nasu, “90°-bent with R=1mm optical fiber technique for optical interconnection,” Proceedings of the SPIE - The International Society for Optical Engineering, v 6891, 7 Feb. 2008, 68910F-1-11.
 - [32] Shu-Hao Fan, D. Guidotti, C. Estevez, G.-K. Chang, Ying-Jung Chang, D.D. Lu, “Short-reach flexible optical interconnection using embedded edge-emitting lasers and edge-

- viewing detectors,” Proceedings of the SPIE - The International Society for Optical Engineering, v 6899, 7 Feb. 2008, p 689905-1-11.
- [33] J. Wang, “Optical Ethernet: Making Ethernet carrier class for professional services,” Proceedings of the IEEE, vol. 92, no. 9, pp. 1452-1462, September 2004.
 - [34] A. Meddeb, “Why Ethernet WAN Transport,” IEEE Communications Magazine, Vol. 43, No. 11, pp. 136-141, Nov. 2005.
 - [35] D. Ferrari and D.C. Verma, “A Scheme for Real-time Channel Establishment in Wide-area Networks,” IEEE Journal on Selected Areas in Communications, vol. 8, no. 3, pp. 368-79, April 1990.
 - [36] H. Chamas, W. Bjorkman, and M.A. Ali, “A Novel Admission Control Scheme for Ethernet Services,” Proc. IEEE ICC 2005, pp. 65-69, Seattle WA, May 2005.
 - [37] T.M. Cover, J.A. Thomas, Elements of Information Theory, New York: John Wiley & Sons, 1991.
 - [38] 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, pp. 1-15, Feb. 1994.
 - [39] J. Padhye, et. al., “Modeling TCP Throughput: A Simple Model and its Empirical Validation,” Proc. ACM SIGCOMM’98, Vancouver, Canada, September 1998.
 - [40] A. Falk, T. Faber, J. Bannister, A. Chien, R. Grossman, and J. Leigh, “Transport Protocols for High Performance,” Communications of the ACM, vol. 46, no. 11, pp. 43-49, 2002.
 - [41] Y. Zhang and T. R. Henderson, “An Implementation and Experimental Study of the eXplicit Control Protocol (XCP),” Proc. IEEE Infocom 2005, Miami, FL, March 2005.
 - [42] A. Kesselman, et al., “Adaptive AIMD Congestion Control”, Proc. 22nd Annual Symposium on Principles of Distributed Computing, pp 352-359, Boston, MA, July 2003.
 - [43] R. Hogg, J. McKean, and A. Craig, Introduction to Mathematical Statistics, Sixth Edition, Prentice Hall, 2004.
 - [44] E. Blanton, M. Allan, K. Fall, and L. Wang, “A Conservative Selective Acknowledgment (SACK)-based Loss Recovery Algorithm for TCP,” IETF RFC 3517, April 2003.
 - [45] D. Katabi, M. Handley, and C. Rohrs, "Internet congestion control for future high-bandwidth-delay product environments," Proceedings of ACM SIGCOMM'02, Pittsburgh, PA, Aug. 19-21, 2002.

- [46] Q. Wu, Control of Transport Dynamics in Overlay Networks. Ph.D. thesis, Dept of Computer Science, Louisiana State University, March 2003.
- [47] Vertical Systems Group, "Business Ethernet Expands 43% in 2008," Vertical Systems News, Feb 2009 (<http://www.verticalsystems.com/news.html>).
- [48] Vertical Systems Group, "Worldwide Business Ethernet Services Market Rises to \$38.9 Billion by 2013," Vertical Systems News, March 2009 (<http://www.verticalsystems.com/news.html>).
- [49] J. Postel, "Transmission Control Protocol," IETF RFC 793, Sept. 1981.
- [50] R. Braden, ed., "Requirements for Internet Hosts – Communication Layers," IETF RFC 1122, Oct. 1989.
- [51] V. Jacobson, S. Braden, D. Borman, "TCP Extensions for High Performance," IETF RFC 1323, May 1992.
- [52] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, "TCP Selective Acknowledgement Options," IETF RFC 2018, Oct. 1996.
- [53] M. Allman, V. Paxson, W. Stevens, "TCP Congestion Control," IETF RFC 2581, Apr. 1999.
- [54] V. Paxson, M. Allman, "Computing TCP's Retransmission Timer," IETF RFC 2988, Nov. 2000.
- [55] M. Allman, S. Floyd, C. Partridge, "Increasing TCP's Initial Window," IETF RFC 3390, Oct. 2002.
- [56] S. Floyd, T. Henderson, A. Gurtov, "The New Reno Modification to TCP's Fast Recovery Algorithm," IETF RFC 3782, Apr. 2004.
- [57] Y. R. Yang and S. S. Lam. "General AIMD congestion control." Technical Report TR-2000-09, The University of Texas at Austin, May 2000.
- [58] R. Rejaie, M. Handley, and D. Estrin. "RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the Internet." Proceedings of IEEE INFOCOM '99, volume 3, March 1999.
- [59] D. Sisalem and H. Schulzrinne. "The loss-delay based adjustment algorithm: A TCP-friendly adaptation scheme." Proceedings of NOSSDAV '98, Cambridge, UK, July 1998.
- [60] J. Kurose, K. Ross, "Computer Networking: A Top-Down Approach Featuring the Internet," 3rd Ed., Pearson Addison-Wesley, 2005.
- [61] R. Tummala, M. Swaminathan. "Introduction to System-on-package (SOP): Miniaturization of the Entire System," McGraw-Hill Companies, Inc., Chapter 6, 2008.

- [62] C. Forrest Tomes. "New avionics ATE considerations for the Boeing 777," Proceedings ATE and Instrumentation Conference, p 491-7, 1991.
- [63] A. Suzuki, et al. "High optical coupling efficiency using 45°-ended fibre for low-height and low-cost optical interconnect modules," *Electronics Letters*, v 44, n 12, p 724-5, 5 June 2008.
- [64] M. Kanda, T. Ogawa, O. Mikami, "VCSEL module with polymer optical output rods to enable high efficiency coupling for optical interconnection," *IEEE Photonics Technology Letters*, v 21, n 11, p 685-7, 1 June 2009.
- [65] M. Popall, et al., "ORMOCERSTM-New Photo-patternable dielectric and optical materials for MCM-packaging," Proceedings of 48th Electronic Components and Technology Conference, Seattle, WA, p. 1018-1025, May 25-28, 1998
- [66] M. Moynihan, et al., "Hybrid inorganic-organic aqueous base compatible waveguide materials for optical interconnect applications," Proceedings of SPIE, vol. 5212, p. 50, 2003.
- [67] Metro Ethernet Forum, "Metro Ethernet Services Definitions Phase 2," MEF 6.1, June 2008. (http://www.metroethernetforum.org/PDF_Documents/MEF6-1.pdf)
- [68] A. Demers, S. Keshav, S. Shenker, "Analysis and simulation of a fair queueing algorithm," *Computer Communication Review*, v 19, n 4, p 1-12, Sept. 1989.
- [69] 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
- [70] H. Stark, J. Woods, "Probability and Random Processes with Applications to Signal Processing," 3rd Edition, Prentice-Hall, Inc. 2002.
- [71] R. Santitoro, "Metro Ethernet Services – A Technical Overview," www.metroethernetforum.org/metro-ethernet-services.pdf, Metro Ethernet Forum, 2003.
- [72] Gubbins, E., "Carrier Ethernet's Growth Curve Continues," *Telephony Online Magazine*, Jan 2008.
- [73] <http://www.telecoms.com/itmgcontent/tcoms/features/articles/20017499804.html>
- [74] http://www.deepikaglobal.com/archives/ENG5_sub.asp?newsdate=03/24/2008&cocode=ENG5&hcode=16623
- [75] Metro Ethernet Forum, "Metro Ethernet Services Definitions Phase 2," MEF 6.1, June 2008. (http://www.metroethernetforum.org/PDF_Documents/MEF6-1.pdf)
- [76] <http://metroethernetforum.org/NewsPressRoom>

Vita

Claudio Estevez was born in Caracas, Venezuela in 1977. He received a B.S. degree (2001) in Electrical Engineering from the University of Puerto Rico (Mayaguez, PR) in the areas of communications and control systems. He received an M.S. degree (2003) in Electrical Engineering from the University of Alabama in Huntsville (Huntsville, AL) in the area of optical communication. Claudio Estevez was admitted to the PhD program in the School of Electrical and Computer Engineering of the Georgia Institute of Technology (Atlanta, GA) where he joined the Optical Networking Research Group (ONRG) in 2005 under the supervision of Dr. Gee-Kung Chang. During the early years with ONRG, Claudio worked mostly on Broadband Access Networks. In the later years, his focus turned to Carrier Ethernet Networks and optical interconnects were his thesis research topic took place.

Throughout the years Claudio Estevez has worked with many organizations. In 1997 he worked with the University of South Carolina (Columbia, SC) researching the potential of lithium batteries. In 1998 he worked for the Federal Aviation Administration (FAA) assigned to a project that dealt with remote monitoring of power generators. In 1999-2000 he researched the potential of nitinol in miniature motorless robots. In 2001 he went to the Georgia Institute of Technology to research high temperature ceramic pressure sensors under the supervision of Dr. Mark Allen, founded by the SURE program hosted by Dr. Gary May. In 2002 he worked on an air-ground unmanned vehicle project at the University of Alabama in Huntsville funded by the Army. In 2003 he returned to the FAA where he was partially in charge of the approval on a novel landing system called Transponder Landing System (TLS). In 2008 Claudio worked for Rockwell Collins, Inc. in Cedar Rapids, IA in an optical power amplifier project funded by the Defense Advanced Research Projects Agency (DARPA). From 2003 to 2009 he worked mainly on graduate research projects and teaching assistantships at the Georgia Institute of Technology, including work related to fiber optic communication, radar communication, optical interconnect, circuit laboratory, transport control protocol (TCP), and Carrier Ethernet networks. These were funded by organizations such as Georgia Institute of Technology – ECE department, National Science Foundation (NSF), Society of Hispanic Professional Engineers (SHPE), and by the Georgia Institute of Technology Presidential Fellowship.

List of Publications

Journal Papers

1. C. Estevez, G. Ellinas, G.-K. Chang, "High Performance QoS-supporting Transport Protocol for Carrier Ethernet Network," IEEE/ACM Transactions on Networking. [submitted 2009]

Conference Papers

2. C. Estevez, G. Ellinas, G.-K. Chang, "Analytical Characterization of the Ethernet Service Transport Protocol Throughput," IEEE International Communications Conference 2010, Cape Town, South Africa, May 2010. [submitted 2009]
3. C. Estevez, G. Ellinas, G.-K. Chang, "Improving Ethernet Services Transport Protocol Performance in Carrier Ethernet Networks," IEEE Infocom 2010, San Diego, CA, March 2010. [submitted 2009]
4. C. Estevez, D. Guidotti, G.-K. Chang, "A Novel Lightwave Device Integration and Coupling Process for Optical Interconnects," Electronic Components and Technology Conference, San Diego, CA, May 2009.
5. C. Estevez, G.-K. Chang, G. Ellinas, "Broadband Data Transport Protocol for Metro Ethernet Services," IEEE SouthEastCon 2009, Atlanta, GA, March 2009.
6. C. Estevez, G. Ellinas, G.-K. Chang, "Broadband Data Transport Protocol Designed for Ethernet Services in Metro Ethernet Networks," IEEE Globecom 2008, New Orleans, LA, November 2008.
7. S.-H. Fan, D. Guidotti, C. Estevez, G.-K. Chang; Y.-J. Chang; D.D. Lu, "Short-reach flexible optical interconnection using embedded edge-emitting lasers and edge-viewing detectors," Proceedings of the SPIE - The International Society for Optical Engineering, v 6899, 7 Feb. 2008, p 689905-1-11
8. C. Xiao, C. Estevez, G. Ellinas, G.-K. Chang, "A Resilient Transport Control Scheme for Metro Ethernet Services Based on Hypothesis Test," IEEE Globecom 2007, Washington, DC, November 2007.
9. C. Estevez, J. Handschuh, "Simulation Study of the Ant-based Routing Protocol," OPNETwork 2007, Washington, DC, August 2007.
10. C. Estevez, C. Xiao, G.-K. Chang, "Simulation Study of TCP Acceleration Mechanisms for Broadband Access Networks," OPNETwork 2006, Washington, DC, August 2006.
11. C. Xiao, C. Estevez, G.-K. Chang, "Performance Evaluation of an SLA-Aware Transport Control Protocol for Ethernet Services," OPNETwork 2006, Washington, DC, August 2006.