A TRULY END-TO-END GLOBAL MULTISERVICE OPTICAL ETHERNET NETWORKING ARCHITECTURE

By

HAIDAR CHAMAS

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy

The City University of New York

2006

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

SUPERVISION COMMITTEE

Date	*Name*
December 15, 2005	Professor Mohammed A. Ali Chair of Examining Committee
December 15, 2005	Dean Mumtaz K. Kassir Executive Officer

Name	*Department*
Professor Roger Dorsinville	Department of Electrical Engineering, The City College of New York, CUNY
Professor Kaliappa Ravindran	Department of Computer Science, The City College of New York, CUNY
Professor Neophytos Antoniades	Department of Engineering, Science and Physics, The College of Staten Island, CUNY
Dr. Stuart Elby	Verizon Communications

Supervision Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

A TRULY END-TO-END GLOBAL MULTISERVICE OPTICAL ETHERNET NETWORKING ARCHITECTURE

By

Haidar Chamas

ADVISER: PROFESSOR MOHAMED A. ALI

This thesis examines the technological requirements and assesses the performance analysis and feasibility for implementing a truly native end-to-end Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. Specifically, this work proposes and devises both short and long-terms innovative, graceful networking transition scenarios to evolve Ethernet into a next generation networking technology that can truly support carrier-class Ethernet services.

The proposed short-term solution addresses the immediate need for legacy systems to support evolving Ethernet services. It utilizes current Layer 2 technology and expands it into the MAN, WAN and long haul networks over DWDM or MPLS Core infrastructures. It also addresses the current Ethernet shortcomings such as spanning tree protocol and the lack of QoS support required by most applications. Specifically, we introduce a new admission control mechanism that addresses key challenges within an end-to-end service architecture. Hence, we define a novel E-UNI and E-NNI with QoS support. The CoS service performance models proposed by this research is vital for expanding Metro Ethernet into the Core and Long Haul as well as provide the opportunity to standardize

Ethernet services (voice, data and video) in the metro, wide, national and global networks.

The proposed long-term solution presents an ambitious vision of how to implement a truly native end-to-end layer-2 MAC frame-based Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. We show that by combining the simplicity and cost effectiveness of Ethernet technology with the ultimate intelligence of WDM-based optical transport layer, Optical Ethernet (Ethernet-over-WDM) could evolve as a next generation networking paradigm that provides a seamless global transport infrastructure for end-to-end transmission of native Ethernet frames.

The proposed Optical Ethernet networking architecture is a true two-layer model, realizing the significant goal of Ethernet-over-WDM, where native Ethernet frames are mapped directly over WDM. It offers significant advantages over existing Layer-2 and MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other layer-2 protocols.

The primary rational behind our vision is decreased cost and complexity in the network. This is how Ethernet won the LAN years ago; it was not necessarily the best technology, it was the most cost-effective and easily implemented. Simplifying network design and reducing costs by utilizing Ethernet as an end-to-end LAN/MAN/WAN protocol is the key for Ethernet to win the MAN and WAN.

ACKNOWLEDGMENTS:

I have been fortunate to have Professor Mohamed A. Ali as my adviser. Professor Ali have been a true inspiring figure in the last three years and provided me the freedom to work on this truly novel idea, shared his great wealth of information in the optical area, and guided me through all of the requirements that enabled me to succeed.

I am thankful to have Dean Mumtaz Kassir as the Dean of the school of Engineering for his encouragement and patience, and his assistant Belkys Bodre for always being there.

I am grateful to my PhD supervision committee for their support and helpful comments, including: Professor Roger Dorsinville, Professor Kaliappa Ravindran, Professor Neophytos Antoniades, Professor Mohammed Ali, and Dr. Stuart Elby of Verizon Communications.

I would like to thank William Bjorkman for his invaluable comments and suggestions in reviewing this thesis. Also, I am grateful for the time we spent in exchanging ideas that inspired me to develop some of the key ideas and contributions made in this research.

I would like to thank Eugene Lubchenko for his support and invaluable knowledge of Microsoft word in developing the thesis template and for demonstrating lots of his nifty routines that helped format the document uniformly across all of the chapters.

I would like to thank my colleagues Vincent Alesi, Lily Chen, Harry Chu, Roman Krzanowski, and Hassan Omar for their review of specific sections in this thesis, and their technical advice and feedback, which have been valuable.

And last but not least, I would like to thank my beloved wife, Maryam Kermani, for her patience and helpful review of the entire thesis and for being a constant source of encouragement. Without her support, this thesis would not have been a reality.

I dedicate this thesis to my father for his endless motivation, encouragement, trust, and for never giving up on me.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1. Introduction

1.1 Overview

Ethernet has been the dominant technology of the LAN networks for over 20 years. The main driver behind Ethernet development was to share printers and computer resources in an office area and to connect them over a wire (segment), where each device attaches (clamps) to the wire to form a bus topology. Later, this prototype led to the initial Ethernet network that used a shared (passive) bus operating at 10 Mbps.

As Ethernet dominated and became the ubiquitous technology in the LAN environment, more and more campuses and enterprises began to pressure the Ethernet vendors and service providers to develop ways to cover multiple buildings across a metro or wide area network. This continues to be a key dominant force behind pushing Ethernet to evolve from the LAN to the Metro in order to span a large geographic area. An Enterprise with multiple LANs was then able to expand its network, by spanning multiple buildings within a Metro, using repeaters (Hubs) or bridges via dedicated private lines.

Today, enterprises have the ability to expand their LANs via a Service Provider into a metro or national area by selecting the appropriate Ethernet service.

LAN segmentation is a key enabler for expansion of Ethernet networks outside of the LAN. The process of splitting LAN domains into two or more separate domains allows a network to span multiple geographic areas and to grow beyond the inherent LAN limitation. Fiber access and coax transport infrastructure is used to span a large geographic area, without compromising the bandwidth needed due to distance limitation

on the type of MDI[1] used [94]. LAN segmentation is achieved by creating logical grouping of users across multiple locations through defining a community of interest or closed user groups. This community of interest is called a Virtual Local Area Network (VLAN). Hence, a VLAN is a logical segmentation of users, by a specific function, regardless of their physical location or access speeds. A VLAN may span a single switch, multiple switches, or across multiple MANs or WANs. The main advantage of a VLAN is its ability to provide segmentation, flexibility and security.

Recent Ethernet expansion into Metro Area Networks (MANs) and Wide Area Networks (WANs) is fueled by the rise of enterprise customer demand for faster networking speeds that traditional and legacy networks are unable to scale to or are too expensive to implement. Ethernet with its low cost, simplicity and ability to scale from 10M to 10GigE has led to the introduction of metro Ethernet services. Also, Ethernet is geographically independent technology and its 802.3 frame is efficient and IP friendly [94]. Service providers' Ethernet services have been growing over the last couple of years at over 50% Cumulative Annual Growth Rate (CAGR). Industry market analysts are predicting that the Ethernet services market will exceed $1B by 2008 [75], [53]. This prediction is consistent with some service providers' experience and projections of double-digit growth trend over the next few years. In addition, recent forecast revision maintains a strong forecast for Ethernet services with a 55% CAGR from 2003 to 2008 or about $1.5B by the year 2008 and 70% CAGR for the Dedicated Internet Access (DIA) or an

[1] Ethernet varying speeds are based on Medium-Dependent Interfaces (MDIs) and are defined by the IEEE 802 physical layer standards. The IEEE standards helped push the emergence of new Ethernet MDI interfaces that included coax, fiber, twisted pair, and wireless.

estimated $2.6B by the year 2008. Hence, Ethernet has evolved tremendously and it no longer resembles its original network concept. It has survived even the most demanding changes, retained its name and added new functionality and capability.

Ethernet continues to go through significant changes to meet new demands such as those spurred by a combination of Service Provider (SP) needs, to meet their enterprise customer requirements, and the technology innovations in the 1990s aimed at improving local area networking scale and performance. Recent years have seen great efforts dedicated to evolving Ethernet-based technology. These efforts have been very important and beneficial in maintaining Ethernet strength in the LAN and in taking it into the MAN and WAN. Amongst the key innovations, standardized by IEEE 802.1 and 802.3, include:

- Ethernet VLAN switching
- Introduction of high speed native Ethernet links – such as 1 Gigabit and 10 Gigabit links
- Optical reach extending up to 100 Km

In addition, the inherent benefits of Ethernet make it a natural evolution of the technology beyond the LAN. The key advantages and benefits that position Ethernet in the metro include:

- Standard access interface with great scalable speeds
- Simple, plug and play at low cost
- Fast service delivery and ease of use

Advancements in Ethernet technology, such as high bandwidth and long reach optics have enabled a whole new generation of Ethernet MANs. These advancements have

provided service providers with a new opportunity to offer native Ethernet services over large metropolitan areas (network diameters of > 100 Km), which allows customers to tie their LAN segments within a metro as if they have connected them through a bridge or a switch. Further technology innovations are allowing service providers to offer Ethernet services to be transported over a variety of switching technologies, including SONET/SDH, RPR, xWDM, dedicated fiber and MPLS systems, which effectively extend the network reach for a service provider nationally.

Despite the promise of Ethernet and its potential of becoming the universal telecom service, a number of challenges remain before Ethernet technology can support true carrier-class Ethernet services. These include:

- End-to-end quality of service guarantees
- Scalability and spanning tree issues
- Rapid, SONET like, restoration

1.2 Thesis Motivation

The tremendous acceptance of Gigabit Ethernet in most Local Area Networks (LANs) has created pressure on carriers and service providers (SPs) to offer native Ethernet services at gigabit rates in the metropolitan area networks (MANs) environment. Enterprises traffic is rapidly transcending existing capabilities, becoming more complex and data-centric with growing percentage of this traffic being time sensitive. Recent advances in Ethernet and optical networking technologies, such as the 10 Gigabit Ethernet standard and long reach optics have enabled a whole new generation of Ethernet networking pushing Ethernet into the MAN and WAN. Further technology innovations

are allowing for Ethernet services to be transported over a variety of media, including SONET/SDH, WDM and Multiple Protocol Label Switching (MPLS) systems, which effectively extend the reach (point-to-point) nationally and globally. Many service providers are looking to expand their metro networks to very large scales or perhaps to inter-metro areas.

The fundamental problem is that the majority of today's MANs are built on legacy Synchronous Optical Networks/Synchronous Digital Hierarchy (SONET/SDH) ring infrastructures. SONET, and its international variant SDH, are deployed throughout the globe. They are the world's defacto standard for physical layer optical transport. Typical legacy SONET/SDH MAN architecture consists of metro core and metro access rings interconnected by a combination of SONET/SDH add/drop multiplexers (ADMs) and digital access cross-connect systems (DACS). SONET/SDH systems allow the transport of voice as well as data traffic, such as IP, frame relay (FR), Ethernet, and Asynchronous transfer mode (ATM) over fiber optics. Optimized for slow growing, narrowband, circuit-switched voice traffic, these networks lack the dynamic functionality and rapid scalability needed to keep pace with the increasing volumes and unpredictability of data traffic.

Data transport is characterized by a lack of statistical multiplexing functionality meaning data transmissions are forced into rigid 51.84 Mbps STS-1/VC-4 increments even though they often occupy less than 20 percent of the available bandwidth. This leads to scalability problems requiring either upgrading existing rings via forklift change-outs of all infrastructure components, or adding dark fiber to deploy additional rings with additional equipment. The voice-optimized nature of this network means that data traffic requires additional switches or routers to map data into time division multiplexed (TDM)

channels for transport across the SONET/SDH network. The result is a complex, multi-tiered, hierarchical architecturally constrained network, ineffective for data-centric metro environments.

Even though ATM and FR are widely used data transport protocols today, circuit switching equipment still processes approximately 80% of the carrier traffic mapped onto legacy SONET networking infrastructure. A new, converging set of metro access, metro transport and metro core service requirements is exposing this equipment's inherent limitations. Traffic in this "Trans-Metro" market is rapidly transcending existing capabilities, becoming more complex and data-centric; and a growing percentage of this traffic (especially that generated by mission critical applications) is delay or time sensitive.

To address many of these limitations, a new set of enhancements (next-generation SONET/SDH) that enable flexible and reliable data transport over SONET have been developed. These include generic framing procedure (GFP), virtual concatenation (VCAT), and link capacity adjustment schemes (LCAS). These developments provide an efficient and standard means of carrying data signals over existing SONET transport networks. However, data signals are still further encapsulated into another new framing format (GFP), which, in turn, are mapped into the SONET frames. The end result is still the same complex, multi-tiered constrained network. Thus, neither traditional SONET ADMs nor next-generation SONET solutions meet all the requirements of the new packet-based network dynamics.

Another emerging option for avoiding the limitations of the rigid circuit-based SONET solutions, is to migrate packet-based technologies from the traditional LAN to the MAN

and WAN. Ethernet is uniquely positioned as the leader for the inexpensive and flexible transport of packet-based technologies. Research shows that the use of Ethernet ports provides up to a 91% Capital Expenditures (CAPEX) savings compared to SONET/SDH because Ethernet interfaces are typically 25 percent to 40 percent less expensive per Mbps of bandwidth than lower speed TDM and ATM ports [98]. According to recent, independent studies conducted by both Gartner and Yankee Groups, Ethernet switching costs will continue to decrease approximately 30 percent annually [98], [59].

Despite the promise of Ethernet and the potential of becoming the universal telecom service, a number of challenges remain before Ethernet technology can support true carrier-class Ethernet services, including rapid, SONET like restoration, the ability to support voice, video and data at wire speed with a predictable quality of service, and greater scalability than traditional Ethernet solutions. Specifically, the vision of implementing a native Ethernet MAC frame-based global transport networking infrastructure that utilizes only pure layer-2 switching across the MAN and WAN still faces several key technical hurdles, including:

1. **End-to-end QoS guarantees**: Currently Ethernet provides best effort traffic delivery. Although IEEE 802.1Q specifies three priority bits, Ethernet has no "hard" quality of service provisioning with guarantees. This requires Ethernet to support end-to-end signaling coupled with control plane support. Ethernet does offer "soft" QoS mechanism such as differentiated services (DiffServ), and therefore can mark packets for prioritization, scheduling, and policing, however, Ethernet's QoS support is not mature enough for a multi-customer environment as in a service provider's environment. Ethernet's lack of

inherent QoS capabilities raises the critical issue of whether a given request to provision a new Ethernet Virtual Circuit (EVC) with a specific QoS requirement can be admitted without compromising the service performance of already provisioned EVCs.

2. **Resilience**: Fast fault detection and recovery are required which is not possible with current spanning tree algorithms. To provide redundancy in Metro Ethernet networks, STP is used to provide loop free connectivity while providing an alternate path if a link, port or switch fails. However, the standard STP imposes several limitations. First, it provides a convergence time of 30 to 50 seconds. Second, because STP blocks links to prevent loops, it can lead to inefficient utilization of expensive fiber links in the Metro network. We can run one STP instance per each VLAN in order to utilize all the fiber in the Metro network. However, as the Metro networks grow and scale to accommodate more customers, the number of STP instances and span of the STP instance can become a bottleneck. Rapid Spanning Tree protocol (RSTP) (IEEE 802.1w), builds upon the original 802.1D STP standard, and was introduced to achieve faster recovery time. However, most implementers remain convinced that RSTP recovery is rarely much below the one second mark—far slower than the carrier standard.
3. **Scalability:** Ethernet's major scalability bottleneck is the use of a spanning tree to forward traffic and the lack of load balance. Second issue is the flat addressing structure and lack of routing hierarchy, which can lead to very large routing tables.

- **Spanning tree issues:** A single tree allows only one loop-free path, which can result in uneven load distribution and potential bottlenecks. Multiple spanning trees (MST), an IEEE 802.1s standard, addresses STP limitations by allowing more links to be utilized in the network, but still has its own scalability problems and is not CoS aware.
- **Addressing structure issues:** Lack of routing hierarchy leads to very large routing tables. Introducing the virtual LAN (VLAN) concept in an enterprise domain provides the ability to logically partition distinct user groups over the same physical network. However, the limited VLAN tag space (the IEEE 801.Q standard defines an address space of only 4096 available tags) still poses major scalability bottleneck.
- **Lack of fault isolation capability:** Ethernet has no built-in alarms, such as SONET's Loss of Signal (LOS) and Remote Defect Indicator (RDI), which allow fault isolation to the malfunctioning section, line, or path. However, Ethernet Link OAM per IEEE 802.3ah does provide Remote Fault Indication.

4. **Frame Size Limitations**: Whether or not Gigabit Ethernet should support frame sizes larger than 1518 bytes has been a topic of hot debate [97], [98]. Alton's [98] proposed to increase the Ethernet frame size from 1518 to 9000 bytes (jumbo frame). Most of the debate about jumbo frames has focused on LAN performance and the impact that frame has on host processing requirements, interface cards, memory, etc. [97], [98]. The impact that frame size has on MAN/WAN performance has not received any attention. With

Ethernet operating at an aggregated high bit rate (multi-Gigabit) in the WAN environment, a small frame size might be an issue.

1.3 Thesis Statement

Enterprise data traffic is nearly all Ethernet. But once it leaves the corporate LAN (or campuses) and heads onto the wide area, it's translated into some other protocol, only to be translated back into Ethernet once it reaches its destination. What's lost in the translation, in this instance, is time and money. These conversions are inefficient and expensive, requiring specialized software on both carrier and customer switches. They are also unnecessary—if native Ethernet can be transported end to end. In addition, Ethernet framing preserves VLAN IDs, QoS tags, and virtually every other packet-level control function available at the MAC layer. Since well over 90 percent of all data traffic originates and ends on an Ethernet LAN, the envisioned data-centric next generation networking infrastructure must have the capability of transporting native Ethernet frames across any segment of the network. Thus, transporting native Ethernet frames end to end from the access network through the metro and core networks to another access network is the most cost effective, simple, and efficient solution.

It is the main objective of this thesis to scale metro Ethernet networks into a global multi-services infrastructure. Specifically, this work proposes and devises both short and long-terms innovative, graceful networking transition scenarios to evolve Ethernet into a next generation networking technology that rival Frame Relay, ATM and Private Lines. The proposed short-term solution addresses the immediate need for legacy systems to support evolving Ethernet services. It utilizes current Layer 2 technology and expands it into the MAN, WAN and long haul networks over DWDM or MPLS Core infrastructures. It also

addresses the current Ethernet shortcomings such as spanning tree protocol and the lack of QoS support required by most applications.

The main characteristics of the proposed short-term Ethernet networking solution are:

1. The introduction, for the first time to the best of our knowledge, an Admission Control Scheme for Ethernet Services (ACES). ACES is a centralized system that works closely with the Ethernet Provisioning Server (EPS). This solution is used in conjunction with Multiple Spanning Trees (MST) to control EVC admission into the SP's network and determine the most efficient path selection through the network.
2. The proposed ACES solution provides a centralized admission control and path selection function for provisioning each EVC in a Service provider's switched Ethernet network, while controlling CoS service level guarantees.
3. The proposed ACES scheme could provide the entire EVC provisioning capability for a large scale, hybrid, Ethernet-over-WDM-based optical transport network. These include: a) Building a dynamic optimum layer-2 logical topology on top of the OTN; b) Identifying specific trunks connecting GigE switches (logical connections) with bandwidth limitations and/or underutilized trunks and, then, signal the OTN to add/delete a λ (lightpath) on these trunks or communicate directly with the optical network elements; c) Providing the signaling reservation to the optical elements or interface with a separate Provisioning Server and optical control system.

4. Presentation of a Dynamic Queue Weighting (DQW) Algorithm that modifies the Weighted Round Robin (WRR) weights on the silver and bronze queues to optimize the utilization of the network resources.

These proposed short-term solutions support true carrier-class Ethernet services, including the ability to support voice, video and data at wire speed with a predictable quality of service, provide greater scalability than traditional Ethernet solutions and at the same time can coexist with both legacy Ethernet services and switching technology

The long-term solution abandons the conventional strategy that has always addressed existing layer-2 limitations and introduces drastic changes to legacy Ethernet technology and services. Specifically, we present an ambitious vision of how to implement a truly native end-to-end layer-2 MAC frame-based Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. We show that by combining the simplicity and cost effectiveness of Ethernet technology with the ultimate intelligence of WDM-based optical transport layer, Optical Ethernet (Ethernet-over-WDM) could evolve as a next generation networking paradigm that provides a seamless global transport infrastructure for end-to-end transmission of native Ethernet frames.

Unlike today's notion of supporting "IP directly over WDM" (IP/MPLS-over-WDM interconnection models), which is little more than cleverly disguised marketing; "IP-over-WDM" for example, is almost invariably IP packets mapped into SONET/SDH, coupled with SONET/SDH-based point-to-point DWDM systems [97], [98], [59]-[61]. The proposed "Ethernet-over-WDM" model is truly a two-layer model where native Ethernet frames are mapped directly over WDM. It offers advantages over existing Layer-2 and

MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other layer-2 protocols.

The overall long-term research vision includes a simplified network model that [11]:

- Integrates L1/L2 control planes with a CAC function for bandwidth on demand (BoD) service offerings
- Optimizes L1/L2 switches for CO deployment to maximize service availability and facility utilization
- Utilizes a model that exploits the huge bandwidth availability in the fiber by splitting it up into multiple non-overlapping wavelength channels
- Addresses redundancy, scalability, and network topology changes that ensure QoS service metrics are within the customer service SLA.

The key for realizing such an ambitious initiative rests on resolving the following six critical issues:

1. How to replace the legacy layer-3 switching (routing) and hierarchal IP addressing scheme with layer-2 switching and flat (non-hierarchal) MAC/VLAN-based addressing scheme?
2. How to provide comprehensive end-to-end Operations, Administration, Maintenance, and Provisioning (OAM&P) in a unified Ethernet-Optical environment?
3. How to realize a converged Layer 2/1-network model in terms of control plane and management functionality?

4. How to totally eliminate the reliance on ST/RST/MST routing and redundancy functionality?
5. How to reliably transport native Ethernet frames that have no overhead capability to perform network OAM&P across the WAN?
6. How to devise a novel global layer-2 MAC and/or VLAN ID-address structure and space that is unique, hierarchal, and scalable with a source and destination addresses.

The main characteristics of the proposed Ethernet-over-WDM model are:

- Conventional Ethernet MAC frames and/or jumbo Ethernet frames must be transported end-to-end natively (translation into some other protocol is not allowed) from the access network through the metro and core networks to another access network.
- Only pure layer-2 switching at the packet/frame granularity is allowed throughout the entire network including Access, MAN, and WAN.
- Support for an IP/GMPLS-based unified control plane that offers a tighter integration between layer-1 and layer-2, leading to the collapse of the two layers into a single integrated layer managed and traffic engineered in a unified manner.
- A unified control plane that supports real-time provisioning and restoration of both full lambda and EVCs by running a single instance of an integrated routing and signaling protocols (use of ST, RST, and MST routing are totally eliminated).

1.4 Organization of the Thesis

This thesis is organized as follows:

Chapter 2 covers Ethernet background and protocols;

Chapter 3 covers overview of the current MEF Ethernet services and related work on L2 services and describes the key challenges and inhibitors to realizing the potential of Ethernet services;

Chapter 4 introduces the short-term solution that includes the ACES model and its simulation results; also, introduces the overlay model: Ethernet over WDM and BoD over optical transport network;

Chapter 5 covers Ethernet QoS and performance monitoring;

Chapter 6 covers the long-term vision and introduces a truly integrated L1/L2 hybrid GigE/OXC model.

Chapter 7 provides summary and conclusion.

2. Ethernet Background

2.1 Ethernet

Ethernet is the most widely used technology with approximately 90% of all data originating and terminating over Ethernet. It was developed by Xerox Corp in 1972 by Dr. Robert Metcalfe (see Figure 1). In May 22, 1973, Mr. Metcalfe wrote his famous memo on Ethernet Potential at Xerox PARC. Over several years, work continued leading to a publication in ACM entitled “Ethernet: Distributed Packet-Switching for Local Computer Networks”, and a patent (# 4063220) issued in December 1977. In 1980, Dr. Metcalfe helped form the consortium of DEC, Intel and Xerox to make Ethernet a standard, which led to the development and introduction of the Blue Book, the original Ethernet standard. However, in 1985, Ethernet earned a global recognition with the help of the Institute of Electrical and Electronic Engineers (IEEE), which introduced a series of standards for Local Area Networks called the IEEE 802 standards in 1983. These standards have found widespread acceptability and currently form the core of most LANs. In addition, the International Standards Organization (ISO) in an ISO 8802-3 series has adopted these IEEE standards. The ISO was created in 1947 to construct worldwide standards for a wide variety of engineering tasks. Adoption of ISO standards is required to allow manufacturers to produce equipment that is interoperable and provides guarantees to operate anywhere.

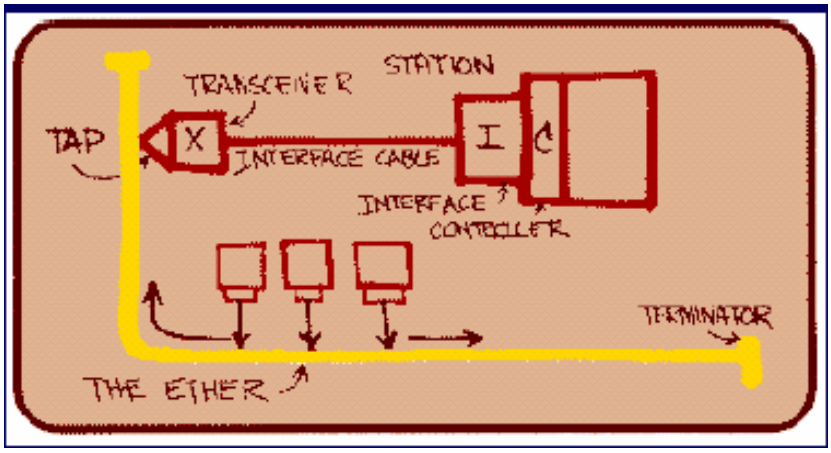

Figure 1: Robert Metcalf Ethernet presentation at the NCC Conference in 1976

2.2 Types of Ethernet LANs

The Original Ethernet network used a shared (passive) bus operated at 10 Mbps. Also, it used a simple access method to prevent two computers trying to transmit at the same time [37]. Each computer connected to the network will listen (if the line is clear) before sending any traffic. Also CSMA/CD ensures that both computers can retransmit any frames, which gets corrupted by simultaneous transmission. There-transmission follows the exponential back off algorithm. Figure 2, shows computer *A* sends a message to computer *E*. This message propagates along the coaxial cable and all of the nodes attached to the cable will see the message.

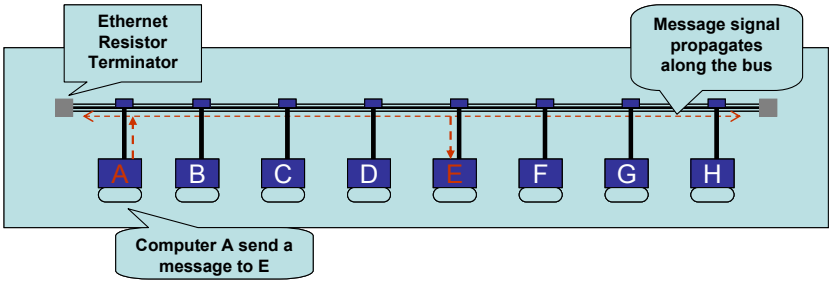

Figure 2: Ethernet Bus topology

2.3 Legacy Ethernet Network Topologies

An Ethernet network topology defines how its nodes and devices connect to it, through network cables and Network Interface Cards (NICs). Also, the type of network topology is determined by the geometric arrangement of its nodes and cables. There are several Ethernet Network topologies, such as bus, tree, ring, etc.

2.3.1 BUS

An Ethernet bus is a cable that forms a segment in which all devices attach to it, thus forming a bus or a backbone. The bus is formed from a 50-Ohm coaxial cable allowing computers in the LAN to attach to it via transceivers and network interface cards. One or more pieces of coaxial cable are joined end to end to create the bus, known as an "Ethernet Cable Segment". Each segment is terminated at both ends, as shown in Figure 3, by 50-Ohm resistors (to prevent reflections from the discontinuity at the end of the cable) and is also normally earthed at one end (for electrical safety).

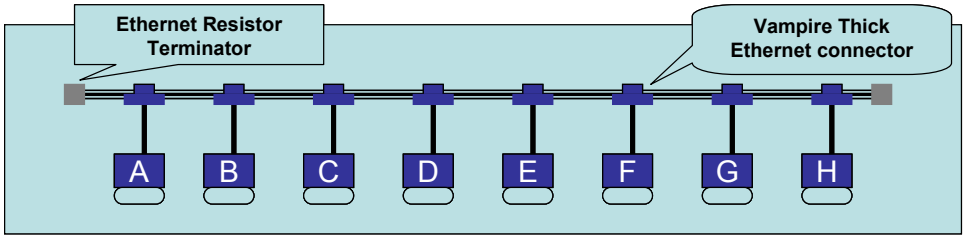

Figure 3: Ethernet Bus (Single Segment over a thick Ethernet cable)

2.3.2 STAR

A star network is a topology where two or more computers or devices are connected together into a central device such as a hub or a switch (see Figure 4). The star topology has one central connection point for other devices to attach and communicate with other devices.

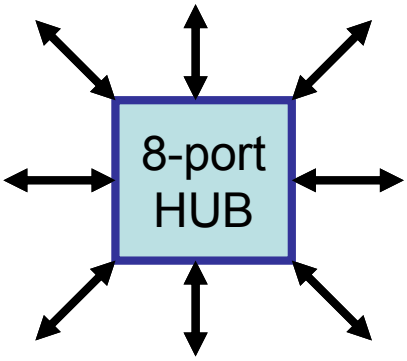

Figure 4: Star topology using a Hub or a Switch

2.3.3 TREE

A tree topology is a hybrid or a combination of multiple Star networks over a shared bus topology (see Figure 5).

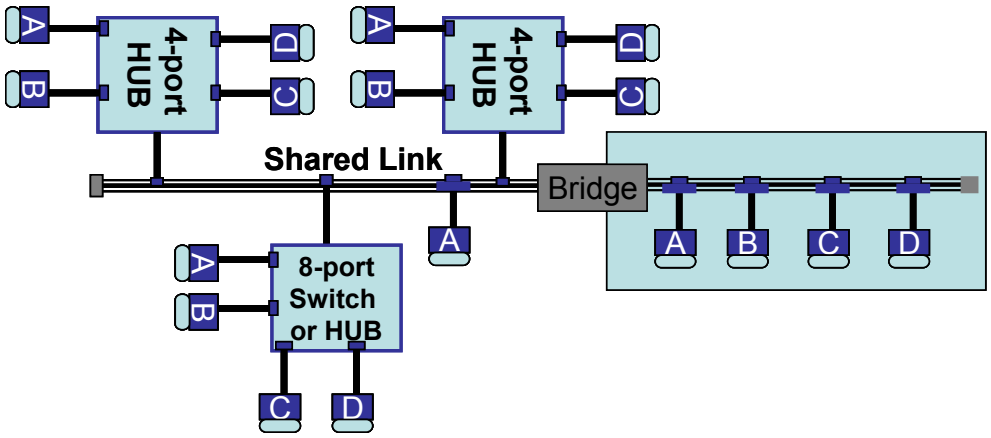

Figure 5: Ethernet Tree Topology

2.3.4 Ring

A ring is a network that allows its nodes to connect to one another in a closed loop forming a ring. Nodes connected along the ring may communicate with each other in the same direction as “clockwise” or counter clock-wise. Each node has exactly two adjacent nodes, on either side of it (see Figure 6).

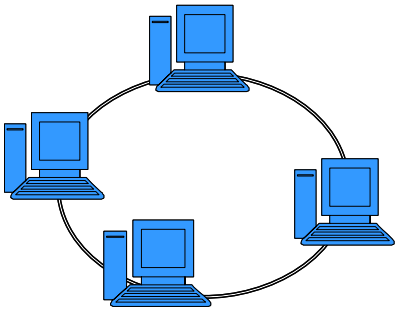

Figure 6: Ring Topology

2.4 Ethernet Transmission Media

Ethernet can be transmitted over a wide range of media types and physical layer segments. In some instances, a media converter is required. Ethernet's flexibility and support for wide range of media types contributed to its success. Ethernet transmission media plays an important part in the design and installation of an Ethernet network. Amongst the Ethernet media types include: copper twisted pair (UTP), coaxial cable (thick and thin), fiber optics and wireless mainly for full duplex. In Figure 7, the Media Access types for Gigabit Ethernet (GigE) are shown to include shielded and unshielded copper twisted pair, single and multimode fiber. For more details on these types and others, please refer to IEEE-802.3-2002 [68].

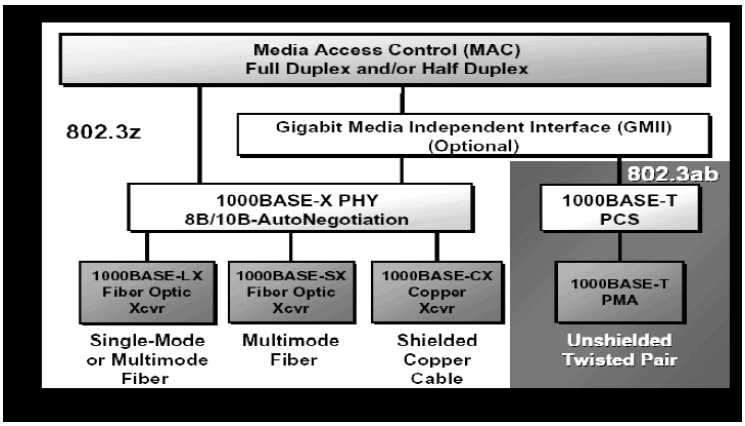

Figure 7: GigE Media Access Types

2.5 Common Ethernet Devices

2.5.1 Hub

A hub is a central device used to connect together two or more Ethernet segments in a star topology, and thus leading to a larger network topology. A hub is a layer-1 relay. This means that a hub takes an incoming signal, of any media type, and repeats it on all of its ports even if the message is intended to only one destination. Thus, a hub functions as a repeater by amplifying an incoming signal to extend its reach. Also, a hub functions as a single collision domain or a single network segment.

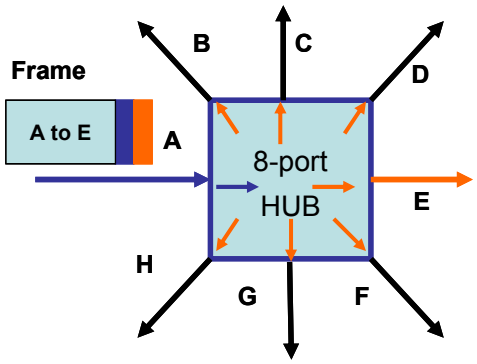

Figure 8: A Hub forwards a frame destined to port E on all of its ports

2.5.2 Repeater

A repeater is a device that allows multiple cables to be joined together to extend the network reach for a greater distance. The repeater accomplishes this by amplifying and retiming the signals. A repeater is a Hub with two ports (in-out)

Figure 9: Repeater

2.5.3 Bridge

A bridge is a device that connects separate networks together. Usually, it is a "Store and Forward" device that allows two LAN segments of different types, such as thin or thick Ethernet segments, to join together by forwarding Ethernet frames between each other. A bridge is a layer-2 relay device that is self-learning, which builds a forwarding table based on user MAC addresses on a segment [40]. Also, a bridge divides a network into separate collision domains while retaining the broadcast domain [90]. An intelligent bridge is able to filter traffic passing between two LAN segments and enforce policy separating the two LAN segments into separate collision domains. In addition, a bridge distributes spanning tree algorithm to help prevent loops, (see Figure 10).

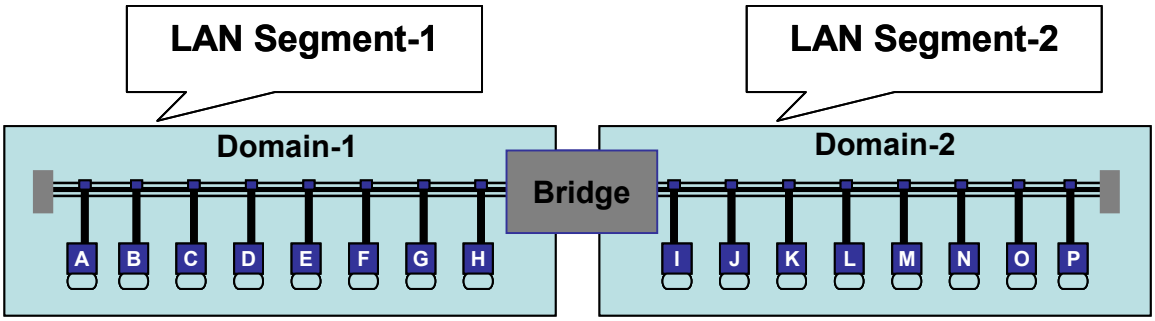

Figure 10: An intelligent bridge connecting two LAN segments

2.5.4 Switch

A Switch is a bridge device (layer-2 relay) with two or more interface ports that allows the bridging of several network segments together. A switch also allows each computer to be connected to it directly via a single dedicated port as one segment and thus eliminate segment collision. Hence, multiple computers can transmit and receive simultaneously via store and forward or cut through architecture. Store and forward examines the entire

packet before forwarding it to its destination, whereas, cut through examines only the destination address before forwarding the packet to its destination segment.

A Switch examines each packet and processes it accordingly rather than simply repeat it on all of its ports like a hub, (see Figure 11).

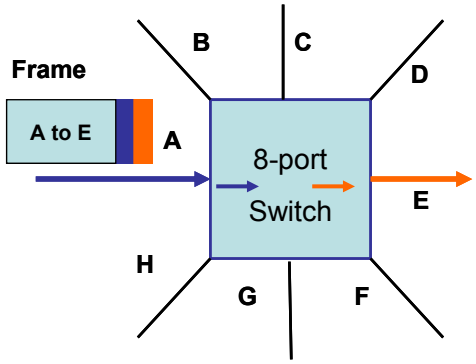

Figure 11: An Ethernet Switch forwarding a packet from A to E

2.6 Ethernet Frame

Figure 12 illustrates the Ethernet frame format (IEEE 802.3 MAC frame) [68], which is used to communicate between computers and has the following format structure: The preamble; Start Frame Delimiter (SFD); the addresses of the frame's source (SA) and destination (DA) addresses; a length or type field to indicate the length or protocol type of frame; the data field of n bytes ranging from 46 to 1500 bytes + Padding; and a 4-byte cyclic redundancy check known as Frame Check Sequence (FCS).

1. 7 bytes of Preamble Field;
2. 1 byte Start Frame Delimiter (SFD) used for synchronization;
3. 6 bytes of destination address: 48 bits Ethernet destination address (Destination Address);

4. 6 bytes of source address: 48 bits Ethernet source address (Source Address);
5. 2 bytes of message/protocol encapsulated type of data being sent (Length/Type);
6. 46 to 1500 bytes of data: up to 1500 bytes of packet data;
7. 4 bytes of cyclic redundancy check (CRC) used to check for error detection in the transmission known as Frame Check Sequence (FCS).

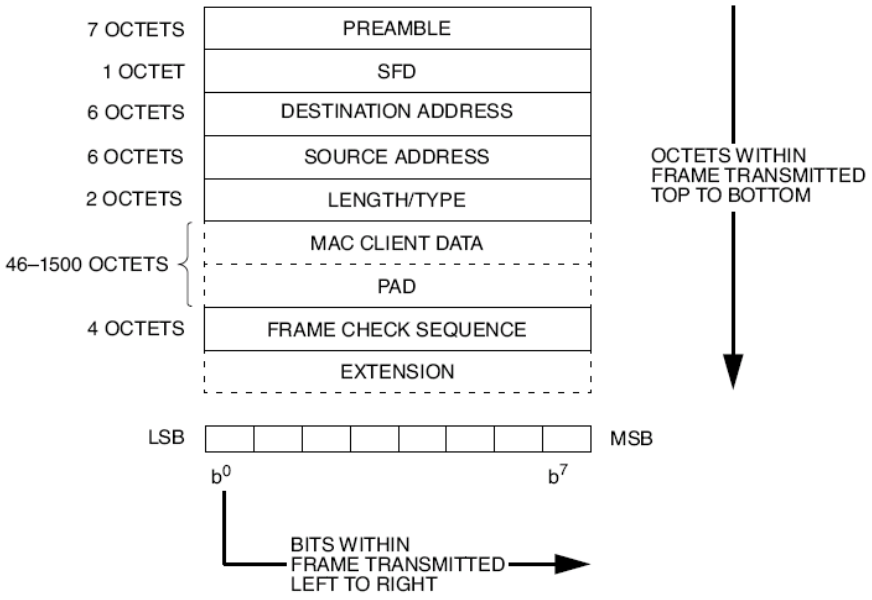

Figure 12: 802.3 MAC Frame Format

The packet format above defines the syntax and semantics of the various components of the MAC frame. In addition, an 802.3 frame with Q tag is known as .1Q Frame format and has the format shown in Figure 13 [68].

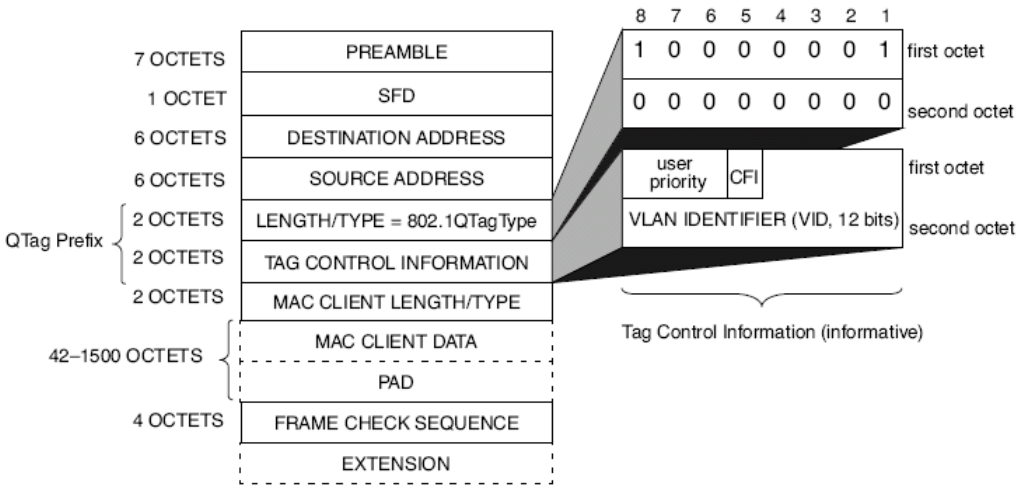

Figure 13: .1Q MAC Frame format

2.6.1 802.1*Q* MAC frame format

An IEEE 802.1Q frame format, (see Figure 13) is based on the IEEE 802.3 frame with the insertion of:

A 4-byte Q Tag Prefix between the end of the Source Address and the MAC Client Length/Type field of the MAC frame. The Q Tag Prefix comprises of two fields:

A 2-byte constant Length/Type field value consistent with the Type interpretation and equal to the value of the 802.1Q Tag Protocol Type (802.1 Q Tag Type).

A 2-byte field containing Tag Control Information.

Following the Q Tag Prefix is the MAC Client Length/Type field, MAC Client Data, Pad (if necessary), FCS, and Extension (if necessary) fields of the MAC frame.

The length of the frame is extended by 4 octets by the Q Tag Prefix.

The Q Tag control information field contains the following information:

1. 3 bits User Priority Field
2. A Canonical Format Identifier (CFI)
3. 12 bits VLAN Identifier

2.6.2 Ethernet Inter Frame Gap

After transmission of each frame, a transmitter must wait for a period of 9.6 microseconds (at 10 Mbps) to allow the signal to propagate through the receiver electronics at the destination. This period of time is known as the Inter-Frame Gap (IFG). While every transmitter must wait for this time between sending frames, receivers do not necessarily see a "silent" period of 9.6 microseconds. The way in which repeaters operate is such that they may reduce the IFG between the frames that they regenerate [52].

2.6.3 Ethernet Inter Frame Space (IFS)

IEEE 802.3 Ethernet frame size varies in size between 64 and 1518 bytes. This size increases as we add a VLAN tag (4 bytes); hence the frame size becomes 1522 bytes. These frame sizes are standard, but larger frames are available and are supported on switches and routers. The larger frames are referred to as Ethernet jumbo frames and range in sizes that exceed 9000 bytes.

Inter Frame Size (IFS) is a layer 2 overhead between adjacent Ethernet frames, which carries no useful information in a full duplex link [37]. IFS is defined as the sum of the inter-packet gap (at least 12 bytes), preamble (7 bytes) and start of frame delimiter (1 byte). Hence, IFS size is at least 20 bytes. IFS plays an important role in determining the actual throughput carried on a link. As Ethernet is transported over transport networks,

such as SONET, it is important to calculate the throughput in order to ensure that the Ethernet data rate is being transported correctly. For example, the dominant Ethernet port speeds today are 100 Mbps and 1 Gbps and they refer to the medium access control (MAC) sub-layer signaling rates. However, due to the physical layer block encoding of 4B/5B for 100Base-TX, or 8B/10B encoding for 1000Base-X, the resulting physical layer speeds are in reality 125 Mbps and 1.25 Gbps respectively, which require 25% higher transmission rate. Therefore, it is best to transport Ethernet frames that maintain 802.1 and 802.3 essential content such as control frames, spanning tree BPDUs, pause or slow protocol frames, discovery frames, etc. directly into the SONET time division multiplexing (TDM) and frame mapped - generic frame procedure (GFP-F)

Hence, a service provider can offer SONET bandwidth options in basic TDM increments (VT1.5, STS-1, or STS-3c), and the customer can buy just the bandwidth that best meets their needs independent of the Ethernet port speed. The impact of a frame-mapped approach is that certain Ethernet mechanisms cannot be transported transparently over the SONET network. While the GFP standard allows for both frame mapping (GFP-F) and transparent mapping (GFP-T) for EoS, the GFP-T standard is continuing to undergo modifications to efficiently transport line rate and subrate Ethernet. The standard time-division multiplexed (TDM) rates over SONET transport are payload capacities of 1.536 Mb/s (DS1), 1.600 Mb/s (VT1.5), 44.210 Mb/s (DS3), 48.384 Mb/s (STS-1), 149.76 Mb/s (STS-3c), 599.04 Mb/s (STS-12c), 2396.16 Mb/s (STS-48c), and 9584.64 Mb/s (STS-192c) [32] [78].

As observed, Ethernet physical layer rates are different from the TDM rates. When attempting to carry Ethernet over SONET, one approach could be to map Ethernet into a

larger SONET payload capacity. However, such an approach would waste a large portion of the customer paid SONET transport bandwidth (e.g., nearly 35 percent when 100 Mb/s Ethernet is mapped into an STS-3c and nearly 60 percent when 1Gb/s Ethernet is mapped into an STS-48c). Virtual concatenation will alleviate this problem to a large extent where individual VT1.5, STS-1, or STS-3c time slots can be logically grouped together to obtain a SONET bandwidth.

2.7 Ethernet Physical Layer

Ethernet evolved from a shared 10-Mbps medium over thick coaxial cable, to a switched 10-Gbps per link over fiber, with 100 Mbps and 1 Gbps as incremental speeds. Ethernet is strictly based on a layered OSI model (see Figure 14). The Physical Layer (Transmit a bit stream across a physical transmission medium) and the Data Link Layer (Transmit data reliably from one node to another) are the two layers, which are mainly used in Ethernet and considered the key local area network layers. The IEEE Ethernet reference model is slightly different and uses a Physical Medium Attachment as the network interface. Hence, Ethernet physical layer, defines the type of cable connectors, data encoding scheme, the physical line rate or speed (whether it is half or full duplex), electrical signaling, symbols, line states, and clocking requirements.

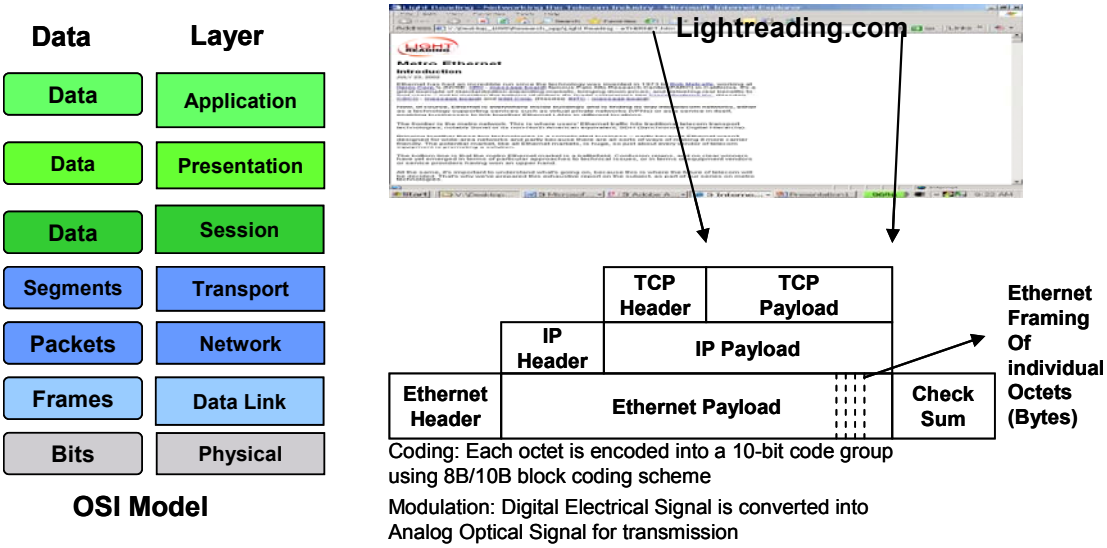

Figure 14: OSI Model

Ethernet physical interfaces that are standardized by the IEEE 802.3 are [53]:

2.7.1 10BASE-T

10BASE-T, as defined by IEEE 802.3-2002 [68], specifies Ethernet over unshielded twisted-pair. The IEEE 802.3i specification addresses the Medium Attachment Unit (MAU), the repeater, and the Twisted-Pair media for 10Mbps CSMA/CD LAN. 10 Base-T utilizes two pairs of unshielded twisted-pair (UTP) cable with one pair of wiring to transmit data, and the second pair of wiring to receive data. The connector type is RJ45 and the UTP cable used is of category 3 or better with a maximum segment length of 330 feet (100m). The signaling frequency is 20MHz. Links between repeaters are also limited to a maximum of 330 feet (100m). It is recommended that no more than of 4 repeaters and 5 segments are to be used; therefore 10BASE-T LAN can have a maximum diameter

of ~ 1650 feet (~500m). 10 BASE-T requires a star-wired configuration with a central hub.

2.7.2 100BASE-T/F/L FAST ETHERNET

100BASE-T, as defined by IEEE 802.3-2002 [68], specifies Ethernet over a variety of cable types. 100BASE-T is the 100Mbps version of the classic Ethernet standard. The IEEE 802.3u specification couples the IEEE 802.3 CSMA/CD MAC with a family of new physical layers for 100Mbps Ethernet, including:

- 100BASE-TX, which requires two pairs of Category 5 UTP or Type 1 shielded STP cabling with a maximum cable distance of up to 330 feet (100m)
- 100BASE-FX, which uses two strands of multimode fiber with a maximum cable distance of up to 1600 feet (412m)
- 100BASE-LX10, which uses two strands of single mode fiber with a maximum cable distance of up to 33,000 feet (100km)

Many optical vendors have developed new transceivers coupled with new GBICs[2] or SFP which extend the fiber reach to 100 kilometer (km).

All 100BASE-T PHY (physical) specifications require a star configuration with a central hub and a signaling frequency of 125MHz.

The maximum segment length for fiber-optics connections varies. For multimode fiber, it can be 412m while repeater diameters can be up to 320m. The ratio of packet duration to network propagation delay for 100BASE-T is the same as for 10BASE-T. This is due to the fact that in 100BASE-T the bit rate is faster, bit times are shorter, packet transmission

times are reduced, and cable delay budgets are smaller; all in proportion to the change in the bandwidth.

2.7.3 1000BASE-S/L/C/T GIGABIT ETHERNET

1000BASE-X, as defined by IEEE 802.3-2002 [68], specifies Ethernet over a variety of cable types. 1000BASE-X refers to the 1000Mbps MAC and the LX/SX/CX transceiver technology. As shown in Figure 15 [94], the IEEE 802.3z specification defines a family of 1000 Mbps PHY physical layer entities such as:

- 1000BASE-SX to support multi-mode fiber-optics cabling with a maximum cable distance of up to 1800 feet (550m)
- 1000BASE-LX to support single mode fiber-optics cabling with a maximum cable distance of up to 16500 feet (5000m)
- 1000BASE-CX to support shielded copper cabling. This type of cabling is typically not used and has been replaced by 1000BASE-TX
- 1000BASE-TX PHY to support four pairs of Category 5 balanced copper cabling with a maximum cable distance of up to 330 feet (100m)

1000BASE-SX is cost effective and is targeted at shorter backbone or horizontal connections. It uses the same physical layer as LX and uses affordable 850nm short-wavelength optical diodes. It uses only multimode fiber.

---

[2] GBIC: GigaBit Interface Converter

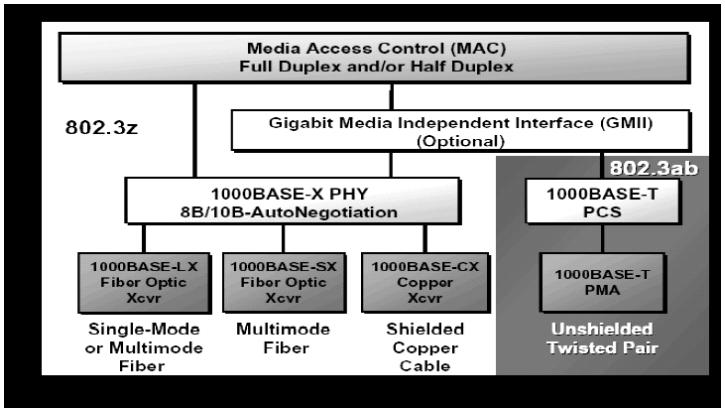

Figure 15: Gigabit Ethernet Layer Diagram

1000BASE-LX is targeted at longer backbone and vertical connections. It requires a single mode or multi-mode fiber using 1300nm lasers.

Gigabit Ethernet extends the ISO/IEC 8802-3 MAC beyond 100 Mb/s to 1000 Mb/s. The bit rate is faster, and the bit times are shorter, both in proportion to the change in bandwidth. In full duplex mode, the minimum packet transmission time has been reduced by a factor of ten.

2.7.4 10GBASE-XX/10 GIGABIT ETHERNET

10 GigE interface is based on the IEEE 802.3ae standard [69] [96], which specifies four physical medium dependent sub-layer (PMD) interfaces that operate at various distances over single or multimode fiber, as shown in Figure 16 [94]. The standards include two new 10 Gigabit physical layer specifications (PHY) that support LAN (LAN PHY) and WAN (WAN PHY) services. The LAN PHY has two interfaces (10 GBASE-X and 10 GBASE-R) and WAN PHY has one interface 10 GBASE-W.

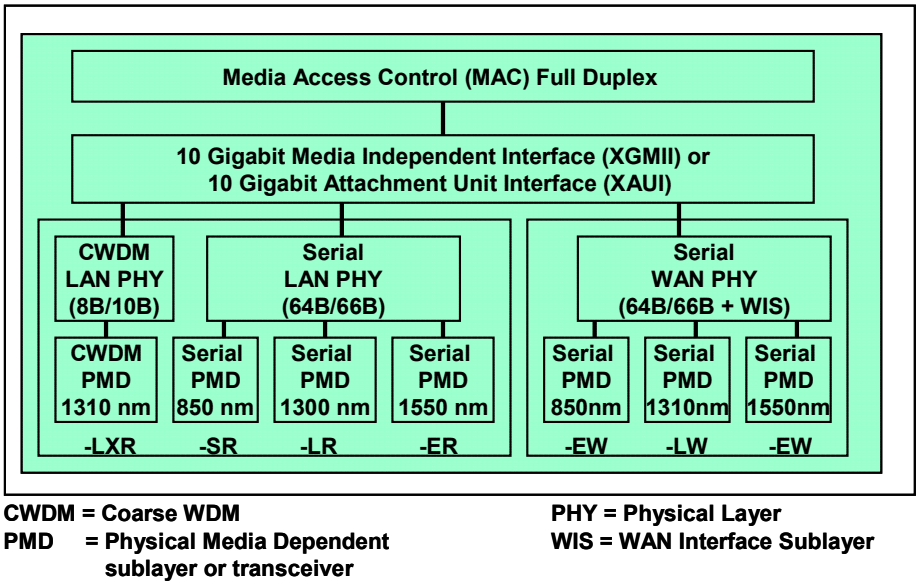

Figure 16: 10 Gigabit Ethernet Layer Diagram

Two PMDs are defined to support single mode fiber:

- 10 GBASE-L (operates at1300nm with distance up to 10 Km)
- 10 GBASE-E (operates at1550nm with distance up to 40 km).

One PMD is defined to support multi-mode fiber:

- 10 GBASE-S (operates at 850nm with distance up to 82 metes).

WAN PHY (9.953 Gbps) differs from the LAN PHY data rate (10.3125 Gbps) in order to ensure its encoded payload fit into a SONET concatenated STS192c or SONET OC192. The encoding on the LAN side is 66B/64B while on the WAN side it is 64B/66B [53]. Therefore, the effective line rates are:

- Effective line rate on the LAN side is = 66/64 * 10000 = 10, 3125 Gbps
- Effective line rate on the WAN side is = 64/66 * 10000 = 9.953 Gbps

Achievable topologies for 10Gbps operation are comparable to those found in 1000BASE-X full duplex mode and equivalent to those found in WAN applications. In

legacy Ethernet, IEEE 802.3 standard for Ethernet operating at 10 Mbps specifies a maximum bus length of 2.5km with repeaters. This translates to approximately 50 microseconds or 500 bits or 512 bits with safety margin as the minimum packet size (64 bytes). The same is true for fast Ethernet operating at 100 Mbps; the standard specifies a maximum bus length of 200 meters. This means that the minimum packet size is 64 bytes. Hence, the collision probability and throughput in 10M and 100M are similar but it is expected that frames in 100M be longer by a factor of 10. In order to maintain the interconnected networks compatible, the maximum frame size of 1522 bytes in both 10M and 100M standards are kept the same. This Ethernet legacy-based limitation is overcome by new switching capability that allows one station or a single hub connected directly into a port on a switch. Also, the use of fiber cables with repeaters has taken Ethernet beyond the LAN into the MAN and the WAN.

2.8 Ethernet Protocols

Advancements in Ethernet Technology, such as high bandwidth and long reach optics have enabled a whole new generation of Ethernet networking pushing Ethernet into the MAN and WAN networks. These innovations have offered a new opportunity for Service providers (SPs) to offer native Ethernet services over metropolitan areas (network diameters of > 100 Km). Further technology innovations are allowing for Ethernet services to be transported over a variety of transport technologies, including SONET/SDH, WDM and Multiple Protocol Label Switching (MPLS) Pseudo Wire (PSW), which effectively extend the reach nationally and globally [32].

Many ad hoc committees are working on solutions and variations around IEEE 802.3 standards as well as ITU-T such as X.86 and GFP. The need to maintain standards

compatibility (backward and across) is important. Also, we need vendor interoperability to conform to Ethernet characteristics such as jitter, delay, quality of service, flow control in burst mode, and remote monitoring to make the Ethernet service ubiquitous amongst service providers using different vendor equipment. To accomplish this, Metro Ethernet Forum (MEF) is actively defining Metro Ethernet Services. Hence, improved Wide Area service parameters utilizing standardized technologies with flexible and diversified options are the key to the future expansion of Ethernet in the WAN and globally. Similarly, this architecture research work conforms to the current Ethernet standards as much as possible, which makes it more valuable and practical.

At this time, both the IEEE and IETF standards are working on defining protocols and standards for use in metro Ethernet networks. Presently, there is not yet a solution for the essential problem facing Service providers, which is how to guarantee service performance while maintaining the low cost structure of a 'native Ethernet' network. Although, Ethernet has evolved drastically and has changed its reach characteristics, i.e. it is no longer distance limited, or a pure CSMA/CD based technology. It still lacks the followings:

- Support for the much needed QoS/CoS capability
- Support for hop count
- Support for maintaining a loop free network without the limitations of spanning tree protocol

2.8.1 Spanning Tree Protocol (STP)

Spanning Tree Protocol is a bridge-to-bridge protocol that maintains a loop free network. Loops create never-ending data paths, resulting in excessive system overhead [70]. Hence, STP is an essential element of enterprise and carrier infrastructure and is defined in the IEEE 802.1D standard.

STP provides a fault tolerant capability by allowing an alternate path to be selected should a network failure occurs. However, spanning tree protocol has several limitations that include:

1. It is a LAN based protocol where SLA and QoS concepts are not available.
2. It is a best effort algorithm with limited prioritization queues - No support for QoS/CoS.
3. It supports low link usage (approx. 55 to 65 % in a large full-mesh network).
4. It is slow to converge (30 to 50 seconds).
5. Its slow restoration time does not meet voice grade service.
6. It determines link cost based on speed and not resource utilization.
7. It does not scale efficiently to span larger networks.

For a given network, Spanning Tree is build by running the Spanning Tree Algorithm, which consists of four key steps [95]:

- Election of the root bridge
- Election of root ports
- Election of designated ports
- Changing of the port states

2.8.1.1 Election of the root bridge

As the network is turned up, the Spanning Tree algorithm runs through these four steps and initially selects a bridge to act as the root bridge for the tree buildup. It makes this selection based on the lowest assigned Bridge ID[3] [20]. If more than one bridge exists with the same BID, then the lowest bridge id is determined by the combination of Bridge ID + MAC address. The root bridge is selected via a series of exchange messages called Bridge Protocol Data Unit (BPDU) between the network devices. These are special frames that are broadcasted by all switches and/or bridges in the network. The BPDU are transmitted every two seconds by each port on a bridge or switch and contain information relevant to building and maintaining the spanning tree topology.

2.8.1.2 Election of root and designated ports

Once the root bridge is elected, all other bridges in the network must determine which ports are the root ports and which ports are designated ports.

Root Port

A root port is the port with the least cost to the root bridge.

Designated Port

A designated port is a port on a bridge that is connected to a segment on the network, which has the least cost to the root bridge

If a port is neither a root-port nor a designated-port, then it is considered a redundant port.

2.8.2 Port States

There are five port states, which are:

[3] BID, is an 8-byte,field, which its value is normally assigned by a network administrator

- Block: A port in the *block* state does not forward any traffic except some control frames
- Disable: A port in the *disable* state is administratively disabled and no traffic will pass through it. The disable port does not participate in the spanning tree negotiation.
- Forward: A port in the *forward* state is able to forward (send) and receive data
- Learn: A port in the *learn* state is able to receive frames and learn MAC addresses on a particular port and begin to build its "forwarding table". No user data pass through this period.
- Listen: A port in the *listen* state receives and forwards BPDUs related to the spanning tree topology. During this state, root port and designated ports are identified for each bridge.

2.9 Spanning Tree Topology

STP, as defined in IEEE 802.1D-2004, allows bridges to listen to each other and determine only one root-bridge in the network. The root bridge allows all of its ports to be in the forwarding state with the ability to send and receive data. All other bridges are considered as non-root bridges, with the ability to designate only a root port, which has the lowest cost path to the root-bridge. The root port is in the forwarding state. On each segment in a network topology, there is one designated port, which has the lowest cost to the root bridge. Designated ports are normally in the forwarding state. STP utilizes lowest path cost to the root bridge based on the accumulated cost of the link path and builds a tree-based network that is loop-free. A path cost is determined by C(p) as the sum of costs

of links traversed in the path, where n is the number of links in a path and C_l is the cost pre link:

$$C(p) = \sum_{l=1}^{n} C_l$$

The Spanning Tree Topology is formed using the following steps [95]:

1. Identify the root bridge, once the root bridge is identified, then
2. Identify the root and designated ports,
3. All remaining ports are put into a blocking state.
4. Root and designated ports are put into the forward state
5. Now, the spanning tree topology is formed and Data traffic begins to flow, until a topology change occurs
6. A topology change will trigger the STP recalculation and will run the STA algorithm again to settle on a new topology

Figure 17 shows a three-node network with equal link speeds and its resulting spanning tree. In this example, node A is the root bridge and link 3 is blocked or unused to avoid forming a loop [20]. STP continually explores the network for a link state change, (i.e. addition, failure, or loop). When STP detects a change in a link state, it places the blocked ports to the forwarding states. However, the time for the network to converge is an issue. The amount of time required for the convergence of network topology (all switch ports have transitioned to their blocking or forwarding state) exceeds the carrier grade requirement. With STP steady state, all traffic takes the same path. This is a major limitation that faces a service provider environment.

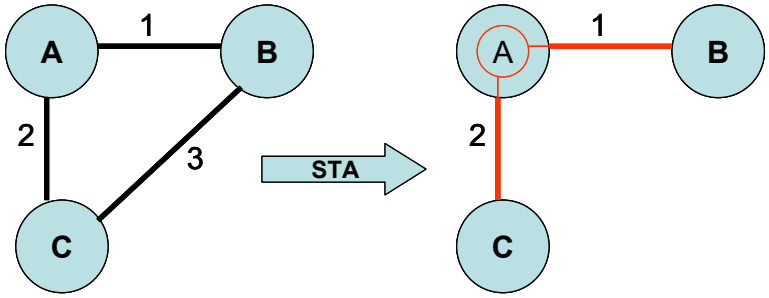

Figure 17: A 3 nodes network and its resulting spanning tree

2.9.1 BPDUs

Bridge Protocol Data Units (BPDUs) are special multicast frames that are broadcast out by all of the bridges and switches on the network [54], [65]. These frames contain all of the vital topology information about each sender that is needed by the bridges to build and maintain the Spanning Tree. They are the lifeblood of the protocol.

Each port on a bridge will transmit a BPDU out of each port every two seconds by default (Hello Time[4]) [73], [74], [95]. Each bridge will examine all of the BPDUs it receives, and sends, on each port to find the best (lowest value) BPDU. It will then save a copy of that BPDU for the port. If the saved BPDU originated from another bridge, the receiving bridge will then stop transmitting BPDUs out of that port. If the receiving bridge stops receiving the "best" BPDU (perhaps due to a link failure), the saved copy will expire in 20 seconds by default (Max Age[5]), and the bridge will begin sending BPDUs out of that port once again (topology change).

---

[4] Hello Time is one of the three distinct timers that STP uses. This timer defines the time interval between transmissions of BPDUs containing configuration frames.

[5] Max Age Time is one of the three distinct timers that STP uses. This timer defines the maximum amount of time a BPDU will be saved for a port.

There are two types of BPDUs:

1. Configuration BPDUs

   The majority of the BPDUs on a healthy network will be of this variety, and are exchanged by all bridges on the network, flooding the network with the topology information needed by STP.

2. Topology Change Notification (TCN) BPDUs

   When a topology change occurs—for example, an active link goes down—this type of BPDU alerts the root that the active topology has changed, triggering a recalculation of the Spanning Tree algorithm and reconvergence. This type of BPDU will originate from the bridge that experiences a change in the state of one of its ports.

2.9.2 A BPDU configuration frame

BPDUs exist only at layers 1(physical) and 2 (data link) of the OSI model. Notice that there is no routing information contained in the BPDU frame [95].

The following fields are contained in a configuration BPDU:

1. Protocol ID: Always contains the value of 0.
2. Version: Always contains the value of 0.
3. Type: Determines the type of BPDU, configuration, or TCN (Topology Change Notification).
4. Flags: Used in conjunction with the TCN BPDU type. Indicates either a topology change or a topology change acknowledgment.

5. Root BID: Contains the bridge identifier of the bridge that has been designated as the root bridge.
6. Root Path Cost: Defines the cumulative cost from the bridge that originated the BPDU, across all links to the root bridge.
7. Sender BID: Contains the bridge identifier of the bridge that generated the BPDU.
8. Port ID: Identifies which port the BPDU left the transmitting bridge on.
9. Message Age: The amount of time since the root bridge advertised a BPDU based on the current topology information.
10. Max Age: Defines the maximum time that a BPDU will be saved for a port.
11. Hello Time: The amount of time between BPDU broadcasts.
12. Forward Delay: The amount of time that a port will spend in the listening and learning STP port state.

2.9.3 Link Cost

The IEEE assigned a cost value per Ethernet link speed as shown in Table 1. These values are default, however, a network administrator can modify them as needed [68] [95].

Table 1: IEEE STP Cost Values Related to Bandwidth

Bandwidth (Link Speed) STP Cost	*Cost*
10 Mbps	100
100 Mbps	19
1000 Mbps (1 Gbps)	4
10000 Mbps (10 Gbps)	2
> 10 Gbps	1

2.9.4 STP Timers

STP uses three distinct timers:

- Hello Time: This is the time interval between transmissions of BPDUs containing configuration frames. The value default is 2 seconds.
- Forward Delay Time: This timer defines the amount of time that a port will spend in either listening or learning states. The default setting is 15 seconds
- Max Age Time: This defines the maximum amount of time a BPDU will be saved for a port. The default time is 20 seconds. Max Age timer provides STP the ability to detect a link failure.

2.10 Rapid Spanning Tree Protocol (RSTP)

Rapid Spanning Tree Protocol, as specified in IEEE 802.1D-2004, is a distributed algorithm that selects a single bridge/switch to act as the spanning tree's root. The algorithm assigns port-roles to individual ports on each bridge/switch. A port role is determines whether a port is a root-port or designated-port. The role of the root-port is to become a part of the active topology connecting its own-bridge to the root bridge. However, the role of the designated-port is to connect a LAN segment through its own-bridge to the root bridge. RSTP ensures rapid recovery of connectivity following the failure of a bridge/switch, bridge port or LAN.

The key difference between STP and RSTP is the time it takes to converge. Once a link is lost or the topology has changed, STP requires 30 to 60 seconds to detect the changes and reconfigure, which affects network performance. RSTP improves the slow Spanning Tree

recovery (tens of seconds interval)) algorithm typically implemented in Ethernet LANs significantly[6] (sub second interval). This is critical to Ethernet service infrastructures supporting classes of service and multimedia applications that are sensitive to session timeouts and data loss.

2.11 Multiple Spanning Trees (MSTs)

Multiple Spanning Trees (MSTs) , as specified in IEEE 802.1Q-2003, are logical topologies of multiple spanning tree protocol (STP) that reduce the limitation of STP by providing the ability to efficiently utilize the unused network links. Unlike the spanning tree switching, multiple spanning tree switching provides the ability to utilize the unused network trunk links and allows a service provider to overlay their physical topology with a logical topology based on the multiple spanning trees and to assign Ethernet Virtual Circuits per MST as needed. Using our previous example, shown in Figure 17, a second spanning tree needs to be created to utilize the unused trunk link between nodes B and C. The second root bridge can be either B or C and not A due to the lowest cost between B and C. Hence, our second spanning tree shown in dotted lines, in Figure 18, is based on node C as the root bridge and utilizes trunk links 2 and 3.

[6] When properly implemented, RSTP reduces the time it takes to reconfigure and restore service on link failures and restorations to sub-second levels, while retaining compatibility with equipment based on STP.

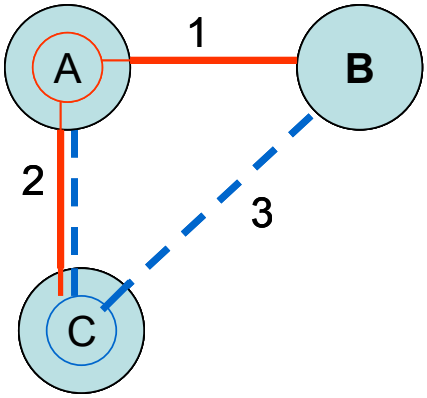

Figure 18: Multiple Spanning Trees

2.12 Provider Bridge

IEEE 802.1ad, “Provider Bridges” working group is defining a hierarchical structure for Ethernet bridges that would allow SPs to use “provider bridges” to carry Enterprise customer bridged traffic. In such a model, the control plane (Spanning Tree) for the SP’s switched Ethernet network is de-coupled from the customer’s control plane. This activity is expected to resolve some of the problems facing SPs that use native Ethernet switching networks for delivering Ethernet services.

This standard builds on the IEEE’s 802.1Q (Virtual LANs) to enable stacked VLANs, commonly referred to as “Q-in-Q” tag stacking, a vital tool for service providers to scale beyond the limitation of 4096 VLANs addresses. With stacked VLANs, customer’s traffic is tagged with a service provider VLAN tag at the edge of the network. This allows a significant scalability of provider infrastructure, and separates subscriber spanning tree restoration from the service provider’s protection methods. In addition, 802.1ad will allow Layer 2 control protocol tunneling. 802.1ad protocol will become a component of virtual private LAN service (VPLS).

2.13 Layer 2 Control Protocols (L2CP)

The Layer 2 control protocol frames are needed for specific applications [32]. The following lists a few standardized L2CPs:

1. STP/RSTP/MSTP
2. Link aggregation control protocol (LACP), which allows for the dynamic bundling of multiple Ethernet interfaces; LACP provides load sharing and protection, amongst links in the bundle, for critical data or headquarter centers
3. IEEE 802.3x MAC control frames (Xon/Xoff flow control that lets an Ethernet interface to send a PAUSE frame in case of traffic congestion on the egress port).
4. Others, such as: GARP, All Bridges, 802.1X Authentication, LAMP, 802.3AH Link OAM, and future standards such as 802.1ag: Connectivity Fault Management, ELMI: Ethernet Local Management

BPDUs are not layer-2 control protocol, but are layer 2 processing frames used in establishing the spanning tree.

3. Ethernet Services

3.1 Overview

Ethernet is the dominant LAN technology, with the majority of all data traffic terminating on an Ethernet port. Today, there are over 6 billion Ethernet ports and the demand for ever-faster networking speed is on the rise, and Ethernet has gone through significant changes to meet this demand. Although Ethernet started with a half-duplex shared media at a speed of 10Mbps in the Local Area Network (LAN), it has matured to a full-duplex switched 10/100/1000/10000 Mbps in the LAN and the MAN Networks. The IEEE 802.3ae 10GigE Standard has been in place for over three years and it is expected that within the next 5 years, 100GigE link speeds will be used in the core backbone.

Carrier Ethernet is a fairly new concept. It started in 2001, when a group of companies decided to push Ethernet outside the Enterprise and as a result, Ethernet first mile (EFM) (last mile) was formed.

Advancements in Ethernet Technology, such as high bandwidth and long reach optics have enabled a whole new generation of Ethernet MAN networks. SPs are tapping into the low cost Ethernet technology and are developing Ethernet services to attract Enterprise customers. Among some of the key service features that are currently under development include Class of Service (CoS) and guaranteed SLAs. The Next Generation Ethernet Switched Networks should provide the ability to bring to the end-user an opportunity to access high bandwidth speeds with great flexibility. This includes residential users with services based on Ethernet and Gigabit passive optical network (EPON and GPON) technologies. This opportunity depends on an end-to-end Ethernet

network infrastructure and the evolution to carry voice, video and data with a predictable quality of service. The short-term solution presented in this chapter enables the service providers to introduce advanced Ethernet services with the help of a centralized admission control system that maintains and tracks the bandwidth utilization of the physical links in the logical topologies. Given the flexibility of Ethernet and its ability to offer seamless interoperability and adaptation to changing environment, these advanced services can be transported over existing legacy transport networks such as SONET. In order to evolve Ethernet and to converge it with other layers, it is important to understand the characteristics of the layers above and below Ethernet.

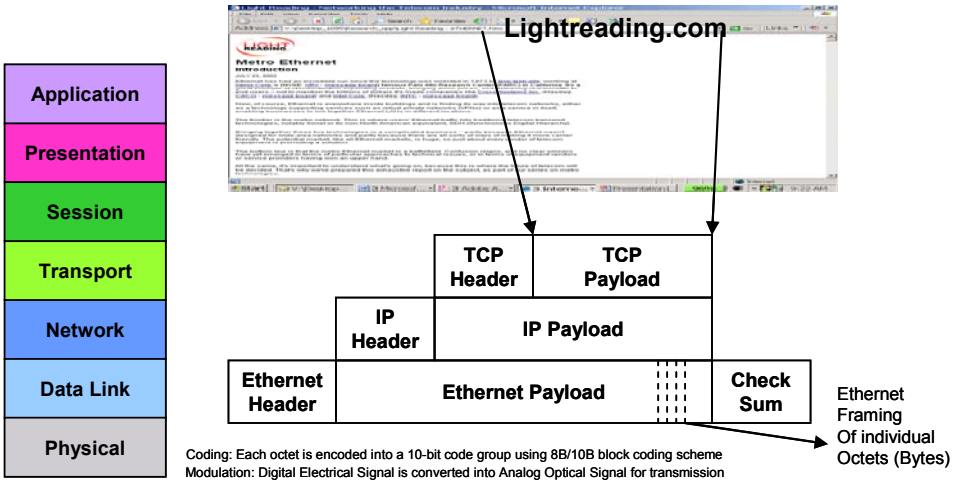

Figure 19 OSI model: Data transported over a GigE

Networks are based on a series of layers such as the OSI layers. As shown in Figure 19, an IP Packet fits nicely into the payload of an Ethernet frame. In addition, Ethernet and IP also share a set of characteristics such as connectionless based mode, packet based technology, and statistical multiplexing for sharing network resources. In addition, the services that IP provides over the Internet are the same as those offered by Ethernet in a

LAN network. Ethernet can provide the means for data convergence and potentially reduce network protocols interoperability issues made by other transport technologies.

Ethernet, as a long-term solution, can leverage the optical transport network control and management planes in the core network to support the emergence of bandwidth on demand services and enable it to overcome some of its critical limitations in becoming a carrier class network.

The Next Generation Ethernet Switched Networks should provide the ability to bring to the end-user an opportunity to access high bandwidth speeds with great flexibility. This opportunity is dependent on an end-to-end Ethernet network infrastructure and its evolution to carry voice, video and data with a predictable quality of service. However, Ethernet technology, in its current form, is not truly ready to support next generation Ethernet Switched services. Specifically, using “Pure” or “Native Ethernet” or VLAN bridged networks, as a transport medium in the metro and WAN still faces several key hurdles, including:

1. End-to-end QoS guarantees: currently Ethernet provides best effort traffic delivery. Although IEEE 802.1Q specifies three priority bits, Ethernet has no true class of service provision, such as DiffServ, although it can mark packets for prioritization, but it cannot schedule or police them. In addition, Ethernet QoS is not mature in a multi-customer environment as in a service provider environment. Ethernet’s lack of inherent QoS capabilities raises the critical issue of whether a given request with a specific QoS requirement can be admitted without compromising the service performance of already accepted Ethernet Virtual Circuits, which have been provisioned.

2. Scalability: Ethernet's major scalability bottlenecks include:
    a. Spanning Tree: The use of a spanning tree Protocol (STP), an IEEE 802.1D standard, to route traffic and the lack to load balance. A single tree allows only one loop-free path, which can result in uneven load distribution and potential bottlenecks. Ethernet's path establishment via STP results in non-optimal routing paths, which could introduce packet loss, jitter, and delay. Multiple spanning trees (MST), an IEEE 802.1s standard addresses STP limitations by allowing more links to be utilized in the network, but MST is not class of service (CoS) aware.
    b. VLAN address space: Ethernet's VLAN address limitation of up to 4096 VLANs is Ethernet's bottleneck for a service provider environment to scale the network.

The next generation Ethernet services require guarantees of service performance, typically specified in a Service Level Agreement (SLA). Consequently, for Ethernet to evolve as a next generation networking technology and rival Frame Relay, ATM and Private Line, it must be able to provide QoS controls that enable a service provider to guarantee a service class performance.

3.2 Standards-based, Carrier-grade

Ethernet services standards have come a long way in the past two years, but much more needs to be done to enable multi-vendor, carrier grade networks to be built. The relevant standards bodies (MEF, IEEE, ITU and IETF) need to continue to focus on the definition of a full set of Ethernet service Operation, Administration, and Maintenance (OAM) capabilities. Functionalities such as Ethernet link OAM, end-to-end EVC OAM

(including connectivity management, performance monitoring and L2 ping/trace), Ethernet Local Management Interface (E-LMI) and fault propagation across multi-layered networks will be critical to the success of metro Ethernet services. In addition to OAM, other Ethernet standards-related issues need continued focus, namely, QoS, scalability, definition of an External Network to Network Interface (E-NNI) and a more enriched set of Ethernet service definitions.

Vendors need to quickly embrace these emerging Ethernet standards, which need to be implemented fully in each of the relevant network element types – switches, routers, multiplexers, Network Interface Devices (NIDs) – as well as in Element Management Systems (EMS). Full support for Generic Framing Procedure-Frame (GFP-F), Virtual Concatenation (VCAT) and Link Capacity Adjustment Scheme (LCAS) are also critical.

3.3 Service Availability

Metro Ethernet services are attractive due to their high-speed access options (10 Mbps to 10 Gbps), ability to prioritize traffic and the relatively low cost per Mbps as compared with other services. Most customer sites require optical fiber access (there are some Ethernet over copper applications, too). There are several optical fiber-based access methods either in use today or planned for near-term deployment. The following optical access methods help to broaden the available market for Ethernet services:

- Dedicated fiber pair per customer
- Dedicated λ per customer over shared fiber pairs using CWDM or DWDM systems

- Dedicated TDM channel over shared fiber pairs using Next Generation (NG) SONET systems
- Shared transport, using L2 switching at the customer site
- Shared transport over Passive Optical Network (PON) technologies

In general, the shared fiber pair access methods provide two important characteristics for customers - increased reliability in the access network and lower cost per Mbps. It is the convergence of Layer 2/1 requirements on network elements that presents the biggest challenge in broadening the reach of these new access methods.

3.4 Converged Layer 2/1 Network

To date, most network element implementations have done very little in supporting the integration of Ethernet services in a converged Layer 2/1 network model. Router and switch vendors have not delivered to date the full suite of Layer 1 (L1) features required (GFP-F, VCAT, LCAS) to enable SONET access to switched Ethernet networks, and most L1 vendors have not implemented the full set of service attributes required for switched Ethernet services support. As a result, a service provider needs to deal with the added cost, reduced reliability and increased complexity of disjoint networks.

The converged Layer 2/1 network demands that vendors develop a common set of L1 and L2 features that provide a high degree of reliability and manageability, including end-to-end service management. In addition, ‘virtual UNI’ type functionality that provides a multiplexing function on an aggregated interface will likely be an essential element of this network.

3.5 Integrated Service Management

The traditional Service provider Operational Support System (OSS) infrastructure is complex, and predominantly focused on a network layer for a particular service.

Ethernet service provisioning systems need to have visibility and control of L1 network resources to quickly and efficiently provision the switched Ethernet service end-to-end. This implies the need for a tighter coupling of the inventory management and service activation functions across each network layer.

3.6 Metro Ethernet Forum (MEF) Network Models

The basic Metro Ethernet Network (MEN) model, as shown in Figure 20, utilizes customer equipment that uses a standard Ethernet interface attached directly to a port on a switch at the service provider network edge. The customer premises equipment known as the Customer Equipment (CE) connects to a User-Network-Interface (UNI), which is in turn connected to a (10, 100, 1000 Mbps) port, on the Service provider edge switch. A UNI is a standard Ethernet interface that is the point of demarcation between the customer equipment and the service provider network.

The MEN models have been defined by the Metro Ethernet Forum (MEF), a forum of over 65 companies around the globe, that has been driving the Ethernet services model, definitions and architectures to help realize this growth potential [MEF10].

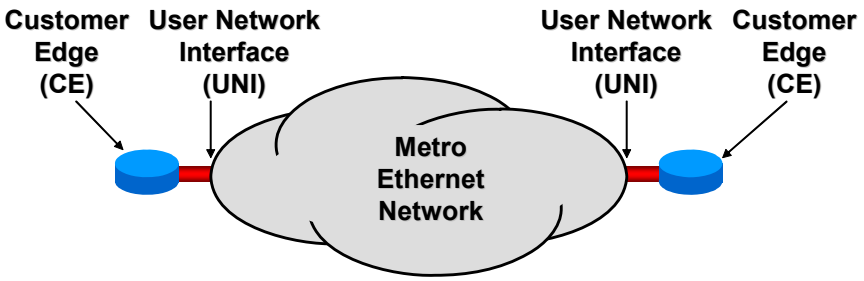

Figure 20: Basic Ethernet Service Model

Ethernet can be modeled closely with Frame Relay and ATM. However, to do this modeling alignment, the concept of Permanent Virtual Circuit (PVC), which is a point-to-point connection, needs to be created in an Ethernet model. The Metro Ethernet Forum proposes an Ethernet Virtual Connection (EVC) to be analogous to that of an ATM PVC. An EVC is an association between two or more customer UNI.

3.6.1 USER NETWORK INTERFACE (UNI)

A User Network Interface (UNI), as defined by MEF technical specifications 11 [92], is used by a service provider to deliver service to a subscriber. The UNI terminates at a demarcation point, which is the technical and operational interface that divides the service responsibility between a provider and a subscriber. The UNI can be further divided into two sub-sets of functional capabilities supported by the UNI-C (at the subscriber equipment) and the UNI-N (at the provider facility).

Service providers rely on standards bodies (ITU-T, IEEE) and forums (EFMA[7], MEF) to standardize the protocols, functional definitions, and architectural framework capabilities such as connectivity, fault isolation, connectivity verification, statistical management, signaling, operational management, and traffic management. MEF has defined a UNI

[7] Ethernet in the First Mile (EFM is now part of MEF

architecture model (see Figure 21), which describes all aspects of the interface between a subscriber network and service provider network. The UNI is physically implemented over a bi-directional link that provides the various data, control and management plane capabilities.

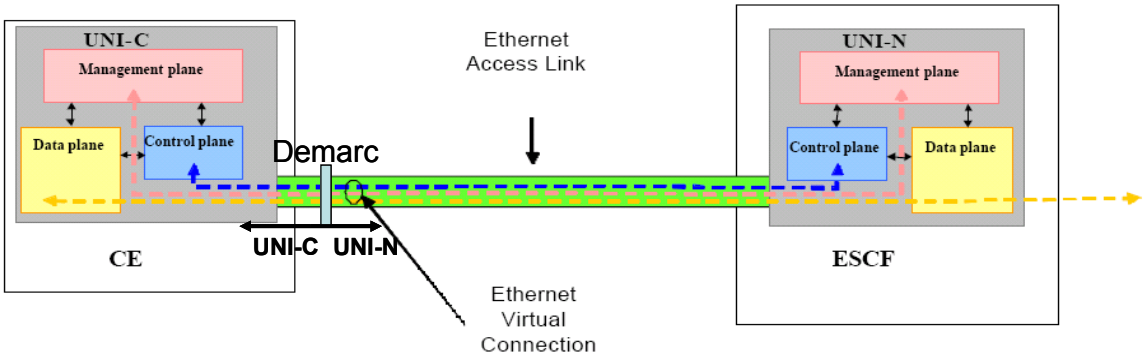

Figure 21: MEF UNI model

Traditionally, subscribers are responsible for all aspects of the UNI on their site up to the demarcation point, unless a subscriber out sources the function to a provider as part of a managed service.

The UNI allows the automation of service provisioning between the UNI-N (on the service provider side) and UNI-C (on the subscriber side). The UNI allows the provider to support service operations, administration and management without direct access to the subscriber equipment. Also, another purpose for the UNI is to facilitate the configuration of Ethernet Virtual Connections (EVCs) across the service provider network for Ethernet based Layer 2 Services. It also provides a framework to access other services such as Internet, voice, or video via the provider's layer 2 services [11], [79], [93].

The MEF UNI model specifies the sub-set of functional capabilities supported by the UNI-C and the UNI-N, which are described in the next sections. These functional

capabilities define how to operate, administer, manage and provision the service across MEN boundaries.

3.6.1.1 UNI Reference Model

The UNI building blocks consist of three main planes: Data, Control, and Management planes. Each block is defined by its various layers, capabilities and processes.

Figure 22: UNI building blocks

3.6.1.2 UNI DATA PLANE

The UNI data plane defines the means of transporting information across the UNI reference point. It provides the mechanism to transport various user traffic flows across the UNI demarc interface and includes the definition of the physical and data link layer such as:

- Ethernet frame flow

  The physical and data link layers across the UNI are based on the IEEE 802.3/Ethernet PHY & MAC. The subscriber sends Ethernet frames over the UNI across an Ethernet Virtual Connection (EVC).

- Tagging

  A subscriber may tag the Ethernet frames using IEEE 802.1q VLAN tags before entering the service provider domain. The service provider may use tagging to segregate customer traffic or for traffic management purposes. Tagging should be

based on agreed upon delineation method between the provider and customer that can be identified by VLAN ID.

- Traffic Management

  Based on the service provider offerings and what the customer purchased, the EVC identification is used to bind the subscriber to certain traffic management policies agreed upon in the SLA. Furthermore, a subscriber may be provided with one or more classes of service as determined by the QOS marking values for a given service.

3.6.1.3 UNI CONTROL PLANE

The UNI control plane defines the means for the customer and the service provider to agree on how to make use of the UNI data plane subject to a bilateral contract agreement. It also provides the means for agreement on the types and characteristics of the Ethernet service(s) that will be provided. To accomplish this task, an in-band or out-of-band communication mechanism is used.

The UNI control plane consists of two functions:

- Static service discovery

  The static service discovery function allows the auto-configuration of the subscriber equipment. In this manner, the Ethernet services may be activated without manual configuration intervention by the customer. An Ethernet service requires some inherent configuration to be made on the customer and service provider equipment such as the creation of the EVC between the communicating UNIs, or configuration of the service attributes and/or addressing information to reach a specific service endpoint. To perform this manually may be a cumbersome

task for the customer. Therefore, it is useful to provide an auto-configuration function to provision the customer's equipment with any negotiated service-specific parameters. This process is static in the sense that the service is already provisioned by the MEN provider, and does not require further provisioning information from the customer at the time the customer retrieves the configuration parameters.

- Dynamic connection setup

  A dynamic connection setup allows for the activation of the service dynamically. Hence, the provider can provision the network at the time the service request is processed (received from the customer). This obviously increases the complexity of the service activation process, but it also increases the service provider's configuration flexibility and manageability of the service up to the customer equipment.

3.6.1.4 UNI MANAGEMENT PLANE

The UNI management plane, as defined by MEF technical specifications 11 [92], controls the operation of the UNI data and control planes. This includes the provisioning and management steps required for the UNI data plane to support the contracted service. Also it controls how the control plane may operate, i.e., it controls when static service discovery and/or dynamic connection setup are allowed.

- Provisioning and QoS management

  The management plane ensures that all elements in the service provider network and the customer equipment are configured to support the EVC's service attributes. Also, the management plane may interact (out-of-band) with a network

management system in the service provider domain to determine whether a particular configuration should be used. Alternatively, the management plane may also interact directly with the UNI control plane. The control plane (if present) will then be used by the customer to either statically retrieve or dynamically request the configuration parameters, whereas, the management plane will then be concerned with the provisioning of the provider's network elements.

QoS management is a subset of the service provisioning function as it includes the provisioning of service attributes, traffic policers, shapers, etc. The QoS Management enables the correct behavior in the UNI data plane and the Ethernet service activation.

- Protection and restoration

  Protection is the ability to provide an alternate facility in the event of a failed link. Restoration is the ability to repair failures and restore service in the event of a failure. Protection and restoration is mainly implemented in a service provider environment, however, it can also be implemented in a customer network. A customer or a provider with two physical links may be activated as part of the same service for protection (link aggregation or dual homing). If one link fails, then the other link is used as an alternate facility. It may be required that the right interaction between the management planes of both links take place for the link switch to occur. Also, load balancing of the two links can be performed during certain times if desired.

Operation, Administration and Maintenance (OAM)

Operation, Administration and Maintenance (OAM) is primarily concerned with fault and performance management. It includes connectivity verification, performance monitoring and statistics gathering. A management channel is required between the subscriber equipment and the provider network over the UNI to accomplish OAM functionality, such as:

- Connectivity verification to make sure that the Ethernet traffic is sent over the UNI and is not lost
- Performance monitoring to assert to what extent errors have or are occurring over the UNI
- Statistical gathering to collect the number of frames successfully sent, lost, or corrupted for verification purposes over the UNI

3.6.1.5 UNI ETHERNET VIRTUAL CONNECTION (UNI EVC)

The UNI Ethernet Virtual Connections (UNI EVC), as defined by the MEF 11 technical specification [92], refers to the logical segment that interconnects the subscriber equipment to the provider facility. The UNI allows these UNI EVCs to be automatically provisioned to the subscriber device without customer involvement in setting up the subscribed services. The UNI EVC may operate in a static or dynamic configuration mode. In the dynamic mode, UNI EVC may have automated self-provisioning attributes of the subscribed service. As each subscriber enrolls into a service that the provider offers, the UNI provides the logical connectivity to these services using an automated signaling of the service via the UNI EVC.

3.6.1.6 UNI MODES

In order to make use of the Ethernet services, the subscriber equipment needs a number of parameters that describe the Ethernet service. Instead of configuring these parameters manually, as it is done today, the introduction of an Ethernet control plane protocol would automate this process. The UNI may operate in a permanent virtual connection (PVC) or switched virtual connection (SVC) mode. The following describes the two EVC modes:

- PVC Mode

  The PVC mode of operation allows the provider edge switch to provision, configure, and distribute UNI EVCs and their associated service attributes to the subscriber equipment statically in a uni-directional manner similar to Frame Relay LMI and ATM ILMI. The subscriber equipment in PVC mode can retrieve certain information from the network through an automated link management interface, such as Ethernet local management I/F (ELMI). Upon powering up, the subscriber equipment uses the link management interface to learn about the connections at a given UNI and configures itself appropriately for those connections. Potentially, this could allow a customer equipment using Internet Protocol utilizes inverse address resolution protocol (ARP) to obtain adjacency information at the far end of the EVC. This enables a more effective self-configuration of the two or more subscriber equipment for the resolution of IP addressing to EVCs.

- SVC Mode

  The SVC mode of operation allows the subscriber equipment to request, signal and negotiate UNI EVCs and its associated service attributes to the provider edge switch dynamically.

3.6.1.7 UNI Service Attributes

UNI Service Attributes are identified as objects. Each object describes a specific service attribute that can be negotiated between provider switch and customer equipment at the interface level. The UNI service attributes serve as a repository of object information defining service capabilities in which the UNI is able to perform add, modify, or delete operations. UNI Service Attributes include:

- Site Identifier
- Physical Medium
- Speed
- The maximum number of EVCs that may be supported
- Ingress Bandwidth Profiles for traffic policing, and
- Performance

3.6.1.8 EVC Service Attributes

EVC Service Attributes are identified as objects. These objects are dynamic or static depending on the type of EVC (SVC, PVC). EVC Service Attributes include:

- EVC Type (Point-to-Point or Multipoint-to-Multipoint)
- UNI List
- CE-VLAN ID Preservation
- Class of Service Identification
- EVC Performance, (e.g., Service Activation Time, Frame Delay, Frame Loss)
- Layer 2 Control Protocol Processing (tunnel or discard)

3.7 Summary of Metro Ethernet Services

Figure 23 depicts the current summary of metro Ethernet services that include the two service families: E-line and E-LAN. Next sections will cover metro Ethernet services and their corresponding transport technology as specified by the MEF technical specifications 1-12.

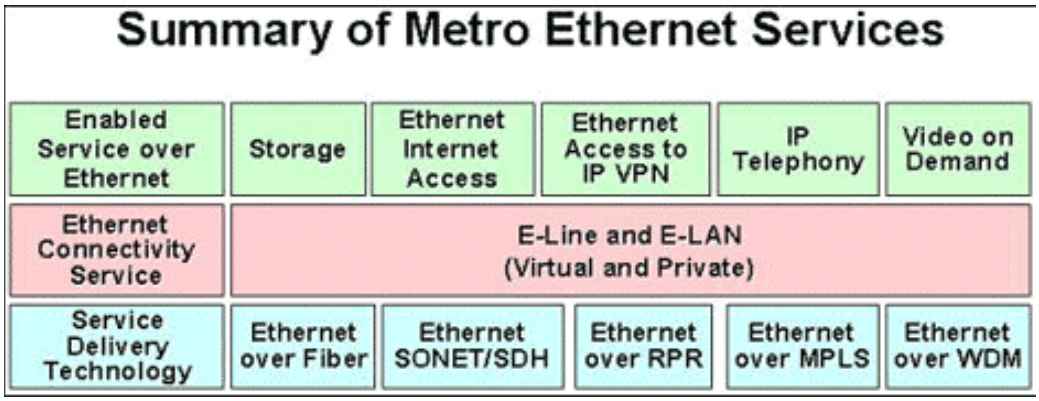

Figure 23: Summary of Metro Ethernet Services

3.7.1 Ethernet Connectivity Services

3.7.1.1 Ethernet Local Area Network (E-LAN)

E-LAN is a switched Ethernet service designed to support LAN interconnect applications (see Figure 24).

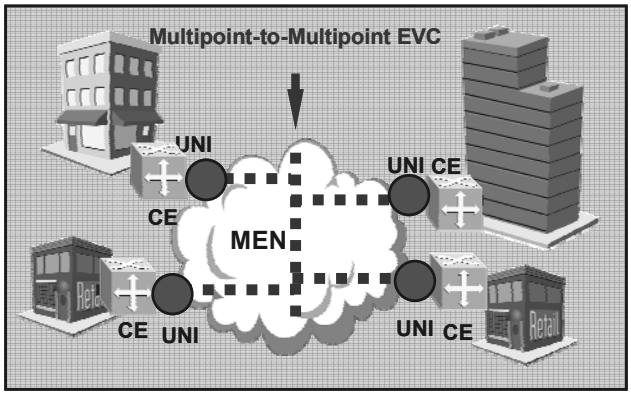

Figure 24: Switched Ethernet LAN Network Model

E-LAN uses a single multipoint-to-multipoint Ethernet Virtual Connection (EVC) to connect two or more User Network Interfaces (UNIs). This allows the customer to extend their LAN across the metro network at native Ethernet speeds of 10Mbps, 100Mbps or 1000Mbps (GigE).

An Ethernet LAN has specific service parameters or attributes applied to its virtual connection or UNI. These attributes include:

1. Bandwidth Profiles
2. Physical Medium (Physical Interface: Copper, Coax, Fiber)
3. Port Speed (10M, 100M, 1000M, 10000M)
4. Mode of the UNI Speed (HD, FD)
5. Traffic Parameters (CIR, CBS, PIR, MBS)
6. Performance Parameters (Quality of Experience (QoE), availability, service activation SLA, delay, loss and jitter)
7. Class of Service (Port, VID, 802.1p, Diffserv, source/destination MAC address

3.7.2 Ethernet-Line (E-Line)

Ethernet-Line is a point to point service with two key services: Ethernet virtual private line (EVPL) and Ethernet private line (EPL). These two services are analogous to private line and frame relay services respectively.

3.7.2.1 Ethernet Virtual Private Line (EVPL)

EVPL provides transport and switching for point-to-point Ethernet Virtual Connections (EVCs) among two or more UNIs, allowing customers to build logical networks similar

to the existing L2 services, such as, frame and cell relay services. Figure 25 depicts a typical EVPL application without showing the network details.

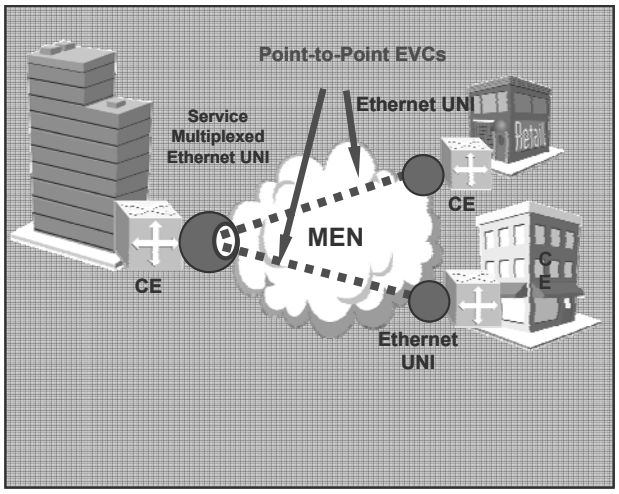

Figure 25: Typical EVPL application

EVPL offers access speeds of 10Mbps, 100Mbps or 1000Mbps (GigE) and granular EVC increments as low as 1Mbps in each of three different classes of service (Basic, Priority Data, Real-time).

3.7.2.2 Ethernet Private Line (EPL)

EPL is a dedicated point-to-point private line Ethernet service designed for customers requiring a dedicated bandwidth channel across a shared fiber based infrastructure. EPL provides customers with a high degree of reliability, service transparency, and performance end to end. Figure 26 depicts a typical EPL application, without showing the network details.

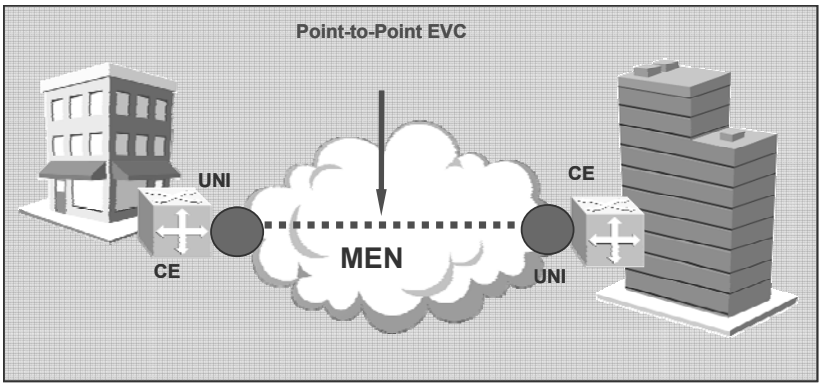

Figure 26: Typical Ethernet Private Line (point to point EVC)

EPL offers access speeds of 10Mbps, 100Mbps or 1000Mbps (GigE) and transport channel granularities as low as STS-1. When a customer application uses a larger frame size over native Ethernet frames, interframe gap and transport rates need to be cared for in order to map them over the SONET infrastructure.

The short-term solution proposed by this research augments the current high-level end-to-end Ethernet services architecture models with an admission control for Ethernet Services (ACES). The current Ethernet Services Architecture (see Figure 27) includes the following:

- UNI Loops/Metro Access: This is the last mile (a.k.a first mile) access to the network and there are multiple access methods such as: Dedicated fiber, xWDM, and NG-SONET.
- Metro Edge: This is the switching function that determines whether to switch packets locally or transport them to the MEN backbone or the Wide Area Network.
- MAN/WAN Aggregation: This is the switching function that determines whether to process packets within a Metro or transport them outside of the MEN into the

national L1/L2 end-to-end or MPLS cloud or into one of the Value Add Services networks such as IP-VPN or the Internet.

- Service Provider with a National L1/L2 or MPLS: This is a Service Provider's national MPLS network that transports a customer Ethernet Virtual Circuit or domain across a national MPLS backbone connecting one MEN to another.

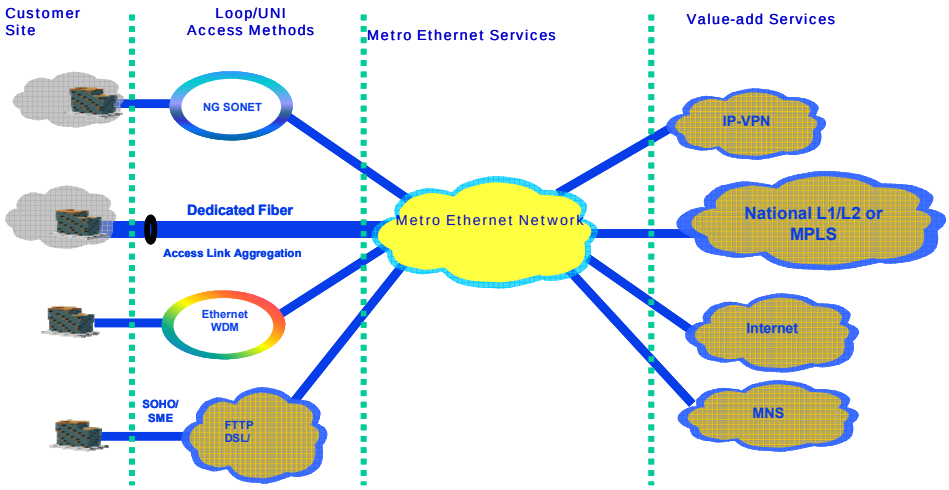

Figure 27: High level end-to-end Ethernet architecture

3.8 Service Delivery Technology

Ethernet services can be delivered over multiple transport technologies. The underlying technology determines the existing analogous services and characteristics. The following sections cover the most common Ethernet transport technologies used in metro areas.

3.8.1 Ethernet over Fiber

Ethernet over fiber is a low cost service provided that fiber is abundant and other lower cost alternatives do not exist. A service provider fiber plant technician determines the type of access topology and the physical transmission medium used to connect to a

customer. In areas where fiber is abundant, a pair of fiber is used to access the customer premises and provide connectivity to the nearest provider point of presence. However, in areas where fiber is not widely available, it is important to utilize existing fiber with wavelength technology in order to maximize their use; or it may be necessary to exploit other technologies that help reduce the number of fibers needed. Such technologies are: SONET, CWDM or DWDM in order to maximize the available fiber strands.

3.8.2 Ethernet over SONET

Synchronous Optical NETwork (SONET) is a time division multiplexing (TDM) transport technology. It is the dominant technology deployed in the optical transport network in support of traditional voice circuits. It carries many signals of different capacities through a synchronous, flexible, optical hierarchy. The SONET base signal is referred to as Synchronous Transport Signal level-1, or simply STS-1, which operates at 51.84 Mb/s. In addition, SONET has higher-level signals that are integer multiples of STS-1, creating the family of STS-N signals as shown in Table 2. STS-N signal is composed of N byte-interleaved STS-1 signals and the optical counterpart for each STS-N signal is designated OC-N (Optical Carrier level-N).

Table 2: SONET Electrical/Optical signals and associated line rates

Electrical Signal, or Optical Signal	*Line Rate*	*Capacity*
*STS-1 or **OC-1	51.84 Mbps	1 DS3 = 28 DS1
STS-3 or OC-3	155.52 Mbps	3 DS3 = 84 DS1
STS-12 or OC-12	622.08 Mbps	12 DS3 = 336 DS1
STS-48 or OC-48	2,488.32 Mbps	48DS3 = 1344 DS1
STS-192 or OC-192	9,953.28 Mbps	192 DS3 = 5376 DS1
STS-768 or OC-768	39,813.12 Mbps	768 DS3 = 21504 DS1
* STS-N = Synchronous Transport Mode-N ** OC-N = Optical Carrier level – N		

A service provider can offer any type of service over SONET but is required to map the traffic to a SONET STS-N signal. Ethernet over legacy SONET does not provide for an efficient transport for 10 Mbps interfaces as they need to be mapped to STS-1 or 51.84 Mbps. A 100 Mbps Ethernet needs to be mapped to an STS-3 signal or 155.52 Mbps. Also, a Gigabit interface will need to be mapped to either STS-12 or 622.08 Mbps or 2,488.32 Mbps. Clearly this mapping is inefficient and expensive. However, the new SONET WAN PHY (although is it is not SONET compliant), does adapt the 10 Gbps data rate to SONET OC-192, but not the SONET specifications for timing and jitter requirements.

As discussed, Ethernet physical layer rates are different from the SONET TDM rates. When attempting to carry Ethernet over SONET, one approach could be to map Ethernet into a larger SONET payload capacity. However, such an approach would waste a large portion of the customer paid SONET transport bandwidth. Based on a mean frame size of 400 bytes in a typical metro area network, it is expected that an additional 20 bytes is

added to each frame resulting in about 5% overhead on each line interface. Virtual concatenation will alleviate this problem to a large extent where individual VT1.5, STS-1, or STS-3c time slots can be logically grouped together to obtain a SONET bandwidth. Therefore, next generation SONET (NG-SONET) presents a service provider with an alternative to utilize virtual concatenation in order to increase the efficiency of transporting data over SONET. Virtual Concatenation (ITU-T G.707) puts Ethernet frames into SONET payloads, rather than using contiguously concatenated SONET payloads, it uses the base SONET payloads and groups these payloads to create a larger "right-sized" aggregate payload. Hence, it is best to use this technique to carry Ethernet over SONET. In addition, Virtual Concatenation enables the payload capacity to vary thus right-sizing the payload to match that of the customer data rate. This sizing allows a greater number of Ethernet channels to be mapped into the SONET signal.

In NG-SONET networks, new standards such as generic frame procedure (GFP), Virtual Concatenation (VCAT) and link capacity adjustment scheme (LCAS), are developed to facilitate the mapping of client signals into SONET. Ethernet frames are encapsulated one at a time into generic frame procedure (GFP) frames, which are then mapped into a SONET channel using virtual concatenation for carriage across the embedded network. LCAS can be used to keep a connection running at a reduced rate if members of the virtual concatenation group fail, or add more members if the customer requests additional bandwidth.

The impact of a frame-mapped approach is that certain Ethernet mechanisms cannot be transported transparently over the SONET network. While the GFP standard allows for both frame mapping (GFP-F) and transparent mapping (GFP-T) for EoS, the GFP-T

standard is continuing to undergo modifications to efficiently transport line rate and sub-rate Ethernet. The standard time-division multiplexed (TDM) rates SONET systems transport are payload capacities of 1.536 Mb/s (DS1), 1.600 Mb/s (VT1.5), 44.210 Mb/s (DS3), 48.384 Mb/s (STS-1), 149.76 Mb/s (STS-3c), 599.04 Mb/s (STS-12c), 2396.16 Mb/s (STS-48c), and 9584.64 Mb/s (STS-192c).

SONET, like other transport technologies, rely on the physical layout of the fiber plant to determine its logical topology. Therefore, the physical and the logical topologies may differ significantly. Figure 28 depicts an example of a service provider metro area with fiber and SONET network layout [94].

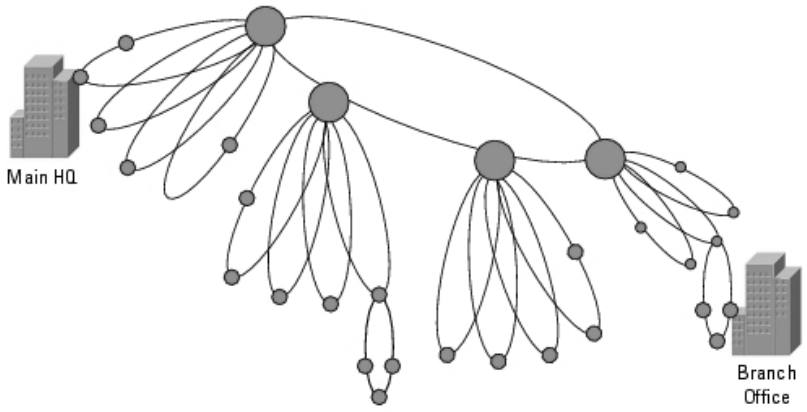

Figure 28: Service Provider SONET metro network over fiber facility layout

SONET topology is ring based and has up to three distinct tiers in a large metro area:

1. The Access Ring: The access ring provides access out to the customer premises and will support single premises (one or more customers).
2. The Collector Ring: The Collector Ring is a termination point for the access rings. The access-collector ring aggregates multiple access rings into a single fiber distribution frame.

3. The Hub Ring: This is a central point of presence in a CO that aggregates collector rings. Each CO POP is typically interconnected over a high-speed (OC-48 or 192 ring) inter-office and/or regional ring.

SONET/SDH networks are evolving to support Ethernet Virtual Private Lines and Ethernet Virtual Private LANs through integrated Ethernet switching or RPR technology.

The virtually concatenated SONET payload is viewed as a set of byte lanes, where each byte is carried in a separate channel. The data stream to be transported is mapped, byte by byte, into the lanes of the virtually concatenated payload. The individual SONET channels that make up the virtually concatenated group can then be independently transported over the SONET network.

Aligning the SONET paths requires buffering each path. The buffering required should be sufficient to store all paths up to the point that the slowest path is received. The slowest path is the path that experiences the greatest network latency between the transmitter and the receiver. The difference in latency between the fastest and the slowest path is a result of the different paths taking diverse routes through the network. Once the slowest path has been received, then the alignment can begin and the data can be extracted.

A second standard developed recently is the Generic Framing Procedure (GFP, ITU-T G.7041.) This provides the means of delineating packet data in a SONET payload without the need for variable bandwidth expansion. Each data packet has an 8 byte header added which indicates the start, type and length of the data frame. GFP supports an extended header, for future uses that may need addressing and multiplexing functions.

Virtual Concatenation and GFP together with necessary Ethernet and SONET path processing functions can be handled in a single device, which may sit on a line card in an Add/Drop Multiplexer (ADM) or a SONET cross connect. This adds Ethernet transport services to the existing SONET network. The expensive alternative is replacing the SONET circuit based network with an all-packet infrastructure, requiring replacement of every box in the network, and inspection and switching of each and every packet at every node. Also, supporting traditional TDM services becomes very difficult in a packet-only network. It will be some time before packet technologies are able to approach the characteristics of today's SONET-based networks such as high performance, reliability and manageability.

3.8.3 Ethernet over Resilient Packet Ring (EoRPR)

Resilient Packet Ring (RPR) is an IEEE (802.17) standard that supports sub 50ms ring-based resiliency on packet switched network architectures. RPR can be implemented over SONET/SDH or native Ethernet transport networks. It supports a significant degree of bandwidth efficiency on rings through the implementation of bandwidth sharing, spatial reuse, and statistical multiplexing. RPR is used to support Ethernet Virtual Private Lines and Virtual Private LANs either shared or dedicated within a metro area network.

3.8.4 Ethernet over MPLS

Ethernet over Multiple Protocol Label Switches support Ethernet Virtual Private Line over a WAN that spans long distances. This is done via a pseudo wire tunnel, where Ethernet frames are encapsulated, and are then transmitted through the packet network connecting the two tunnel endpoints. Hence, pseudo-wire emulates layer 2 Ethernet service in a point to point, simple, and flexible service.

In addition, MPLS supports layer-2 virtual private network (L2 VPN) via a provider provisioned pseudo-wires called Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS).

Virtual Private Wire Service (VPWS) is a layer-2 service that provides Layer-2 point-to-point connectivity (e.g. point-to-point Ethernet) across an MPLS-enabled IP network. VPWS is setup over a nailed pseudo wire virtual circuit across an MPLS cloud. If the provider edge equipment has the right kind of intelligence, this service model can allow inter-working between different Layer 2 attachment circuits. Depending on whether service multiplexing, VLAN transparency or All-to-One Bundling is employed, VPWS can be used to create port-based or VLAN-based Ethernet private lines.

3.8.5 Ethernet over WDM

WDM provides the ability to transmit a high number of closely spaced frequency signals or light paths over a fiber cable. There are three flavors of WDM [94]:

1. Coarse Wave Division Multiplexing (CWDM)
2. Dense Wave Division Multiplexing (DWDM)
3. Wide Wave Division Multiplexing (WWDM)

Ethernet over CWDM is best suited in a metro access network with access speeds up to 10 Gbps, however, the most common speed supported is up to 2.5Gbps. Ethernet over DWDM is best suited in a larger metro and long haul networks with access speeds up to 10 GigE. Ethernet over WWDM is best suited to carry 10G in large metro and long haul backbones.

3.8.5.1 Ethernet over CWDM

Ethernet metro access is amongst the most important CWDM applications for Enterprise customers. CWDM frequencies are widely separated in the 1550nm band in the order of 20nm per wave length. CWDM provides up to 32 channels into the fiber transmission window. Ethernet over CWDM is a plug and play with CWDM providing a point-to-point connectivity on individually multiplexed frequencies.

3.8.5.2 Ethernet over DWDM

Ethernet over DWDM provides for high reliability, high protectiveness and high effective bandwidth scalability in the metro and wide area networks. Also, it is cost efficient as it provides the ability to maximize the provider's fiber investment. DWDM packs far more optical channels into the fiber transmission window (~200 channels) than does CWDM. DWDM eliminates the bandwidth bottlenecks for enterprise applications running over LAN and WAN backbones, such as storage.

3.8.5.3 Ethernet over WWDM

A proposed, IEEE P802.3ae, standard for 10GigE Ethernet transceiver.

3.9 Summary

The majority of network data today exists as an Ethernet packet, which originates from a personal workstation or a network server (file server, application server or web hosting server). These computers typically connect to a network with a 100 Mbps or Gigabit Ethernet interface. Hence, an Ethernet packet starts and terminates as an Ethernet packet. Therefore, it is highly desirable to keep Ethernet format end-to-end and prevent the need for any protocol conversions. The problem, however, is that Ethernet was designed mainly for LANs and doesn't readily scale to WANs extending beyond a building or

campus. Our attempt is to improve the scalability and reliability of Ethernet networks in the immediate to near term period. Also, our vision is to propose a longer term solution which enables Ethernet networks to stretch around the globe and offer predictable performance and reliability.

The benefit of Ethernet over transport technologies such as Ethernet over SONET (EoS) is that it is evolutionary and uses existing protocols and networking equipment at the customer premises. Also, it allows data packets to be transported over the same infrastructure internally within a customer site, or externally to a service provider network without protocol conversion.

Ethernet services can be provided over a wide variety of transport technologies and infrastructure such as SONET/SDH, MPLS, WDM and OTN (optical transport network). Often, the result is a combination of multiple protocols. In addition, Operations, Management, and Administration (OAM) functions are needed across these protocols.

3.10 WHAT ARE THE CHALLENGES?

Enterprise customers' experience with Ethernet in the LAN over the past 25 years has contributed to maintaining its current simplicity and maturity status within the LAN. However, these Enterprise customers are now pushing Ethernet into the MAN and the WAN to extend their LANs and to keep things simple without the need for protocol conversion and above all, to maintain low cost. In addition, customers are pushing for the convergence of their information (voice, video streaming, and data) onto a single pipe.

Some of the key service requirements include:

1. UNI characteristics that allow an enterprise to retain its traffic characteristics through a service provider network;
2. Enterprise customers view of Ethernet technology as a simple and mature technology with scalable interfaces that are easy to plug and play;
3. Cost.

However, Ethernet technology is not truly ready for Service Provider Environment. The following lists some of the key service concerns for the short-term service rollout:

- Ethernet QoS is not mature enough in a multi-customer environment as that in a service provider environment
    a. Availability: 3 9's to 5 9's
    b. Jitter and Delay: 50msec to few seconds
    c. Data delivery (99.0 to 99.999%) or Packet Loss (.01% - 1%)
- Ethernet LAN based switching protocol is not currently suitable as a WAN based protocol
    a. Switching protocol is slow to converge in the network
    b. SLA concept is not available
    c. Best Effort with limited prioritization queues
    d. Priority bit in the LAN that specifies enterprise priority traffic
- LAN traffic does not have the concept of a service definition that is bandwidth based. For example, MEF E-Line/E-VPL service definition that may have multiple classes with performance guarantees over a UNI. Hence, in order to offer these services then we need to address the following questions:

    a. How to handle QoS, bandwidth guarantees, and priority
    b. How to perform a fail-over architecture and maintain service level guarantees

- Does not have the mechanism to support the definition of new services and network requirements, such as:
    a. Preservation
    b. Type of service (ToS) bits
    c. Differentiated Service (Diff Serv code points)
    d. Reservation and service guarantees via signaling such as RSVP

Among the key inhibiting factors in realizing the potential of a global Ethernet services include:

- Lack of a standards-based carrier-grade set of capabilities for switched Ethernet services
- Need for multiple access methods to improve the ability to deliver services on time to all customer sites
- Need for network elements that provide L2/1 convergence of the transport plane with the right hooks into control and management planes
- Need for Admission Control, routing, and bandwidth adjustment requirements for a given EVC across the converged L2/1 network.

4. Short-term Solution

4.1 Overview

Ethernet's wide deployment in the enterprise LANs, its continuous expansion in the network deployments, and its scalable transmission speeds have provided increased optimism that Ethernet is expanding everywhere and is becoming the key technology for data convergence. This is not because it is the best technology, but because it is the cheapest, simplest, and provides the most flexibility and scalability in the access networks. In addition, Ethernet's scalable interface speed, ranging from 10 Mbps to 10 Gbps, has led to an increased gap between access and backbone speeds deployed in a carrier access environment such as Frame Relay, ATM and IP. To bridge this gap, Ethernet traffic must reach service provider core optical networks such as SONET & DWDM, or ride the MPLS core network with a minimum number of hops. In addition for Ethernet, to become a true carrier grade service and a true plug and play WAN technology, it needs to evolve from a best effort service to a service that supports classes of service with guaranteed service level performance.

To alleviate Ethernet's lack of support for carrier grade services, we introduce a novel admission control for Ethernet services (ACES) that ensures existing and new initiated virtual circuits meet the agreed upon performance criteria and are within acceptable service metrics. ACES will select the most efficient path through the L1/L2 network; if sufficient bandwidth is not available; ACES will signal the optical transport network for additional bandwidth capacity, and if this bandwidth capacity is not available, ACES will then reject or put the service request in a wait state. A service request enters the wait state if the nature of the request is not urgent and or on demand.

ACES is a near term solution for a service provider that enables offering new Ethernet services over existing legacy systems. ACES admission control model can also be applied into different Ethernet virtual private line architectures such as Ethernet VLAN bridging, Ethernet over SONET, Ethernet over Optical Transport Network (light path), Ethernet over MPLS, and Ethernet over Resilient Packet Ring (RPR) without the need to replace the existing infrastructure.

This chapter discusses the short-term solution in support of new Ethernet services over existing legacy infrastructure. This includes Ethernet architecture in the MAN and the WAN. In addition, for Ethernet to succeed as a carrier grade transport technology with support for triple play, quad play and enterprise critical applications, Ethernet needs to offer quality of service and differentiated classes of services with guarantees.

This chapter will first discuss the ACES basic workings and architecture. Then, it will cover the three network topologies that were used for ACES simulations, which are:

1. Basic network model - A 3-node network architecture model
2. A typical metro area network model – A metro Ethernet network architecture model, which consists of two core switches and six edge switches demonstrating a typical metro area network
3. An Enhanced model - A metro Ethernet network inter-working with Optical Transport Network

These models will demonstrate the basic need for a ACES, whether as a centralized or as a distributed admission control system.

At the end of the chapter, a summary is provided to illustrate the comparison of the simulation results achieved. Further, it discusses the enhancements needed for ACES, which include: 1) Performance measurements feedback from network probes to ensure service level guarantees; 2) Converged L1/L2 management and control planes.

4.2 Introduction to ACES

The proposed ACES solution provides a centralized admission control and path selection function for provisioning EVCs in a Service Provider's switched Ethernet network, while controlling CoS service level guarantees. Previous solutions focused primarily on LAN environment and did not address QoS capability for Metro/Core Ethernet Services in a Service Provider network [74], [91].

The ACES solution is extensible to various network technologies such as: Provider Bridge, Multi Protocol Label Switching (MPLS) based Layer-2 Virtual Private Network (L2 VPN), Resilient Packet Ring (RPR), Ethernet Passive Optical Network (EPON), etc.

ACES provides a centralized admission control with efficient path selection function for provisioning each EVC request in a multiple spanning tree network. ACES could provide the entire EVC provisioning capability for a given network or it may interface with a separate Provisioning Server, as shown in Figure 29, in the case of larger scale networks.

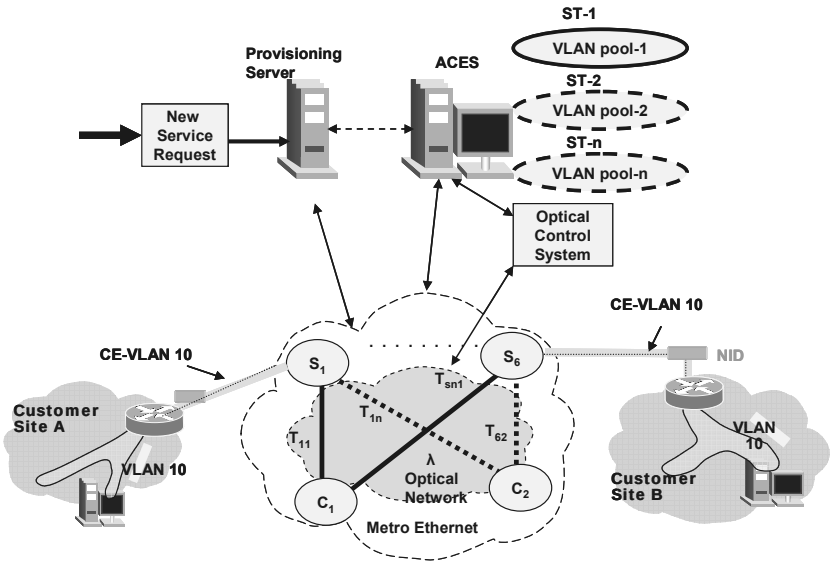

Figure 29: ACES Architecture Diagram

ACES key functions are:

1. Maintain current view of the physical topology;
2. Maintain current view of the logical topologies;
3. Maintain trunk link bandwidth allocation rules per CoS;
4. Assign S-VLAN ID for each EVC from the appropriate VLAN pool;
5. Assign a back up EVC as a redundant path for link failure scenarios;
6. Request the OTN Optical Control System to add a λ between switches or communicate directly with the Optical Network Element.

ACES allows for percentages settings of the classes of service offered based on known historical or empirical data that a service provider has at its own disposal and wishes to use. This empirical data represent service demand based on historical offering in a given metro area. For example, the bandwidth allocations per CoS can be assigned a percentage of the trunk-link speed as follows: 10% Gold (stringent voice grade requirements), 20%

Silver (stringent voice, video or data requirements) and 70% Bronze (best effort to low guarantees), see Table 3.

ACES allows for the topology definitions of metro Ethernet network nodes and associated network interfaces with varying speeds of 1 GigE and up to 10 Gbps capacity per trunk link. In addition, ACES has the flexibility to assign a number of classes of service on each trunk link with a percentage of the link speed. ACES allows some service classes (e.g., bronze and silver) to be oversubscribed, while other service classes (e.g. platinum and gold service classes) are not oversubscribed. ACES provides full flexibility in configuring and modifying the over-subscription ratio for each CoS. Therefore, the sum of the aggregate service class bandwidths provisioned on a given trunk-link may exceed its physical bandwidth capacity.

Table 3: Example of a Trunk-link bandwidth and CoS assignments

Trunk 1			Trunk 2			Trunk 3		
Gold	Silver	Bronze	Gold	Silver	Bronze	Gold	Silver	Bronze
100	200	700	100	200	700	100	200	700

4.3 ACES Algorithms

ACES has the following built-in algorithms:

1. **Spanning Tree:** The basic spanning tree algorithm is used as the reference model with only one root-bridge per network.
2. **Round Robin:** The round robin algorithm is used to assign the VLAN ID in the service provider network to a spanning tree instance. This algorithm allows ACES to toggle amongst the multiple spanning trees (MSTs) instances.

3. **Class of Service**: The class of service algorithm is used to assign one or more service classes to a given MST instance. For example: Gold is assigned to VLAN Pool-1 (spanning tree instance ST-1); Silver is assigned to VLAN Pool-2 (spanning tree instance ST-2); and Bronze is assigned to either pool based on Round Robin algorithm.
4. **Shortest Path:** The shortest path algorithm is used to determine the shortest path between the source node and the destination node regardless of CoS bandwidth requirements. If more than one ST within the set of MSTs have the same shortest path computation, then round robin is added amongst those instances to further balance the load.
5. **CoS-Traffic Engineering (TE)**: The CoS-TE algorithm is used to determine which path (MSTs) have sufficient bandwidth, and route the EVC based on the most (or least efficient) available bandwidth given in a service class.

An EVC service request consists of three attributes, they are:

1. Node pair (Source, Destination),
2. Class of service assigned to the EVC (such as gold, silver, bronze)
3. The bandwidth associated with the EVC CoS (ranges of 1 to 10000 Mbps)

In the ACES simulation, as we assign an EVC, we generate the above attributes randomly. In our simulation models, we use an EVC bandwidth request that ranged between 1 Mbps and 200 Mbps. A simulation run generates up to a specified number of EVC requests (in our example, we used 1000 EVC requests).

Each EVC request passes through each of the ACES algorithms and is processed based on the algorithm's rules. As each EVC is processed and admitted into the network, each trunk-link residual bandwidth along the EVC path, is updated on a per CoS basis. ACES will track each EVC request processed and, before it is admitted into the network, determines whether the needed resources are available to meet the request. If an EVC is not processed for that particular algorithm (indicated by a 'No' in Table 4), then all relevant information up to that point are collected for comparison and for statistical analysis, see Table 4.

Table 4: Example of EVC requests with associated algorithm

EVC Request ID	**Source Node**	**Destination Node**	**CoS (G, S, B)**	**Bandwidth (Mbps)**	**Associate Algorithm used**
1	*B*	*C*	*Bronze*	*80*	*Yes*
…	…	…	…	…	*Yes*
…	…	…	…	…	
…	…	…	…	…	
994	***A***	***B***	***Gold***	***4***	***No***
1000	*A*	*C*	*Bronze*	*12*	*No*

The example of the EVC process shown in Table 4 is repeated per each algorithm specified for the entire number of EVCs to be admitted into the network. This process can be repeated for as many times as needed. For the purpose of this example, the simulation process was repeated 50 times and its data was collected, stored and statistically analyzed. Also, two figures were plotted per network model used and their results are briefly discussed in the corresponding sections.

The Service Provider network can be configured with up to 64 maximum Spanning Trees (STs) based on the IEEE 802.1Q-2003 (formally IEEE 802.1s) standard. However, it is not recommended to use more than several spanning trees in a small to medium size network due to the configuration complexity. Also, it is recommended to use root bridges that are central to the network with the maximum nodal degree number.

4.4 Basic Network Model

Let us consider a Metro Ethernet Network (MEN) with three switches (A, B, C) – see Figure 30. In order to demonstrate the limitation of basic ST and the need for MST, let us assume that all three-trunk links have the same speed and thus the cost budget associated with each link is the same. Let us designate Switch A as the root by assigning it a lower bridge ID than switches B and C. Initially, all three switches send control messages claiming to be root. However, once switches B and C receive BPDUs from A, they determine that A is elected as root-bridge, hence, they will stop sending control messages.

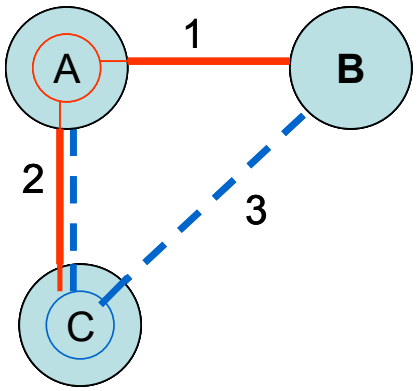

Figure 30: MST over a basic network model

Switches A, B and C will then build their forwarding tables. Switch B will have a path to the root A via trunk link-1 at cost of 4 [70], Switch C will have a path to root A via trunk link-2 at cost 4. The trunk link-3 is unused since both switches block their trunk link-3

interface. Hence, the red spanning tree is formed (solid lines) and data can flow between the switches with no loop in the network.

MST provides the ability to utilize more trunk links in a network where those trunk links were considered blocked using the basic spanning tree. In our example of Figure 30, a service provider can overlay their physical network topology with logical topologies based on MST instances. A given EVC can be assigned to a given MST instance using a variety of algorithms, as needed. For improved efficiency, a second spanning tree (Blue dotted line) needs to be created to utilize the unused trunk link. The root selection plays an important role in the speed of tree convergence. In our example, the second root bridge can be either B or C but not A, in order to utilize the trunk-link between B and C. Hence, our second ST is shown in dotted lines and utilizes trunk links 2 and 3.

4.4.1 Basic Network Model Simulation

This section will discuss the different MST algorithms used in our simulation. A common assumption is that we use two service provider VLAN pools (S-VLAN), one for each ST instance. VLAN pool-1 (S-VLAN IDs of 1 to 2000) is associated with ST-1, and VLAN pool-2 (S-VLAN IDs of 2001 to 4094) is associated with ST-2.

4.4.1.1 Option-1: Round Robin

For each EVC request, the ‘Round Robin’ algorithm assigns an S-VLAN ID from one of the S-VLAN pools based on alternating the ST assignments sequence of equal weights. As shown in Table 5. EVC-1 is assigned the S-VLAN ID value of 20 to make sure it gets mapped to ST-1, and EVC-2 is assigned the S-VLAN ID value of 2001 to make sure it gets mapped to ST-2. In this example, the round robin algorithm is alternating between ST instances based on whether the EVC number is odd or even.

Table 5: Example of ST assignment based on Round Robin

CE-VLAN	Src Node 'A'	Dest. Node 'Z'	CoS	BW Mbps	EVC ID	ST #	S-VLAN ID
10	B	C	NA	80	1	ST 1	Pool-1 20
100	A	C	NA	36	2	ST 2	Pool-2 2001

4.4.1.2 Option-2: Class of Service.

For each EVC request, the 'CoS' algorithm assigns an S-VLAN ID from one of the two S-VLAN pools based on the CoS associated with the EVC. The two S-VLAN pools have been pre-configured for service classes as follows: 'Gold' is assigned to ST-1, 'Silver' is assigned to ST-2, and 'Bronze' is assigned to either ST based on previously described Round Robin algorithm. See the example in Table 6.

Table 6: Example of ST assignment based on CoS

CE-VLAN	Src Node	Dest. Node	CoS	BW Mbps	EVC ID	ST #	S-VLAN ID
10	B	C	Bronze	80	1	ST 2	Pool-2 2001
100	A	C	Gold	36	2	ST 1	Pool-1 20

4.4.1.3 Option-3: Shortest Path

For each EVC request, the 'Shortest Path' algorithm assigns an S-VLAN ID from on the S-VLAN pools based on the shortest path (least number of hops) between source and destination nodes. The algorithm uses VLAN pool-1 for node pairs (A, B) and (A, C), and VLAN pool-2 for node pair (B, C). The algorithm applied is based on determining the shortest path between the source node and the destination node, regardless of the Class of Service or bandwidth requirements. See the example in Table 7.

Table 7: Example of ST assignment based on Shortest Path

CE-VLAN	Src Node	Dest. Node	CoS	BW Mbps	EVC ID	ST #	S-VLAN ID
10	B	C	Bronze	80	1	ST 2	Pool-2 2001
100	A	C	Gold	36	2	ST 1	Pool-1 20
200	A	C	Silver	26	3	ST 1	Pool-1 21

4.4.1.4 Option-4: Shortest Path + CoS

The ‘Shortest Path + CoS’ algorithm is equivalent to the ‘Shortest Path’ algorithm with the addition of the class of service rules. This algorithm is also referred to as ‘CoS TE’ to highlight its traffic engineering (TE) rules. ACES will check whether the requested bandwidth for a given service class is available on the shortest path between the source and destination nodes. If it is not available, then the algorithm considers an alternative path. The ‘Shortest Path + CoS’ algorithm provides the best network efficiency, i.e. enabling the greatest number of EVCs to be assigned to the network before any capacity augmentation of the links needs to occur. See the example in Table 8.

Table 8: Example of ST assignment based on Shortest Path and CoS

CE-VLAN	Src Node 'A'	Dest. Node 'Z'	CoS	BW Mbps	EVC ID	ST #	S-VLAN ID
10	B	C	Bronze	80	1	ST 2	Pool-2 2001
100	A	C	Gold	36	2	ST 1	Pool-1 15
200	A	C	Silver	26	3	ST 1	Pool-1 1025

4.4.2 Performance Model

Each of the trunk links in the performance model (i.e., network shown in Figure 30) was configured for 1 Gbps. The bandwidth allocations per CoS were assigned on each trunk link as follows: 10% Gold, 20% Silver and 70% Bronze. Up to 1000 EVC service requests were generated. In this example, each EVC bandwidth request varied between 1Mbps and 100 Mbps. The simulation was run and the information was then collected 50 times. Table 9 shows an example of a sample run. When an EVC request is not processed by an algorithm due to lack of network resources, it is indicated by a 'No' and is highlighted in the table.

Table 9: Example of an EVC Request and when its associated algorithm is blocked

EVC ID	*Source Node*	*Dest. Node*	*CoS*	*Bandwidth*	*STP*	*R/R*	*CoS*	*SP*	*CoS-TE +*
1	A	C	Bronze	18	Yes	Yes	Yes	Yes	Yes
. . . . .	. . . . .	. . . .	. . . . .	. . . . .	Yes	Yes	Yes	Yes	Yes
232	B	C	Gold	34	No	Yes	Yes	Yes	Yes
293	A	B	Bronze	88	No	Yes	No	Yes	Yes
296	A	B	Bronze	78	No	No	No	Yes	Yes
522	B	C	Gold	22	No	No	No	No	Yes
558	B	C	Bronze	98	No	No	No	No	No
1000	A	B	Gold	4	No	No	No	No	No

Once a simulation run is complete, the results and all relevant data are collected and stored. This process is repeated for as many times as needed. In this simulation, this process was repeated 50 times and its data was collected, stored and statistically analyzed and plotted.

4.4.3 Statistical Analysis

Network efficiency, E, is defined and calculated (in percentage) as the total bandwidth provisioned in the network, B, divided by the network Capacity, C.

$$E(\%) = B / C$$

The total bandwidth provisioned for all EVC requests in the network is defined as:

$$B = \sum_{i=1}^{n} EVC_i$$

Where n is the number of EVCs and EVC_i is the bandwidth assigned for the i^{th} EVC per algorithm used. The network capacity is given by:

$$C = \sum_{l=1}^{n} C_l ,$$

Where C_l is the trunk link capacity for link l, and n is the number of trunk links.

Table 10: Network Efficiency Simulation per algorithm

Case	**STP**	**R/R**	**CoS**	**SP**	**TE+CoS**
1	39%	38%	45%	73%	73%
2	28%	41%	28%	90%	90%
.	31%	37%	42%	87%	87%
.	32%	44%	32%	67%	67%
50	32%	43%	42%	68%	84%

The five algorithm options are then tabulated and compared as shown above in Table 10. The results for the 50 simulation runs were then sorted from minimum to maximum efficiency, using STP as the baseline. The sorted data were then plotted to illustrate the

Network Efficiency achieved per algorithm (see Figure 31 left hand side). Also, a second figure is included to illustrate the delta efficiency improvement over spanning tree (see Figure 31 right hand side). The delta graph demonstrates that CoS+TE algorithm is almost 2 to 1 factor improvement over STP (~100 % improvement).

The increase in network efficiency realized by ACES CoS+TE algorithm translates to a longer lifetime and lower cost for a given network, as measured by the number of EVCs provisioned.

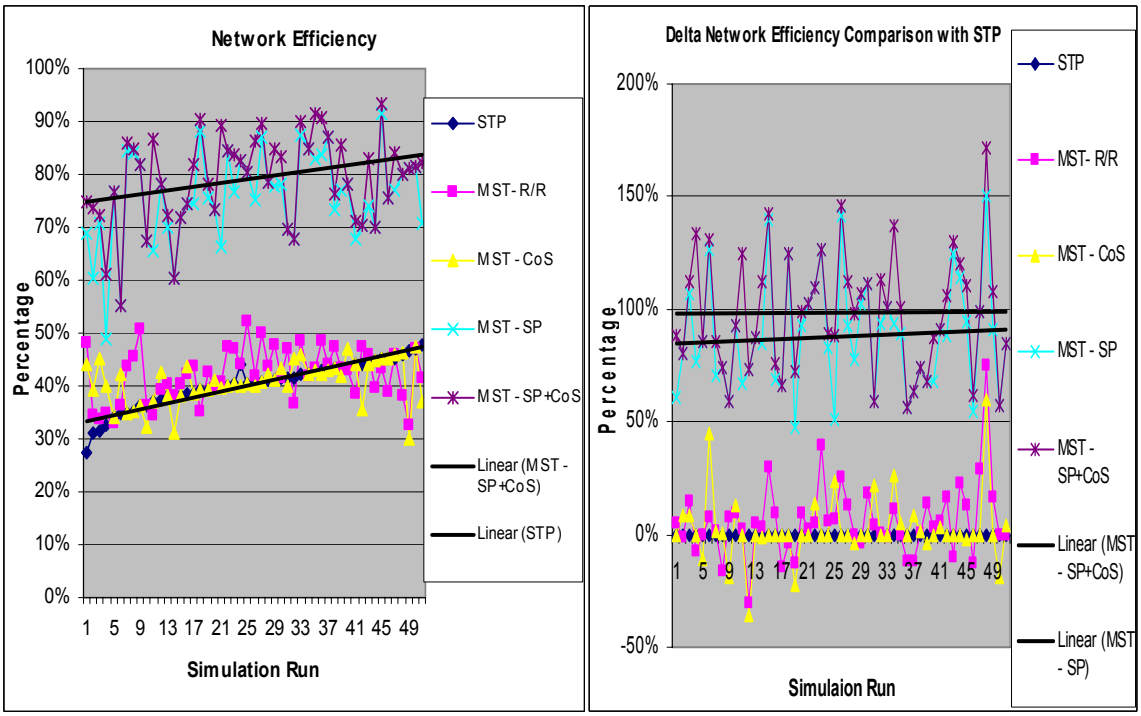

Figure 31: Network Efficiency per algorithm compared to STP as the baseline

Figure 32 shows the simulation results of each algorithm with its number of EVCs provisioned based on the basic network model.

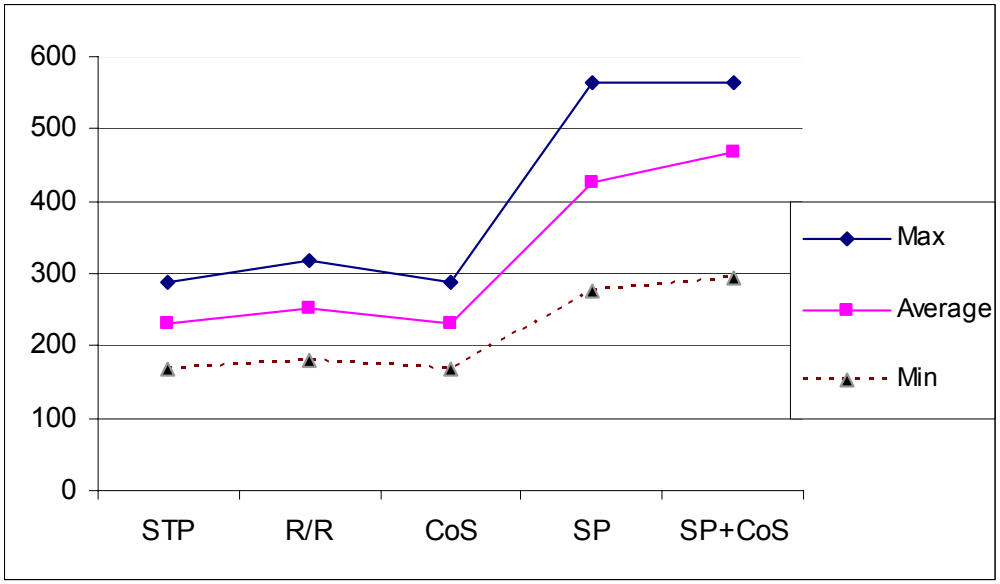

Figure 32: Number of EVCs provisioned per algorithm

ACES 'Shortest Path + CoS' or 'CoS-TE' algorithm demonstrated on average an improvement of approximately 240 EVCs over STP. This improvement could translate to delaying capacity upgrades to a medium-sized metro network by 1 to 2 years.

In the next section, we enhance our basic model to a typical service provider metro area network, which will provide better scalability and reliability.

4.5 Typical Metro Area Network Model

A typical Service Provider network model is shown in Figure 33. This model enhances the basic network model by scaling its network size and improving its network reliability. Also, the trunk-link bandwidth capacity between nodes is increased with the use of link aggregation groups.

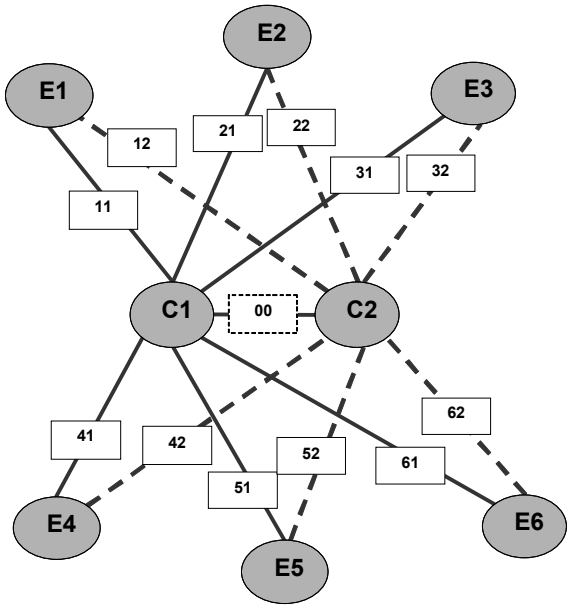

Figure 33: A typical metro Ethernet network: Dual Hub Model

The dual hub model maintains ACES capability in accepting or rejecting the EVC provisioning in a given network. In addition, it expands its functionality to include the signaling reservation of a link aggregation group, or to interwork with the optical transport network elements. This interworking may also include an interface into a separate Provisioning Server as shown in Figure 34, in the case of larger scale networks. ACES may also request from the OTN Optical Control System to add a λ between switches or communicate directly with the Optical Network Element.

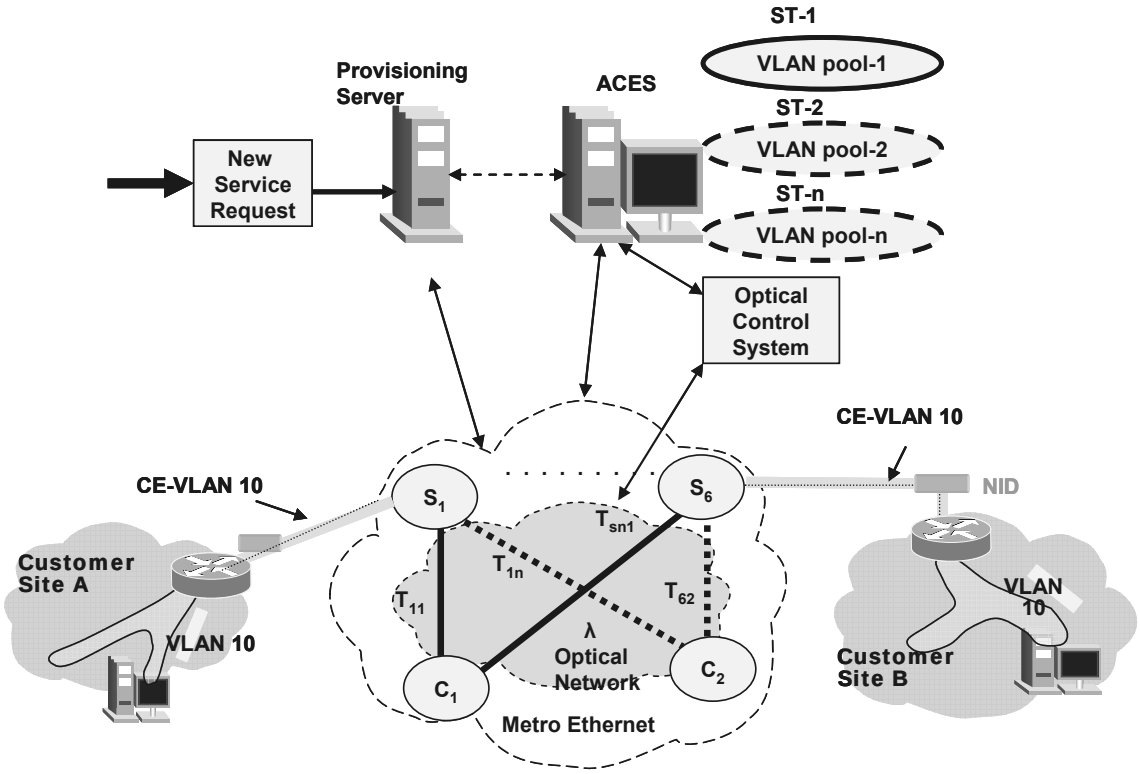

Figure 34: ACES Architecture Diagram with OTN signaling and management layer

4.6 Metro Area Performance Model

Let us consider a typical service provider Metro Ethernet Network (MEN) with n switches ($S_1 \ldots S_n$), in our example, we use n = 6 (see Figure 34). Each switch has two trunks dual homed to two core switches C1 and C2. The trunks are dedicated fiber or optical trunks switched via a λ-based optical transport network (OTN) and are connected to an optical cross connect or multiplexer (up to 8 λs). Let us assume that all trunk links have the same capacity. Each trunk will be represented by the numbers of the two nodes it is connected to: the switch number and the core switch number. For example, T_{11} represent a trunk from switch S_1 to core switch C_1, T_{12} represent trunk link from switch S_1 to switch C_2, and T_{61} represent trunk from switch S_6 to core switch C_1. Let us designate switch C_1 as the root-bridge. Switches S_1 thru S_6 and C_2 will then build their

forwarding tables back to the root for ST1. Switch S_1 will have a path to the root C_1 via T_{11}; Switch S_2 will have a path to root C_1 via T_{12} and so on. However, all trunk links connected to C_2 will be unused since switches will block their trunk interfaces to C_2. A single spanning tree is now formed and data can flow between the switches with no loop in the network. Multiple spanning tree switching provides the ability to utilize more trunk links in a network where those trunk links were considered blocked based on basic spanning tree protocol. In our example of Figure 33, a service provider can overlay their physical topology with logical topologies based on MSTs and assign Ethernet Virtual Circuits to a given MST, as needed.

A second spanning tree needs to be created to utilize the unused trunk links, see Figure 34. In our example, the second root-bridge is C_2. Hence, our second ST is shown in dotted lines and utilizes trunk links T_{12} thru T_{62}.

As discussed in section 4.3, the ACES Algorithms are used and we introduce an additional algorithm that looks at utilizing C1 and C2 trunks. In addition, two algorithms are not used: ‘CoS’, since it is no longer interesting; and ‘SP’, since it is no longer applicable as both ST instances have the same number of hops. The algorithms used in this model are summarized below:

1. **Spanning Tree:** This algorithm is used as a reference model
2. **Round Robin:** This algorithm is used to assign the VLAN ID in the service provider network by toggling amongst the multiple spanning tree instances
3. **Classes of Services – Traffic Engineering (CoS-TE)**: This algorithm determines if the required bandwidth is available is in different MSTs paths and route the EVC based on available bandwidth per given service class. The

algorithm selects the ST instance with the most available bandwidth on the 'worst-case' link

4. **CoS-TE+:** This option looks at packing the spanning trees differently.

Note: Other spanning trees designed to utilize C1-C2 trunks for added reliability did not improve the network efficiency by more than 1%. For example, if a request from S1 to S6 needs to be fulfilled and trunk T_{11} and T_{62} have no capacity available, but the path (T_{12} to C2, C2 to C1, and C1 to S6) has the available bandwidth, then the request will be granted.

4.6.1 Performance Simulation

The simulation model assumes a network as shown in Figure 33, with 10 Gbps capacity per trunk link and a trunk utilization factor not to exceed 73% subscribed bandwidth on any given trunk. Also, we maintain the same assumptions used in the baseline network model in section 4.4.1. However, minor variations on the assumptions were made that include the bandwidth allocation per CoS as follows:

- Service Request for each EVC consists of three randomly generated attributes:
    a. Node pair
    b. CoS assigned to the EVC
    c. Bandwidth associated with CoS
- Service Classes with modified bandwidth ranges for tighter packing of the pipe
    a. 10% Gold, with an EVC request ranging from 1Mbps to 50Mbps
    b. 20% Silver with an EVC request ranging from 2 to 100Mbps
    c. 70% Bronze with an EVC requests ranging from 4 to 200Mbps

- The simulation model generates up to 1000 EVC requests

As discussed previously in the basic network model, once a service request is not met for any given algorithm, the simulation will capture the relevant EVC number and the network utilization data for analysis. In addition, once all 1000 EVCs are processed, the simulation run is completed, then the results are stored and the simulation process repeats itself until it reaches the desired sample runs. The results are then statistically analyzed and plotted as shown in Figure 35.

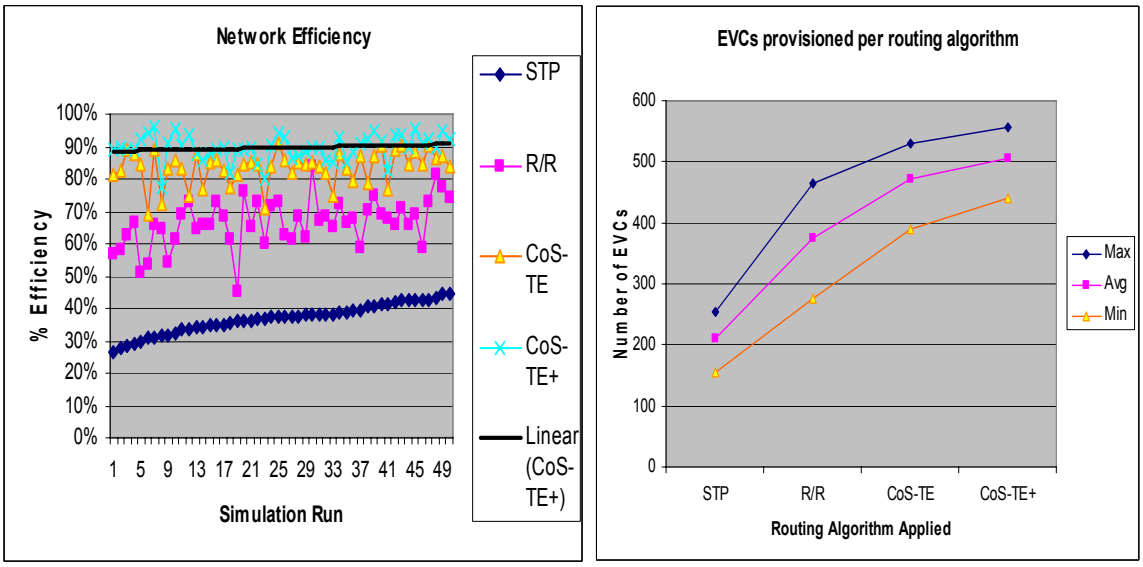

Figure 35: Typical MEN Network Efficiency and EVCs provisioned per algorithm

As illustrated, a service provider will be able to offer a reliable CoS in the Metro and Wide Area Networks utilizing ACES algorithms. Figure 35 clearly shows that the network efficiency percentages achieved in this model are higher than our previous example and on average reached above 90%.

This model proved to be more efficient, reliable and scaled better than the basic model. Also, in this model, ACES will signal or request from the Optical Control System

additional bandwidth for a given path between two switches as needed. This will result in a better utilization of network resources because the congested links are upgraded on demand, and therefore it increases the life of the network.

The results achieved, shown in Figure 35, are consistent with the results discussed in section 4.4.3. The main reason in obtaining similar range of the number of EVC provisioned is that in our current example we allowed an EVC request for bronze class of service to vary between 4 and 200 Mbps, which is double the allowed range in our previous simulation in section 4.4.3. Also, an EVC request for gold class of service varied only between 1 and 50Mbps, which is half of our previous example range. The silver class of service remained the same. Taking into consideration that bronze requests account for about 70% of all requests made and the maximum percent of link utilization used in this example is 73% versus our previous example of 100%. Based on the previous example assumptions, we expect an estimated number of provisioned EVCs to yield a number in the range between 850 and 930 EVCs.

4.7 Enhanced Model

This model is further enhanced to utilize an optical transport network (OTN) where the trunk link capacity utilizes light path and dynamic queue weighting algorithms.

The enhanced model improves the scalability and QoS efficiency in the network. The scalability is addressed by increasing the number of nodes homed in to the dual core switched or by expanding the number of core nodes to three or 4. However, the placement and the mesh requirement become inefficient as distances between core nodes and switches become larger. Hence, the use of the OTN improves network scalability by increasing the trunk-link bandwidth on demand, reducing delay and jitter, and increasing

the scalability and reach of the electrical transport links, operating at up to 100 lambdas per trunk link.

In addition, the enhanced model supports bandwidth on demand requests made to the admission control and is capable of predicting the required bandwidth to ensure the appropriate bandwidth is allocated, or requests the optical network to provision new bandwidth channel to meet the BoD request.

Therefore, ACES forthcoming improvements require a drastic change to the network infrastructure to include the ability to dynamically route assigned EVC paths based on higher CoS requirements and real time feedback received from active probes within the network. The enhancements will allow ACES to identify the trunk links that have bandwidth issues and further utilize the network resources by requesting the needed bandwidth between switches over the optical λ switch network, thus maximizing the number of EVCs provisioned and improving the overall network efficiency. In addition, the enhanced model clearly needs to achieve network scalability with significant efficiency of its network resources utilization based on a new forwarding L2 algorithm other than ST and MST based algorithms. In order to achieve this, we need to integrate L1/L2 functionality and control mechanisms into a single network element. This network element will have all of layer-2 functionality on the UNI side and all of the layer-1 optical transport functionality on the NNI side with an intelligent plane combining L1 and L2 management and control planes.

This model utilizes a new queuing algorithm which dynamically adjusts the queue weights on a given port of a given trunk. The next section describes this algorithm in greater details.

4.7.1 Queue Scheduler

Each trunk link has three queues, one strict priority and two Weighted Round Robin (WRR) queues. Gold traffic is mapped to the strict priority (SP) queue, ensuring low loss, jitter and delay. Silver and bronze traffic are mapped to separate WRR queues that have less stringent service level guarantees, see Figure 36. For a given trunk link L_{ij}, B_{PG} represents the provisioned gold bandwidth and B_{AL} represents the allocated bandwidth for L_{ij}. Silver and bronze traffic are served by the WRR scheduler and share the remainder of the trunk link bandwidth, B_{AW}, which is equals to $B_{AL} - B_{PG}$, where B_{PG} is the bandwidth provisioned for gold.

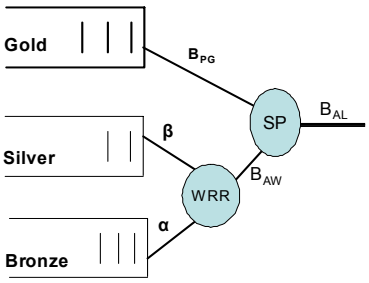

Figure 36: Queue Scheduler

In Figure 36, alpha (α) and beta (β) represent the weightings assigned to the bronze and silver queues, respectively. Initially, α is set to 1 and β is set to 4.3, based on the need for differentiated SLAs and the forecasted demand.

4.7.2 Dynamic Queue Weighting Algorithm

Dynamic queue weighting (DQW) is a new algorithm (see Figure 37) developed to modify the Weighted Round Robin (WRR) weights on the silver and bronze queues to

optimize the utilization of the network resources based on static or dynamic feedback from the network [10].

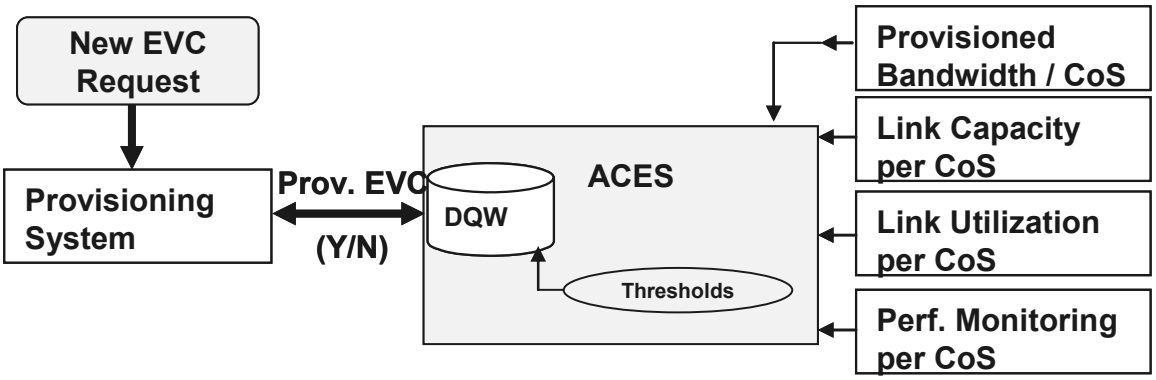

Figure 37: ACES Dynamic Queue Weighting Algorithm Architecture

DQW maintains the original bandwidth allocations specified per service class. A threshold is also configured to determine when the DQW will kick in. Once the threshold is crossed, DQW begins to adjust α and β from their initial settings as each EVC is provisioned in that trunk link. If the available bandwidth on the trunk link is exhausted, then ACES will signal the Optical Controller to request additional transport bandwidth for that trunk link. Figure 38 provides DQW high level algorithm.

Dynamic over subscription weighting factors adjustment (DOWFA) is also a new algorithm developed to modify the service class oversubscription factors. The DOWFA algorithm assigns the full trunk link bandwidth to 'Gold' service class and the remainder is allocated between 'Bronze' and 'Silver' classes. As in the DQW algorithm, once the initial provisioning threshold is met on a given trunk link, then the bronze and silver weightings α and β are adjusted dynamically as EVCs are assigned based on the provisioned bandwidth per service class. ACES then updates the 'available bandwidth' for the 'Silver' and 'Bronze' classes, in real time accordingly. However, the key

difference between DOWFA and DQW is that DOWFA will kick in once a second threshold is reached and determine whether to adjust the CoS oversubscription values.

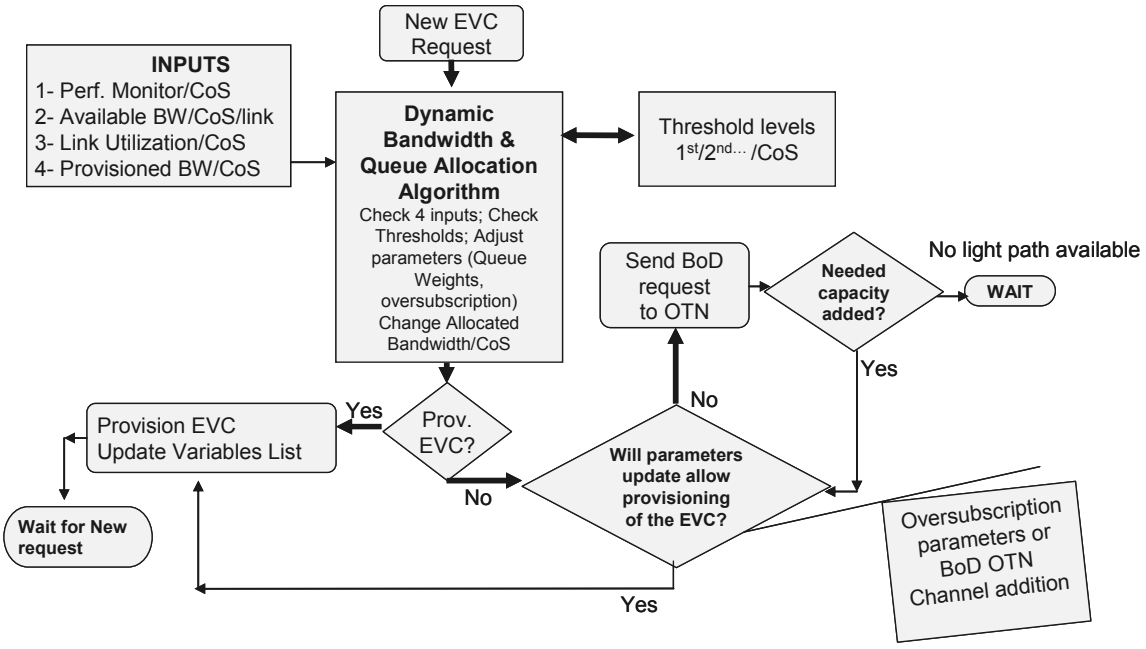

Figure 38: Dynamic Queue Weighting Algorithm

4.7.3 Dynamic Over-subscription Weighting Factors Adjustment

ACES may modify the over-subscription factor of either bronze or silver service classes based on analysis of the trunk link CoS utilization and/or network performance measurements. Once a provisioning threshold is met on a given trunk link, the bronze and silver weightings α and β are adjusted dynamically as EVCs are assigned, based on the provisioned bandwidth per service class. ACES then updates the 'available bandwidth' for the Silver and Bronze classes, in real time accordingly.

The simulation model assumes a network as shown in Figure 33, with a trunk link capacity that may vary between 1 and 10 Gbps. In our simulation, we used a 10Gbps link with ~73% bandwidth allocation, B_{AL}=7325 Mbps. B_{AL} is the physical trunk link capacity

if no over-subscription is used, or the aggregate of the allocated bandwidth per oversubscribed class of service. Note: In the case of link aggregation, ACES applies a load-balancing factor that corrects for sub-optimal load balancing across the links.

A service request for an EVC consists of three randomly generated variables: Node pair, CoS assigned to the EVC and bandwidth associated with the CoS. The node pair is assigned randomly between different edge switches. CoS is assigned randomly based on the following distribution: Gold (15%), Silver (25%) and Bronze (60%). CoS bandwidth is assigned randomly within the following ranges: Gold (1 to 50 Mbps), Silver (2 to 100 Mbps) and Bronze (4 to 200 Mbps). Note: the above ranges used for our simulation can be easily modified. In our simulation example, ACES algorithms have been modified to take into account the network model and the enhancements made with the introduction of DQW and DOWFA (see Table 11).

Table 11: ACES Enhanced Model Algorithms

Solution	# STs	EVC Routing Algorithm	Link BW Allocation	Comments
ST	1	All EVCs => ST1	Static for each CoS: Total link BW = 7325 Mbps, allocated as follows: G (425), S (1300), B (5600)	Not all of the network links are utilized
R/R	2	Based on EVC number => Odd => ST1, Even => ST2	Static for each CoS: Total link BW = 7325 Mbps, allocated as follows: G (425), S (1300), B (5600)	The EVC routing algorithm ensures utilization of all unused links
CoS-TE	2	Check EVC CoS B/W: if smallest trunk link {Ti,Tj} of ST1 > smallest trunk link {Ti,Tj} of ST2 => ST1; else => ST2	Static for each CoS: Total link BW = 7325 Mbps, allocated as follows: G (425), S (1300), B (5600)	The EVC routing algorithm achieves load balancing across ST-1 and ST-2.
CoS-DQW	2	Check EVC CoS BW: if smallest trunk link {Ti,Tj} of ST1 > smallest trunk link {Ti,Tj} of ST2 => ST1; else => ST2 (same as CoS-TE)	Static for each CoS, but reflects actual queue scheduling, i.e., Gold is Strict Priority (never gets starved) and S, B split the difference (7325 - Gold BW) based on a fixed ratio (B/S = 4.31) consistent with forecasted demand per CoS	The EVC algorithm is the same as CoS TE but it allows 'Gold' to have full utilization of the trunk-link B/W and the remainder of the B/W is assigned to 'Silver' and 'Bronze' using a fixed ratio.
CoS-DOWFA	2	Check EVC CoS BW: if smallest trunk link {Ti,Tj} of ST1 > smallest trunk link {Ti,Tj} of ST2 => ST1; else => ST2 (same as CoS-TE)	Dynamic for each CoS - reflects actual queue scheduling, i.e., Gold is Strict Priority (never gets starved). Silver and Bronze split the difference (7325 - Gold B/W) based on a dynamically calculated ratio of B/S on a per-link basis - calculated ratio starts with (4.31) based on forecasted demand, but learns actual provisioning per link over time and adjusts the link weightings to balance the available bandwidth each time an EVC is provisioned	This algorithm learns as we provision EVCs and dynamically changes the WRR weightings for Silver and Bronze. This algorithm achieves the maximum Network efficiency.

A simulation run generates up to 1000 EVC requests into the network. Once a service request cannot be met due to lack of resources, ACES will then identify the EVC ID, the trunk link utilization per CoS, and the algorithm used. The CoS-DQW, will adjust the WRR queue weighting to further improve the network utilization. Also, the CoS-DOWFA will adjust the WRR queues and modify the oversubscription weights based on actual network utilization. Our analysis tool calculates the network efficiency (Eff %) in

percentage as B_P/B_A, where B_P is the aggregate EVC bandwidth provisioned, and B_A is the aggregate trunk link bandwidth allocated in the network between edge switches. It also stores the results for statistical analysis and plotting. A summary of the simulation results is shown in Table 12 for the four algorithms.

Table 12: Summary results of the EVC routing algorithms

Solution	*EVC Routing Algorithm*	*Link CoS BW Allocation*	*Network Efficiency*		
			Min	*Avg*	*Max*
ST	All => ST1	Static	27%	37%	47%
R/R	Odd => ST1, Even => ST2	Static	45%	66%	84%
CoS-TE	CoS Aware EVC routing	Static	69%	83%	90%
CoS-DQW	CoS Aware EVC routing	DQW	71%	85%	93%
CoS-DOWFA	CoS Aware EVC routing	DQW	77%	90%	96%

In Table 12, we also present the min, max and average network efficiency across the 50 simulation runs, for each of the EVC routing algorithms used. As shown in Table 12, the newly introduced enhancements demonstrate an improvement in the network utilization with CoS-DQW on average two percent (2%) and with DOWFA on average of six percent (6%), (see Figure 39).

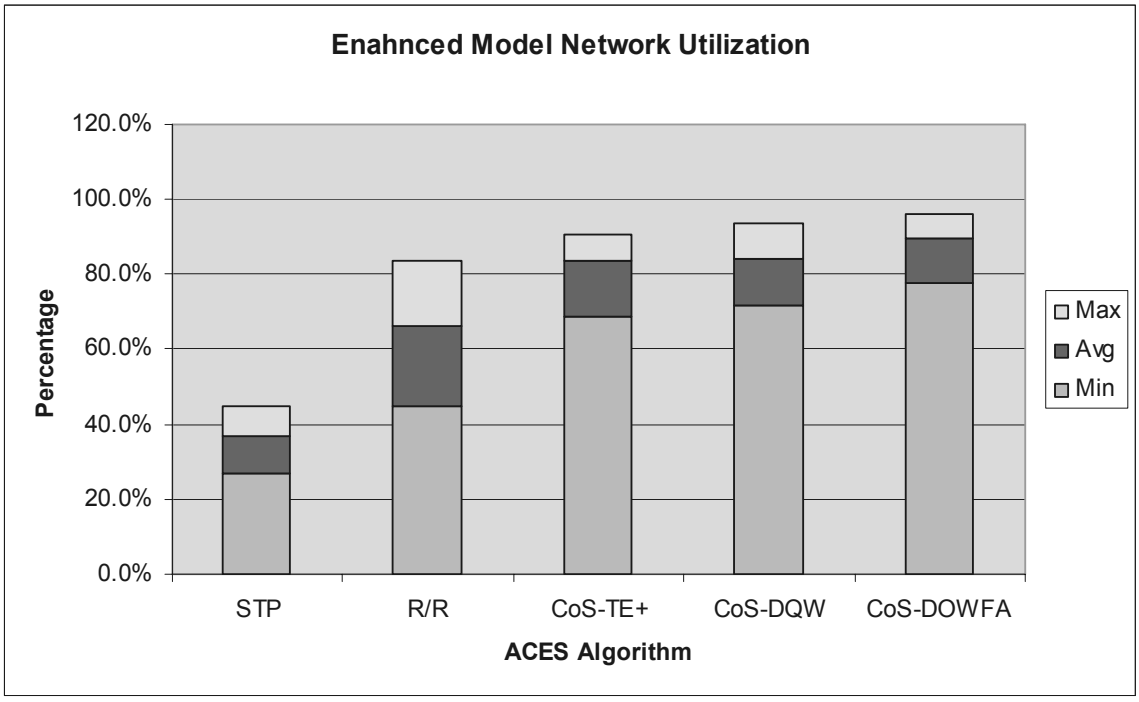

Figure 39: Enhanced Model Network Simulation results

Figure 40 below shows the results of each of the 50 simulation runs, normalized for the baseline single Spanning Tree solution. The CoS algorithms achieve the highest efficiency, with little variability around the average, clearly demonstrating that they can achieve significant improvement in utilization of the network resources over ST and R/R (in this example, this is equivalent to shortest path algorithm) based algorithms. Therefore, by combining CoS TE-based EVC routing, with the dynamic adjustment of the WRR queue weights, and the dynamic adjustment of the oversubscription factors algorithms, results in the most efficient approach to improving the network utilization (see Figure 40).

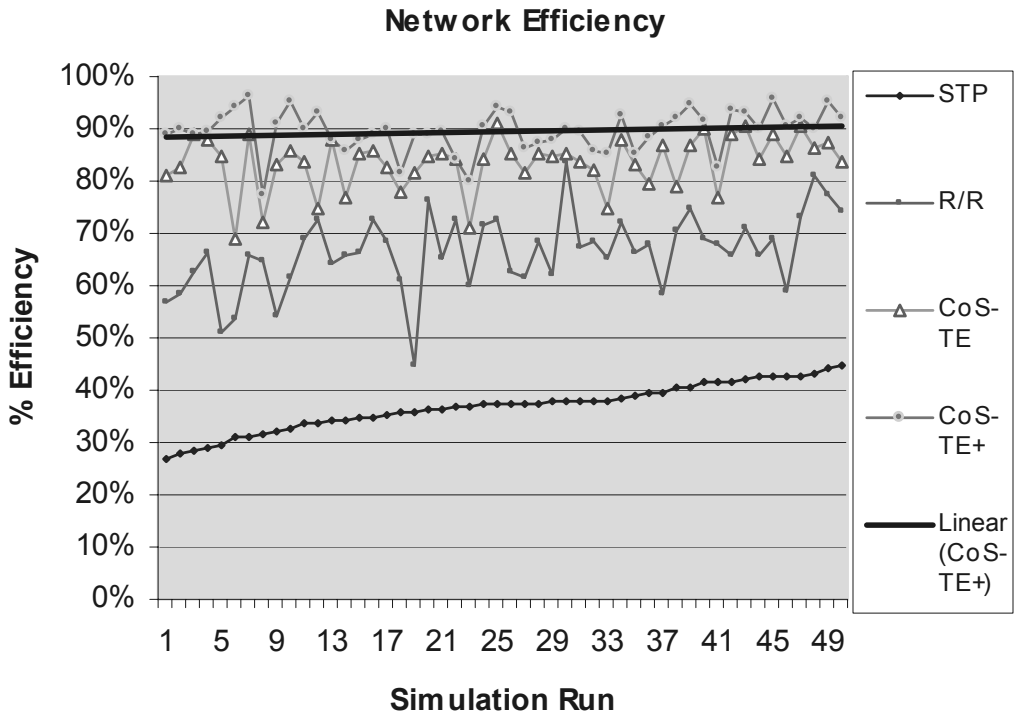

Figure 40: ACES Enhanced Model Network Efficiency

We can draw a similar conclusion when looking at the number of EVCs that can be provisioned before running out of bandwidth resources for each of the algorithms. In our simulation runs, the average number of EVCs that could be provisioned ranged from 210 for STP to 505 for CoS-TE+.

With the introduction of ACES, a service provider will be able to efficiently route EVCs within a switched network while optimizing the utilization of network resources. Real-time feedback from the network can be fed to ACES to further enhance its value. ACES can also be extended to signal a request to the Optical Control System to add bandwidth for a given path between two switches, when required. This idea is explored in more detail in the next section.

4.8 Network Model Scalability

If we assume a tight coverage of metro area, then the dual hub model can be extended to cover additional nodes of at least 10 edge switches. If the number of core switches increases to 4, then we can quadruple the number of edge switches and this will scale up to 40 edge switches. This model should be simulated to validate how well the model will scale and perform.

In a meshed squared network, assume we have m represents the depth of the mesh network, or m^2 represents the number of nodes in the square mesh network, and x number of link, then our formula becomes:

$$X = ((m-2)^2*4 + (m-2)*4*3+ (4*2))/2$$

Whereas, a spanning tree network will only utilize *n-1* links in an *n* node network. As shown in Figure 33, we can easily expand the model to include 10 edge switches or nodes at the top and 10 nodes at the bottom for a total of 20 nodes. This can be further scaled by introducing the two additional core switches for a total of 4, which will form a squared mesh network. With each side of the squared mesh, we can have up to 20 edge switches for a total of 80 edge switches. With this in mind, we can further expand the model as we increase the number of core switches.

Table 13 illustrates the link utilization of a mesh network as compared to the number of links utilized when running STP.

Table 13: Link utilization in a mesh network versus spanning tree

Link Utilization (Mesh vs. STP)					
Length	# Nodes	# links	# links used in STP	% links used	% links not used
	3	3	2	66.7%	33.3%
2	4	4	3	75.0%	25.0%
4	16	24	15	62.5%	37.5%
6	36	60	35	58.3%	41.7%
8	64	112	63	56.3%	43.8%
10	100	180	99	55.0%	45.0%
15	225	420	224	53.3%	46.7%
20	400	760	399	52.5%	47.5%
25	625	1200	624	52.0%	48.0%
30	900	1740	899	51.7%	48.3%
40	1600	3120	1599	51.3%	48.8%
50	2500	4900	2499	51.0%	49.0%
60	3600	7080	3599	50.8%	49.2%
70	4900	9660	4899	50.7%	49.3%
80	6400	12640	6399	50.6%	49.4%
90	8100	16020	8099	50.6%	49.4%
100	10000	19800	9999	50.5%	49.5%

The next sections cover: overlay model and provide simulation results; and bandwidth on demand over optical network proof of concept and share the future model requirements. The main objective of these sections is to highlight the need for integrated L1/L2 control and management planes and the need to address any Ethernet integration issues as a result.

4.9 Overview of Overlay Model over Optical Layer

The two-layer model aims at providing a tighter integration between layer-2 and layer-1 (optical layer), offers a series of important advantages over the current multi-layer

architecture model. To examine the architectural alternatives for the two-layer model, it is important to distinguish between the data plane and control planes over the user-network interface (UNI). The Ethernet-over-optical-network architecture is classified according to the organization of the control plane, i.e., whether there is a single integrated or separate independent monolithic routing and signaling protocol spanning the two layers.

Several industrial organizations including the Optical Interworking Forum (OIF) and the Internet Engineering Task Force (IETF) have already proposed several architectural options on how GigE switches/IP routers must interact with the optical layer to achieve end-to-end connectivity, including overlay, augmented, and peer-to-peer models (interconnection models) [111], [112]. The simplest is to treat the optical layer as completely separate from the client layer (layer 2/layer 3).

Under the overlay model, layer-2/layer-2 domain is more or less independent of the optical domain. Layer-2/layer-3 routing and signaling protocols are independent of the routing and signaling protocols of the optical layer. In this "overlay" model, the optical layer provides point-to-point connections (lightpaths) to layer-2/layer-3 domain. The client routers request high-bandwidth connections from the optical network, via some User-to-Network Interface (UNI), and are provided with no knowledge of the optical network topology or resources. The resources managed by the optical layer include wavelengths and fibers on physical links. The client layer manages bandwidth resources on lightpaths and route traffic (label-switched paths (LSPs)) over the logical topology treating the lightpaths as links. A more sophisticated model that offers a tighter integration between IP and optical layers (peer model) collapses the two layers into a single integrated layer managed and traffic engineered in a unified manner.

The overlay model is the most practical for near-term deployment because it is appropriate for the current telecommunications infrastructure that consists of multiple administrative domains, where there is clearly a need to maintain topology and control isolation between the optical transport and the service layers since they are most likely to be under different administrative controls and policies. However, the simplicity of the overlay model comes at the expense of the inefficient use of network resources due to information hiding at the domain boundaries.

4.9.1 Overlay Model Simulation

As discussed in chapter 4, a service provider can offer BoD services over our proposed model (i.e. integrating Ethernet layer 2 switching and OTN.layer 1 transport network). Simulation results obtained from our simulation model (see Figure 41) indicate that shortest path is still a good selection. Ethernet switches (GigE) are used to aggregate and switch EVCs, whereas, OTN (OXC) are used to provide the light-path between Ethernet Switches. Our simulation mode assumed the followings:

- A network diagram of 16 nodes
- EVC request ranges between 1 and 200 Mbps with a mean of 100 Mbps
- 8 lambdas per OXC (4 transmit, 4 receive)
- 100 Gbps of traffic generated
- Spanning tree based on maximum nodal degree (up to 6), the rest are selected at random up to 16
- Shortest path, Round Robin and First Fit (available) algorithms were used

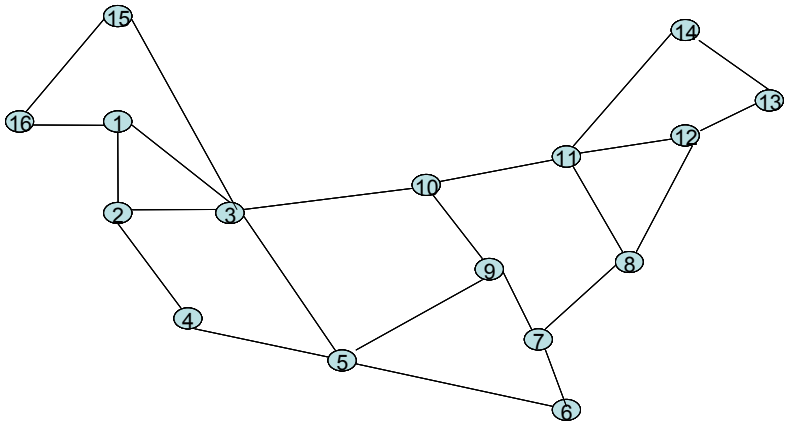

Figure 41: Overlay network topology

The model needs to be enhanced to include CoS and oversubscription traffic, however, the preliminary model looks promising (see Figure 42). Further work is still needed to validate the model to include cost and distance as attributes. .

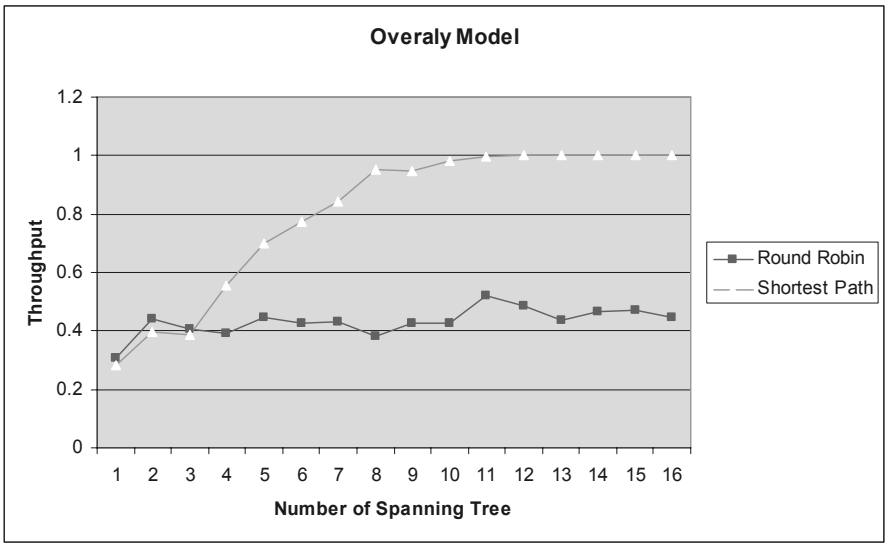

Figure 42: Hybrid model simulation results

The blocking probability is a metric that provides the quality of the network performance (i.e. how much throughput and loss a network is subject to). In our example, we plotted throughput instead. However, we can compute the blocking probability in an optical network as a reference as well as we can compute the blocking probability light path

(With wavelength converter or without). The blocking probability for an h-hop link would be:

$$P_b = [1 - (1-\rho)^n]^w \qquad \text{No Wavelength conversion is used}$$

The blocking probability for an h-hop link would be:

$$P_{b2} = [1 - (1-\rho)^w]^h \qquad \text{with Wavelength conversion}$$

4.9.2 Overlay Model (Incremental)

As shown in Figure 43 , the NSF network consists of 16 nodes with two fiber cables connecting nodes: one for transmit and one for receive. Each fiber link supports up to 8 wavelength channels. The simulation model assumes 1,000 EVC requests normally distributed around 200Mbps with a standard deviation of 50Mbps.

In our simulation model (see Figure 44, we observed that the incremental logic topology performs better than physical equals to logical at lower number of STs. The main reason is that each EVC request requires two light paths (source to root-bridge and then root-bridge to destination) for transmit and the same for the return path or receive. Therefore, for each request initially made, we assign four lightpaths, unless if either the source or destination nodes happen to be the root-bridge.

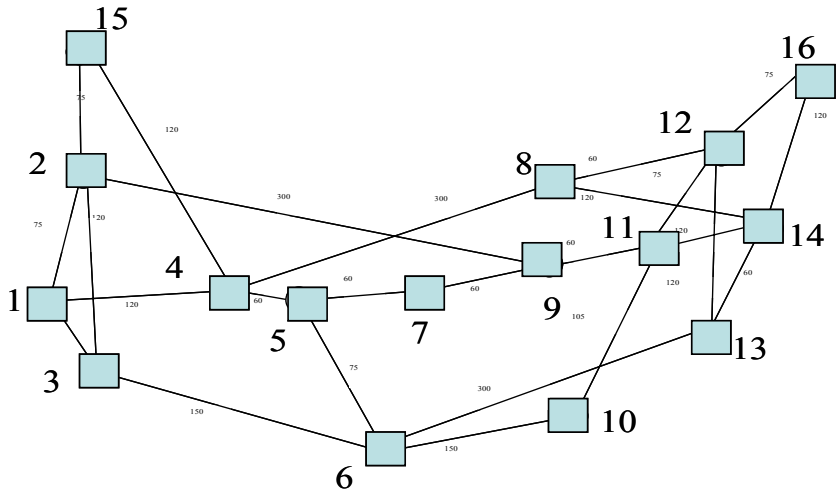

Figure 43: NSF network topology

Ethernet is full duplex requiring bi-directional connectivity and a symmetric logical topology. We simply conclude that the incremental logic topology forms a start network with the root-bridge at the center and thus behaves as shortest path between source and destination nodes. However, as more STs are implemented in the star topology model, it looses its advantage and the logical equal physical topology offers higher degree of efficiency. As observed in the simulation, shortest path achieves zero blocking as the number of spanning trees approaches 12 and above. This is because the number of spanning trees in the network approaches the number of nodes and thus the logical routing of shortest path. Also, our ability to add and drop at each node reduces the grooming issue. In addition, if we elect a root-bridge with the maximum nodal degree, then we are able to further enhance the network performance with fewer number of spanning trees.

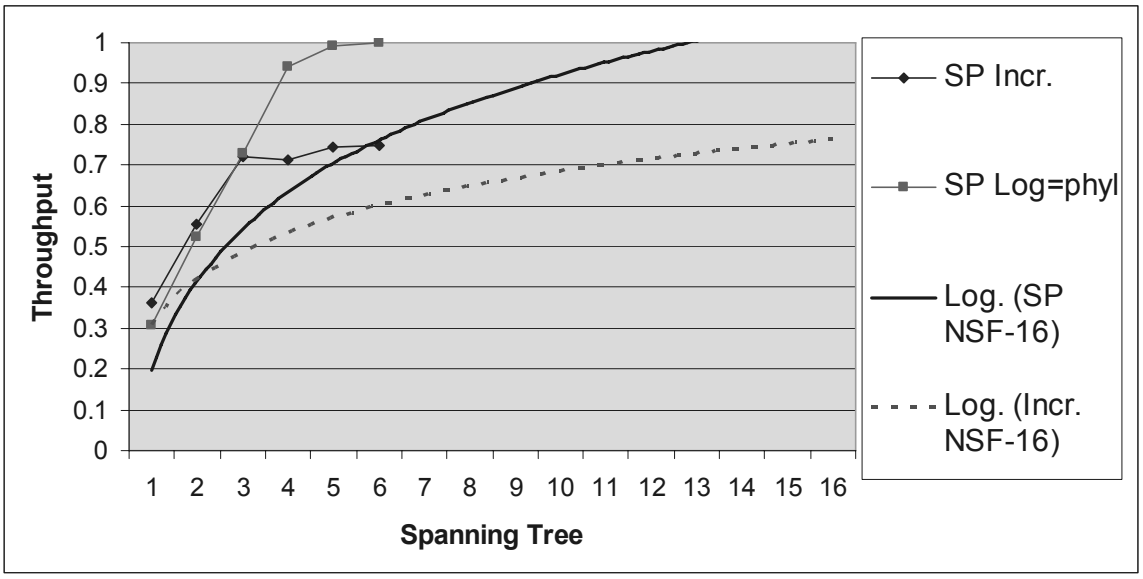

Figure 44: NSF 16 Shortest Path blocking probability

4.10 Bandwidth on Demand (BoD) OTN Prototype

This section discusses a prototype that investigates various OTN service concepts and the relationship of Ethernet services to the evolution of the Next Generation Optical Transport Network (NG-OTN). The trial examined proof-of-concept implementations of service architectures in support of advanced Bandwidth-on-Demand (BoD) service concepts leveraging NG-OTN control plane (CP) and next-generation network OSS capabilities.

4.10.1 BoD OTN Architecture Overview

The BoD OTN architecture (see Figure 45) represents an overall architecture of a typical NG-OTN, which consists of three functional planes: transport, control, and management.

A brief discussion of the capabilities and platforms deployed on each functional plane is as follows:

Transport Plane: Typical transport platforms considered for the NG-OTN are xWDM/NG-ADM in the access networks, Ethernet Switches in the metro network, OXC for regional core networks and ROADM/WXC for photonic core networks.

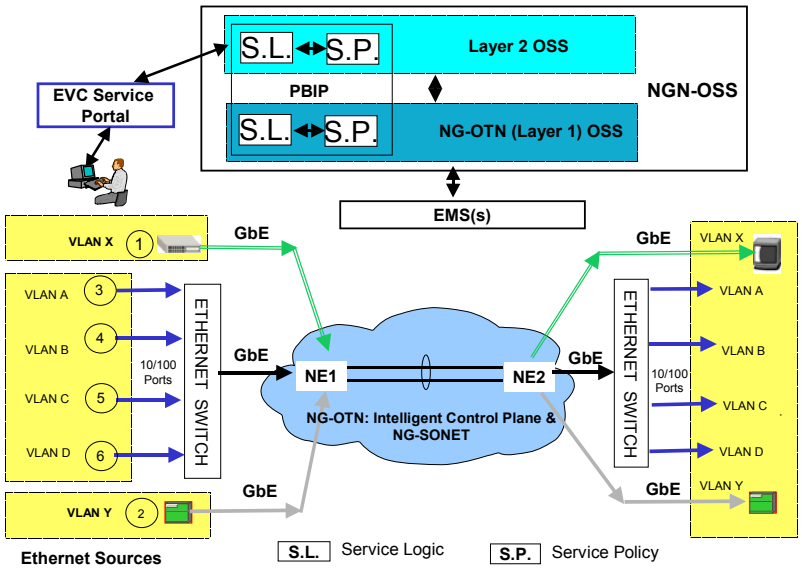

Figure 45: NG integrated L1/L2 OSS

The NG-SONET technology provides the following service adaptation and bandwidth management capabilities:

Service Adaptation via Generic Framing Procedure (GFP): A standardized mapping mechanism to adapt higher-layer (L2 and above) client signals for transport over SONET or OTN networks [99].

Virtual Concatenation (VCAT): Maximizes bandwidth utilization for flexible channel sizing over a SONET network via inverse multiplexing of SONET payload into multiple component STS-1s (High Order VCAT) or VT1.5s (Low Order VCAT) [100].

Link Capacity Adjustment Scheme (LCAS): Provide a mechanism to dynamically adjust the bandwidth of any VCAT'd path by adding or removing component paths in VCAT groups. The bandwidth adjustment is done in a hitless fashion [101].

The above NG-SONET features need only be implemented at the end points of SONET connections, thus allowing Service Providers to leverage their existing SONET core.

Control Plane: The capabilities of an OTN control plane include dynamic routing and distributed signaling [102]-[109]. These capabilities support the following functions:

- Auto-discovery and self-inventory
- Dynamic and fast service provisioning and activation
- Resilient service protection and restoration
- Integrated support of NG-SONET capabilities

A NG-OTN that is built upon integrated control plane and NG-SONET capabilities will achieve better resource utilization, will provide enhanced resilience, and will enable fast on-demand provisioning of L1 bandwidth, thus benefiting all L2 and L3 services that are provisioned over the L1 network.

Management Plane: Next-Generation Network Operations Support Systems (NGN-OSS) are required for managing the dynamic nature of the NG-OTN control plane and for supporting L2 services that are provisioned over NG-OTN network. In addition to the traditional OSS functions in Operations, Administration, Maintenance and Provisioning (OAM&P) and Fault, Configuration, Accounting, Performance, and Security Management (FCAPS), NGN-OSS needs to manage the following new capabilities:

- Control plane capabilities in the NG-OTN

- NG-SONET NEs and VCAT/LCAS activities
- Integrated L1 and L2 functions on single platform

Besides the NGN-OSS, the management plane needs to support the following two new service-related components for cross-layer service offerings:

- Service logic for each value-added OTN service along with a service logic execution environment; and
- A policy server for managing service policies that govern cross-layer provisioning to support enhanced services.

4.11 Next Generation Optical Network

The NG-OTN platforms and NGN-OSS components are summarized in Table 14. The first version of the trial employs NG-ADM platforms that provide CP-based OTN transport service (L1 provisioning) and VLAN-aware Ethernet switches for L2 traffic, aggregation, and grooming. The NGN-OSS component is usually a service provider developed system that supports the policy-based intelligent provisioning function. Other OAM&P functions such as surveillance and billing were not addressed in the lab trial.

Table 14: NG-OTN and NG-OSS components

Platform	*Functionality*
Next generation NG-ADM platforms and VLAN-aware Ethernet switches	Integrated L2 & L1 optical transport with intelligent control plane and NG-SONET capabilities (i.e., GFP, VCAT, LCAS)
Service provider NGN-OSS: Policy-Based Intelligent Provisioning (PBIP) system	L2 and L1 provisioning, activation, and equipment inventory for NG-OTN

As shown in Figure 45, a customer can initiate a service request that is then routed to the Policy-Based Intelligent Provisioning (PBIP) system where service parameters are screened and forwarded to the service logic. The execution of the logic will trigger a sequence of provisioning activities on L2 and/or L1 in order to fulfill the order. Depending on the service type, the service logic may consult with related service policies for provisioning decisions.

4.11.1 OTN Proof of Concept

The OTN proof of concept illustrates how an integrated NG-OTN and NGN-OSS systems work in concert to support advanced L2 services. In this setup, policy-based wide-area VLAN service with class-of-services and over-subscription was examined.

The proof of concept focuses on a BoD point-to-point VLAN service over the NG-OTN network. This point-to-point VLAN service can be established with the requested bandwidth on a real-time basis. A customer request may include various parameters such as: bandwidth ranges, priorities, and class of services (up to eight CoS can be supported). During the trial, the service classes were aggregated into high and low priority groups. For a high priority group, bandwidth is guaranteed, whereas, for a low priority group, over-subscription is determined by service contracts.

Table 3 shows sample test cases executed using the configuration shown in Figure 45. Each test case represents a service request with the required bandwidth (b/w), priority (Pri), and the observed outcome based on the service policy.

All VLAN traffic share bandwidth on the NG-OTN. The VLAN service requires that all L1 entities support Bandwidth-on Demand (BoD) services. The service relies on the BoD

mechanism to support most of its features. The key BoD capabilities required to support the VLAN service include: on-demand, fully automated provisioning, tear-down and bandwidth modifications of L1 end-to-end bandwidth of STS-1 and above.

Table 15: Sample Prototype Test Cases

Test Case	**VLAN**	**B/W**	**Port/ Platform**	**Pri**	**Policy**	**Results**
1	X	5 M	GE on OXC	High	within policy <Threshold%	Policy threshold determined adequate b/w was available for high-priority group.
2	Y	25 M	GigE on OXC	High	within policy <Threshold%	Policy threshold determined adequate b/w was available for high-priority group.
3	A	8 M	10/100 on Ethernet switch	High	within policy <Threshold%	Policy threshold determined adequate b/w was available for high-priority group.
4	B	8 M	10/100 on Ethernet switch	Low	within policy <Threshold%	Policy threshold determined adequate b/w was available for high and low-priority groups.
5	C	4 M	10/100 on Ethernet switch	Low	Congestion experienced, but within policy <Threshold%	Policy determined adequate b/w available for high priority group & congestion is within acceptable threshold for low-priority group. No additional L1 bandwidth is needed.
6	D	8 M	10/100 on Ethernet switch	Low	Over-subscription >Threshold%	Triggered provisioning of additional L1 bandwidth based on policy threshold.

Over-subscription was observed for the low priority VLANs, which were treated as basic service with the oversubscription threshold set by the service policy; (see test cases 5 and 6 in Table 15). As demonstrated in test case 6, layer-1 BoD provisioning was triggered by commands sent from the L2 OSS to provision the appropriate L1 bandwidth over the affected spans based on the service policy decisions.

4.12 Service Implementation

To implement the service, the following components were developed and installed:

- **VLAN service logic:** A service algorithm was defined for the VLAN service to guide the flow-through provisioning carried out in the L2 OSS. The logic defines a sequence of tasks that the L2 OSS must perform and identifies checkpoints where the service policy needs to be consulted for L1 provisioning.
- **VLAN service policy:** A simple L2 over-subscription policy was implemented for the low priority VLAN traffic. The policy resides in the L2 OSS and when triggered, it can interwork with the L1 OSS to initiate BoD provisioning for the required bandwidth.

4.13 Prototype Observations

Through the proof of concept efforts in prototyping various OTN services, several observations were made:

- NG-OTN technology is fully functional, but has not been perfected for BoD type services. Technology gaps in transport and control planes on NG-OTN NEs have been identified.
- There is a greater need for cross-layer management support in the OSSs, particularly with the interactions between the layers in the network.
- Other advanced L2 services can be implemented in a similar fashion using the PBIP NGN-OSS model as described in this section. L2 services such as those mentioned in the earlier sections and a multi-point version of the *Policy-Based Wide-Area VLAN service with class-of-services and over-subscription* can be designed and prototyped by defining a set of service logic and policies.

4.14 ACES Integration with OTN Control Plane

As shown in the ACES simulation, we have demonstrated that a centralized, intelligent Ethernet services provisioning system, coupled with standard data plane functionality, can be used to perform admission control and optimally route point-to-point EVCs within the network. Such a system can greatly improve the utilization of the network resources, while guaranteeing SLA performance objectives. In addition, ACES is expected to be used as a specific implementation of the layer-2 OSS service logic and/or service policy to signal directly to the network elements and request provisioning of required bandwidth channels.

As outlined in the BoD OTN prototype, we propose the implementation of a cross-layer Policy-Based Intelligent Provisioning (PBIP) system, based on L1 and L2 service logic and service policy, that adjusts transport bandwidth on demand, for Ethernet services. Furthermore, we demonstrated the ability to signal the bandwidth adjustment request using an integrated OTN control plane/LCAS implementation in the network.

There are many new standards that will enhance the marketability of Ethernet services to a wide customer base. The key gap we see is the need for a full set of service OAM functionality that operates at the EVC level, providing the Service Provider the ability to monitor the availability and performance of EVCs within the network.

Future directions and research activity include:

1. Prototyping ACES and generalizing it for provisioning over multiple layer-2 technologies
2. Integrating PBIP and ACES functionality into a single implementation system

3. Addressing the impact of Bandwidth on Demand services on key FCAPS functions.
4. Addressing the network scalability issues related to intelligent service provisioning over very large networks in a metro, regional and/or national deployment.
5. Working on fully implementing OTN control planes on network elements that need to interwork in a multi-vendor network environment
6. Working on enabling full migration to NG OSS in support of L1/2 convergence.
7. Ensuring support for key network requirements, with special focus on achieving a full set of service OAM functionality.

Key components in the future metro Ethernet services network need to include:

1. Fully standardized interfaces and capabilities - enabling multi-vendor interoperability for new and emerging services.
2. Multiple transport and switching technologies – Ethernet, MPLS, RPR switching; and SONET, xWDM, native Ethernet transport.
3. Integrated L1/2 control planes
4. Intelligent, policy-based service provisioning systems

The future metro networks will be discussed in greater details in the next chapter and share the long-term vision in implementing Ethernet over WDM and eliminate most of the key inhibitors for Ethernet end to end networking.

5. Ethernet QoS and support for SLAs

5.1 Introduction to Ethernet QoS

Ethernet is a broadband service that supports different applications such as voice, video and data. These applications have different service quality requirements that range from best effort to guaranteed delivery. For example, voice traffic requires stringent end-to-end delay variation and minimum frame loss guarantees, whereas, data traffic such as file transfer requires high throughput but less stringent requirements for delay and delay variation quality, and Internet traffic requires good throughput.

Quality of Service (QoS) is the ability to meet different user applications needs by distributing the scarce network resources amongst them in a fair manner. Native Ethernet is not a QoS based network and hence, it allocates network resources roughly equally (best effort) unless a soft QoS mechanism is used, such as priority bit marking [1].

Supporting QoS is one of the key carrier class issues facing a service provider in deploying Metro Ethernet [10]. Traditional Carrier services, such as SONET, allow a service provider to guarantee Service Level Objectives parameters including very low loss and delay performance coupled with very high availability. However, unlike these traditional services, Ethernet services do not have the ability to make such stringent guarantees and therefore treats them as Service Level Targets or Objectives. This presents the carriers with a major challenge on how to offer, migrate, or converge traditional services (TDM based services) over Ethernet and maintain the same level of guarantees.

This chapter covers the basic definitions of QoS and introduces an enhancement to algorithms used by ACES, whereby, real-time feedback from network probes placed at

the provider's edge switches or located at the end-user premises will ensure that key performance metrics (such as latency, frame loss ratio and frame delay variation) are met. The proposed enhancement is based on a network model within a provider trusted domain, i.e. edge to edge, as opposed to a network model that also spans the customer edge equipment, which is in the un-trusted domain. A provider may extend his 'trusted' domain to a customer premises by offering managed service where the customer edge equipment is managed by the provider.

The first part of this chapter covers what Ethernet QoS is, and what its associated network control functions are (i.e. classification, shaping, policing, etc.) The second part of this chapter introduces Ethernet traffic characteristics, which include Ethernet bandwidth and service profiles. The third part of this chapter deals with SLAs, performance monitoring, queuing, QoS switching and congestion avoidance. The last part of this chapter discusses the proposed enhancements to ACES to make it more efficient for supporting end-to-end QoS and SLAs.

5.2 Basic QoS Concepts

QoS mechanisms enable preferential treatment of packets for certain classes of service by raising their service frame priority in the network. Current QoS mechanisms include priority marking of 802.1p frames and and DSCP marking of IP packets. diffserv assured forwarding (AF) per hop behavior (PHB) [10], [16], [25], [35]. These mechanisms allow a service provider to offer multiple classes of service differentiated by performance targets, such as *platinum, gold, silver, and bronze.* The QoS mechanisms associated with these classes of service do not sufficiently guarantee a service level agreement, and therefore, a service provider must complement them with engineering practices that are

implemented and monitored on a regular basis to minimize any network related performance problems [57].

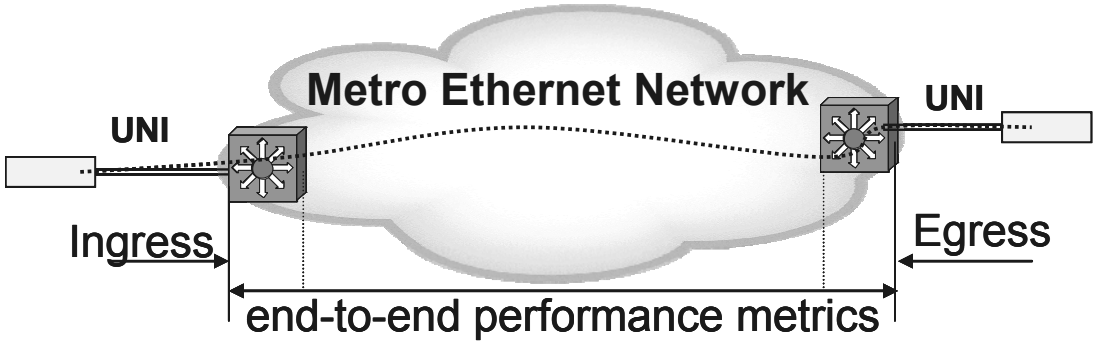

Figure 46: Ingress and Egress in a "trusted" Network

In Figure 46, the ingress node plays the most important role in QoS since it is the first point in the network where traffic is classified and processed using QoS mechanisms along the path as it traverses the network [24], [50].

5.2.1 Ethernet Differentiated Services

Ethernet can use the Differentiated Service (Diff Serv) model in a metro area network to prioritize traffic, however, Diff Serv does not have a topology-aware admission control mechanism to prevent establishing an EVC that might degrade all existing EVCs in the network [32], [50]. The proposed ACES enhancement model will build upon the Diff Serv functionality (enhancing it with the control management that it currently lacks) and will use the code points settings to ensure guaranteed QoS on a per hop behavior. The Diff Serv code points define performance classes that can be charactertized by specific performance metrics, which are achievable under normal and some congestion operating conditions. These classes are based on the IP ToS field, which is translated to an equivalent markings found in 802.1p (3bits) tagged Ethernet frames (see Figure 47), [50].

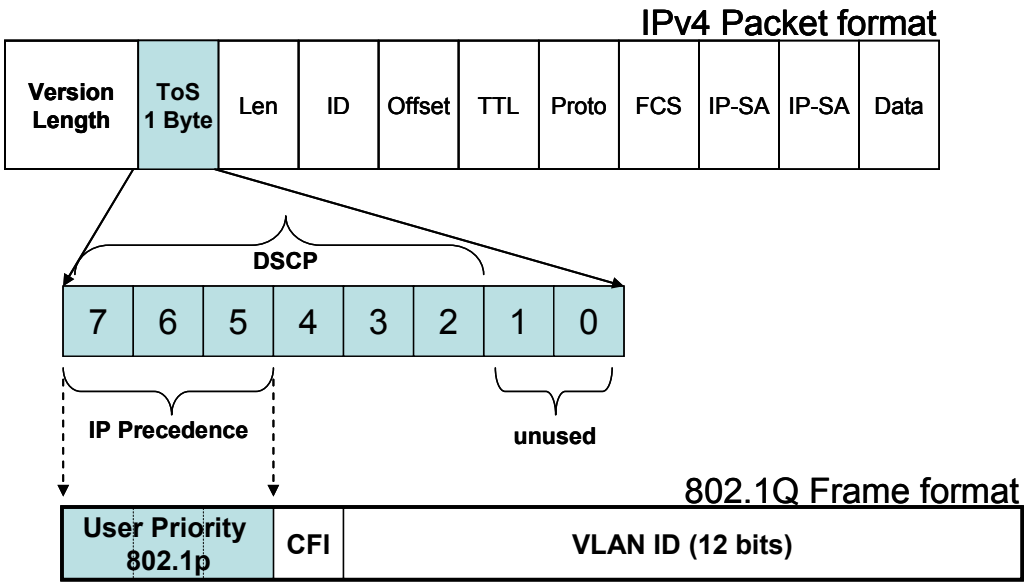

Figure 47: IP Precedence to Ethernet 802.1p bits mappings

Diff Serv defines the quality of service for a given class based on the settings of the type-of-service p-bits included in the VLAN or EVC and then map its traffic into the appropriate queues in the network. Also, you can set the switch port with an appropriate service class, which will then mark all of the incoming traffic with the same class of service. Traffic markings from a “trusted” source can be mapped or copied into a service provider tag and are used in the admission decision into the network. Also, IP precedence/DSCP markings of IP Layer-3 packets can be mapped into Ethernet Layer-2 CoS bits (802.1p bits) as illustrated in our example in Figure 48.

Forwarding Class Name	CoS + Drop Precedence Identifier Value	P-bits	802.1d recommended mapping (3 Queues)	
Premium	7	111	VO	7
	6	110		6
Gold	5 (Green color)	101	CL	5
	4 (Yellow color)	100		4
Standard	3	11	BE	3
	2	10		2
	1	1		1
	0	0		0
VO: Voice; CL: Control Load; BE: Best Effort				

Figure 48: Possible mapping of IP precedence and CoS into p-bits

Traffic from an "un-trusted" source will be re-marked based on the EVC service profile as it enters the service provider network. Ethernet Class of Service parameters can be applied to:

- The UNI physical port, where all traffic entering and leaving the port, share the same class of service.
- Source or Destination MAC address, where a class of service is assigned based on the source and destination MAC addresses.
- VLAN ID (VID inside the 802.1Q tag), where user traffic belonging to this VLAN receives the selected class of service, while other traffic carried over a different VLAN ID, within the same UNI interface, receives a different class of service.
- EVC 802.1p value, where user traffic can share up to three classes of service, depends on the service provider offerings.
- Diff Serv/IP ToS, where user traffic can be classified up to 8 classes based on the IP TOS field.

Diff Serv Code Points allow for up to 64 classes of service and may have different per hop behaviors [90], which are:

- Expedited Forwarding (EF), used for low delay and low loss traffic such as VoIP
- Assured Forwarding (AF), used for bursty real time and non-real time class of service for up to 4 classes
- Class Selector (CS), used for backward compatibility with IP ToS and default Forwarding (FD) for best effort services.

As an Ethernet frame arrives at the first ingress node, it will go to the classifier to be classified into the appropriate queues, (see Figure 49). Then the meter measures the rate of the Ethernet frame stream and passes this information to other elements that trigger a particular action. The marker sets the p-bit QoS value of a frame, effectively adding it to a particular behavior aggregate queue. As congestion in the network increases, overflow of arriving frames will be dropped unless queuing algorithms handle the excess traffic. The queuing algorithms include: priority queuing (ensures that strict priority is provided to important traffic), weighted fair queuing (ensures that traffic gets predictable service and no starvation occurs) and first-in first-out queuing (ensures that frames are sent in order of their arrival and it has no view into a packet priority), [16], [32].

Queue management and congestion avoidance are techniques that monitor the network element traffic loads and try to estimate, predict and avoid congestion, thus improving QoS in a given network. There are additional mechanisms to improve the QoS in a network element, which includes congestion management. Congestion management controls the network congestion after it happens and provides flow control mechanisms that limit or scale back the traffic.

The following QoS functions are defined in the network element, [32], [50], [90]:

Classifier: Contains the frame identifier or classifier and marker sub-functions. If the frame is classified and the marking is not set, then this condition is referred to as per hop behavior basis (PHB), see Figure 49. The classifier determines the type of frame entering the network based on its VALN ID and p-bits value. Once a frame type is determined, then the frame is forwarded as appropriate or it is subjected to traffic conditioning, if necessary; A multi-field classifier selects frames based on one or more fields such as source MAC address, destination MAC address, p-bit field, protocol ID, source or destination port.

Marker is the function that sets the value of the CoS field or 802.1p bit or re-writes it. In addition, if packet coloring is supported, then it will color the frame as per the MEF technical specifications.

Traffic Conditioner is the function which meters, marks, drops, and/or shapes traffic. A traffic conditioner may re-mark an EVC traffic flow, or may discard or shape its packets to bring it into compliance with its traffic profile.

Meter measures the rate of traffic flow selected by the classifier and uses the info for accounting and statistical measurement purposes. In our model, we attach policing to the metering function that evaluates the meter measurements and uses them to enforce the EVC policy-based traffic profile. Policing is currently done at the ingress port of the network element and may use the dual token bucket specified in MEF Technical Specification [10]. There are proposals to include policing at the egress edge of the network to prevent rate mismatch and overcome the queue buffer limitations.

Dropper discards some or all of the frames in a traffic flow and enforces EVC policy compliance. Initially, red frames are dropped, but based on network congestion condition; yellow frames will be dropped and lastly green frames will be dropped depending on how serious the network congestion is.

Shaper delays frames within an EVC flow to cause it to conform to its traffic profile based on the size of the buffer. The larger the buffer size the more delay may be added to the packet delay time. This may impact delay sensitive traffic.

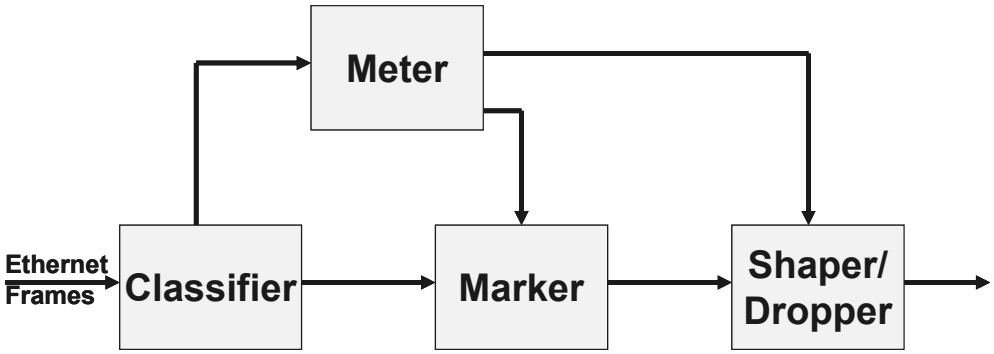

Figure 49: Ingress Ethernet Frame traffic per hop bandwidth profile model

To address end-to-end QoS, there are many issues to be resolved. ACES model does resolve the bandwidth issues and perhaps it alleviates any CoS issues based on engineering practices, however, it does not include the mechanisms to guarantee any SLAs. Dynamic negotiation through signaling protocols such as RSVP allows an EVC to request for a specific QoS, but for a large network, scalability is an issue as it requires resources for setting up the frame classification, scheduling, and end-to-end flow monitoring.

In order to design and develop Ethernet network services, we need to understand the characteristics and requirements of the traffic that will be carried over Ethernet in order to classify the application requirements.

Ethernet traffic characteristics can vary widely depending on the user application. Key criteria to characterize Ethernet traffic is its measured burstiness as the maximum to the average bit ratio. As the traffic burstiness increases, it becomes more difficult for the network to handle the traffic. As the burstiness of a source node increases, it decreases the predictability of its traffic pattern or behavior. In addition, a large variety of applications originating from traffic sources in a LAN such as a personal computer, a VoIP or edge router may have different QoS requirements such as:

- Traffic that is tolerant to delay, but is loss sensitive, such as voice or interactive applications
- Traffic that is tolerant to moderate loss, but is delay sensitive, such as VoIP
- Traffic that is tolerant to both delay and loss, such as Internet traffic

Applications QoS requirements can be categorized into:

- Bandwidth traffic characteristics (peak rate, average rate, and burst size)
- Quality of Service requirements
    a. Connection-oriented emulation (setup delay time, release delay time, blocking probability)
    b. Connectionless (frame loss ratio, end-to-end delay, jitter, throughput guarantees, etc.)

These Ethernet application requirements can be mapped mainly into three performance metrics:

1. Bandwidth sensitive traffic with Frame loss ratio (FLR) performance metric

   Traffic is characterized by the amount of information transmitted and received such as intranet file transfer, etc.

2. Delay sensitive with average or maximum delay performance metric

   Traffic is characterized by rate and duration in real time such as voice, video conferencing, and transaction based applications

3. Mixed LAN traffic that varies

   Traffic parameters vary at all times and depend on long-term behavior. Traffic tends to be self similar but does not have time duration where the traffic is constant.

5.3 Ethernet Traffic Characteristics

Ethernet is a connectionless best effort frame delivery service. In a metro area, Ethernet was introduced as a point to point best effort service with some priority queues that relied mainly on buffer management and frame discard as the means to control network congestion. Typical applications had no knowledge of when Ethernet frames will be delivered or whether they were delivered. To support quality of service and to apply it to end-user traffic with the same performance parameters, service classes were defined for customers to choose from (see section 3.6.1.7) and network resources were made aware of these requirements.

5.3.1 Ethernet Bandwidth Profiles

A bandwidth profile allows a customer to purchase the Ethernet service that best supports its applications requirements [10]. An Ethernet provider specifies a UNI with one or more CoS bandwidth profile parameters, see Figure 50, as follows:

- CIR = Committed Information Rate, in Mbps.
- CBS = Committed Burst Size; Burst size of the committed rate token bucket, in bytes (B)
- EIR = Excess Information Rate, in Mbps
- EBS = Excess Burst Size; Burst size of the excess rate token bucket, in bytes

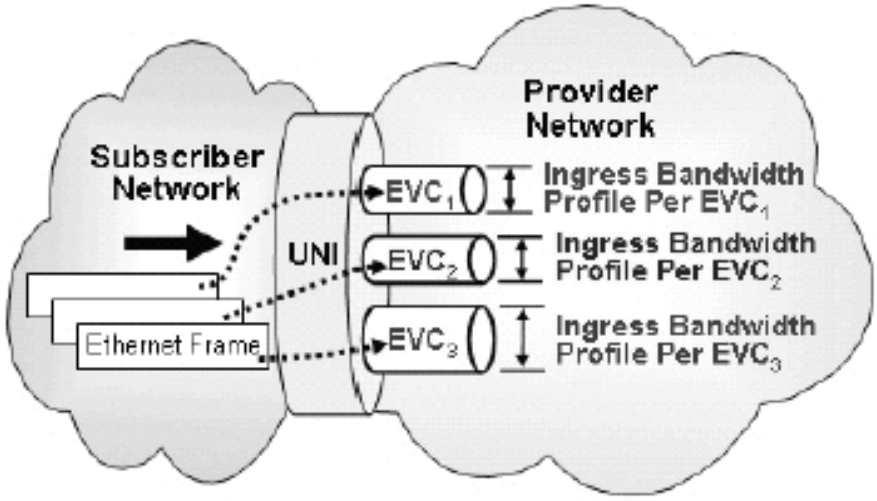

Figure 50: Ethernet Bandwidth Profile

Ethernet service frames are referred to as ‘in-profile’ or ‘conformant’ to the bandwidth profile if they meet the network performance requirements, such as CIR and EIR described below. They are referred to as ‘out-of-profile’ or ‘non-conformant’ to the bandwidth profile when they are allowed into the provider’s network and are delivered without any service performance objectives, or are dropped at any point along the path

due to network congestion. MEF technical specifications [10] defines the following bandwidth performance metrics:

1. Committed information rate (CIR), represent the average guaranteed transmission rate in a defined interval of time that Ethernet service frames are ‘in-profile’.
2. Excess information rate (EIR), specifies the average rate above the CIR rate that may be allowed into the network. These frames are tagged yellow depending on the congestion conditions in the network and are considered as ‘out-of-profile’ Ethernet frames.
3. Committed burst size (CBS), represents the maximum number of bytes allowed for incoming service frames that are still considered ‘in-profile’ frames or CIR-conformant, frames.
4. Excess burst size (EBS), is the maximum number of bytes allowed to be admitted for incoming service frames to be EIR (out of profile) and may not be guaranteed delivery through the network.

Although the MEF technical committee has not yet defined how color marking should be indicated, however, they have defined a useful way of tracking service frames through the use of colors by looking at whether the average rate of the service frames is ‘in-profile’ or ‘out-of-profile’. A Service Frame is ‘green’ (if it is conformant with ‘committed’ rate of the bandwidth profile); ‘yellow’ (if it is not conformant with the ‘committed’ rate but conformant with the ‘excess’ rate of the bandwidth profile) and ‘red’ if it is not conformant with either the ‘committed’ or ‘excess’ rates of the bandwidth profile). Therefore, green service frames have guarantees to be delivered per the service performance objectives and usually should not be discarded, yellow frames are typically

admitted but do not have guarantees and are not immediately discarded and red frames normally are not admitted and are discarded immediately.

Ethernet service frames designated as ‘Green’ are delivered per their service performance objectives. These service frames are not to be discarded because they fall within their performance metrics and are considered ‘in-profile’ frames that conform to the bandwidth profile requirement. On the other hand, yellow service frames are out-of-profile but are typically not immediately discarded. These yellow service frames are not delivered per their service performance objectives and may get discarded by the network under different conditions such as network congestion. Red service frames are designated as ‘out-of-profile’ and are usually not allowed into the network and are discarded immediately.

5.3.2 Service Performance Attributes

This section defines the service performance objectives associated with each CoS offering based on the MEF technical specification [10]. A customer is able to select and purchase the class of service that best meets their application service performance requirements.

Three service performance attributes – Frame Delay, Frame Delay Variation or Jitter (FDV), and Frame Loss Ratio (FLR) – are defined. In addition, the service performance objectives related to the performance SLA are also documented.

5.3.2.1 Network Delay or Latency

Network delay is the time it takes to carry an Ethernet frame between two defined end-points (ingress and egress) within a network. Network delay, N_D, is the sum of

propagation (P_{gd}), transmission (T_d), processing (P_d), and queuing (Q_d) delays associated with an Ethernet frame traversing a network.

$$N_D = \sum_{i}^{n} (P_{gd} + T_d + P_d + Q_d),$$

where n is the total number of nodes in a network along an EVC path; P_{gd} is the propagation delay time for a signal to travel down the communication facility; T_d is the transmission delay time across each link in a communication network; P_d is the processing delay across each switch between the two end points; and Q_d is the queuing delay time associated with the frame waiting in a buffer prior to a given action to take place. In other words, latency is a fixed delay contribution within a network that is beyond the control of a network engineer, such as propagation and switch delays. For simplicity, latency and delay will be used interchangeably.

5.3.2.2 Frame Delay Variation (FDV)

Frame delay variation (FDV) is a measure of the variation in the frame delay between a pair of service frames. FDV is also referred to as frame jitter to represent the fluctuation of the end-to-end frame latency. FDV is defined as the variance in frame delay (in milliseconds) of the actual inter-frame arrival time to the expected inter-frame arrival time between two service frames as measured at the ingress and egress UNIs, and is measured on a pair-by-pair Ethernet frame basis as follows [5]:

$$FJ_{ij} = (t_{rj} - t_{ri}) - (t_{sj} - t_{si}),$$

where t_{ri} t_{rj} are the receive times for frames i and j, and t_{si}, t_{sj} are the send times for frames i and j.

In order to eliminate network jitter, we must first understand how it happens. As discussed, frame delay variation (FDV) is a variable delay introduced by the Ethernet network due to asynchronous nature of switched Ethernet traffic (bursty) and its varying frame sizes traversing the network. This delay can impact voice traffic adversely that may require echo cancellation equipment. Network latency is measured from the ingress point at which an Ethernet frame enters the network to the egress point at which it leaves the network. For example, if a node transmits a message that is longer than the maximum frame size (e.g 1522 bytes), then the message is broken up into frames and each frame is sent into the network one at a time. These frames are sent to the first node in the network path where they are stored temporarily in the input buffer (ingress port), then the switch determines the next node based on the destination VLAN/MAC address by consulting its forwarding table. The switch forwarding engine moves the frame to the output buffer associated with the outgoing trunk-link that the EVC rides on. Each frame is then transmitted to the next node as quickly as possible, similar to a statistical time division multiplexing. However, priority frames may be transmitted after non-priority frames that result in a significant frame delay variation that may adversely impact the service performance. Although all of the message frames will get transmitted to their destination node, the effect of high amount of frame delay variation may degrade the service. The service provider may need to increase the destination buffers (the jitter buffers) to accommodate late or early arriving frames, however, the larger the buffer size the larger the queuing delay.

5.3.2.3 One-Way Delay Time and Round Trip Delay (RTD) Definition

One-way delay (OWD) is defined as the time it takes for a service frame traverse from the ingress UNI to the destination egress UNI. Round-trip Delay (RTD) is defined as the time (in milliseconds) it takes for a service frame to be sent from the first ingress UNI to last egress UNI and back again (includes link insertion delays, propagation delays and queuing delays in the network). The RTD calculation includes only the time the packet is in the network, i.e., the processing time spent in devices attached to the UNI are factored out of the definition. The RTD definition is restricted to in-profile service frames (CIR compliant or green frames). Monitoring of service frame delay will be based on round trip measurements. Specifically, RTD is to be defined for service frames less than 1522 Bytes as 'first bit out' to 'first bit in' at the source and excludes any processing delays at source and destination.

$$RTD = t_r - t_s - t_p ,$$

where t_r is the received time at source, t_s is the sender time at source, and t_p is the processing time at source and destination.

One-way and round-trip delay measurements are performed with the assumed average size of the frame for the given CoS to be monitored. These measurements provide helpful information in establishing the maximum expected throughput for a given EVC CoS in a given MEN.

The expression $T=W/R_d$ is used to calculate a network element or a switch throughput. This expression is a simple system throughput definition, provided that the window size is not too large T is the end-to-end throughput, W is the window size (the number of bits

in the frame transmitted) and R_d is the frame round trip delay within a given system. A more complex expression needs to be developed for the network that is based on either a window-based or rate-based system.

5.3.2.4 Frame Loss Ratio (FLR)

Frame Loss Ratio (FLR) is defined as the number that the network offers as a target at the time of the connection over the life of the connection and applies to two parts [10]:

1. Ingress frames that are 'in-profile' sent during an interval
2. Egress frames that are 'in-profile' received during an interval

FLR is equal to the total number of 'in profile' frames sent – total number of 'in-profile' received during an interval such as the life of the EVC connection and is given by:

$$FLR = \left(\frac{F_{is} - F_{ir}}{F_{is}} \right) \times 100\%$$

Ingress or Egress frames that are 'out-of-profile' sent or received, F_{or} are tagged either red or yellow and do not play any role in the FLR calculation.

5.4 Service Level Agreement

Service level agreements (SLAs) are used to define the service level guarantees between customers and service providers. Service level specifications (SLS) define the network service parameters, how they are measured, and any penalties if they are not met.

A SLA consists of three components. They are:

1. Contractual Obligations

   Contain information about the parties involved such as contact info and procedures

2. Technical Specifications

   Contain product information or service description defining user connections and applicable network rules that may be used to restrict or limit the user traffic such as shaping or policing; Quality of service parameters such as bandwidth guarantees, frame/packet loss, jitter and delay

3. Administrative and Operational Methods & Procedures

   Contain admin information such as billing, pricing and operational details such as fault, trouble reporting, support, and escalation procedures

Ethernet is a connectionless technology with frames that have a 3-bit priority field in which the customer can indicate their preference class of service which can be enforced or bypassed based on the SLS and SLA agreements.

For the hybrid model, we require that both Ethernet and Optical cross connect management and control planes to offer a unified automated platform for provisioning and maintaining connections, and managing network resources. In order to support an automated platform, knowledge of a network's topology (inventory and connectivity) and its available resources is crucial. The network topology and its resources should be auto-discovered, requiring methods for neighbor and end-system discovery, and for sharing this information throughout the network. In order to utilize network resources efficiently, it will require that an inventory of the currently available network resources be updated

and maintained. The automation of connection provisioning requires sophisticated customizable algorithms for route selection, and signaling mechanisms to request and establish connectivity within the network along a chosen route. Once a connection is successfully established, its availability and bandwidth (whether static or dynamic) needs to be maintained subject to negotiated service level agreements (SLAs.)

For the Ethernet Virtual Private Line (EVPL) service, which offers a virtual point-to-point connection, a customer can have up to three classes of service in any combination on a given Ethernet Virtual Connection (EVC) as follows:

- EVPL-Real-time (EVPL-Gold) is designed for packet voice, video and other customer applications requiring tight guarantees of frame delay, frame jitter and frame delivery performance.
- EVPL-Priority Data (EVPL-Silver) is designed for customer data applications requiring guarantees of frame delivery performance, with looser requirements for frame delay.
- EVPL-Bronze (EVPL-B) is designed for customer data applications that require no guarantees of performance, and where the cost per Mbps is the key requirement.

5.4.1 Service Performance Objectives

The service performance objectives for Ethernet virtual private line CoS (EVPL-Gold, EVPL-Silver and E-LAN-Gold) are listed in Table 16 within a medium size LATA network as a reference example.

Table 16: Examples of Service Performance Objectives per Class of Service (CoS)

Service Performance Attribute	*Parameters*	*EVPL-Gold CoS*	*EVPL-Silver CoS*	*E-LAN-Gold CoS*
One-way Delay (OWD)	OWD Objective	10 ms	25 ms	15 ms
	Percentile	99.9%	99.0%	95.0%
	Time interval Δt	5 minutes	5 minutes	5 minutes
Frame Delay Variation (FDV)	FDV Objective	5 ms	15 ms	10 ms
	Percentile	99%	N/A	N/A
	Time interval Δt	15 minutes	30 minutes	N/A
Frame Loss Ratio (FLR)	DDR Objective	99.5%	99%	99%
	Time interval Δt	1 hour	1 hour	1 hour
CoS Guarantee	Monthly/ Quarterly SLA	99% of Δt meet objective	98% of Δt meet objective	95% of Δt meet objective
Service Availability				

These SLA numbers may be conservative as compared to a service provider offering, but are used for discussion purposes and they take into account the possibility of large queuing delays within a given network. While queuing delays are possible, an architecture design is built with enough backbone capacity to minimize queuing delays.

The expected performance under normal operating conditions, in the absence of significant queuing delays within the network, is shown in Table 17.

Table 17: Example of Service Performance Objectives per Class of Service

Service Performance Attribute	*EVPL-Gold*	*EVPL-Silver*	*E-LAN-Gold*
Frame Delay (FD)	< 3 ms	< 5 ms	< 4 ms
Frame Delay Variation (FDV)	< 1 ms	< 5 ms	< 5 ms
Frame Loss Ratio (FLR)	> 99.999%	> 99.99%	99.99%
Service Availability			

To obtain Table 17 numbers, the following assumptions are made:

- 100 Mbps UNIs at each end,
- six hops through the metro network
- 200-mile network diameter.

The delay numbers are mostly dependent on propagation delay (i.e., network diameter). These numbers represent the expected average customer size network in a metro area, however, for most services requiring only one UNI, the numbers achieve much better performance due to the less number of hops and about half of the network end-to-end diameter. The model numbers can be adjusted based on actual field measurements and data monitoring.

5.4.2 Service Performance Results

The performance SLA Data are collected and are correlated to maintain and support an SLA. The data provides a level of abstraction to provide meaningful information. The data is statistical in nature and data collection is done in increments of time, based on the SLA performance parameters offered. The following sections describe the parameters that are considered for an SLA.

5.4.2.1 Per Δt Conformance

For each Δt in the month (5 minutes increments up to x hours), a determination is made as to whether the performance objectives are 'Met' for the CoS attributes related to the CoS instance on a given EVC. So, for a given Hour (e.g., hour-1), the overall performance objective is 'Met' if the performance objectives for each of the one way attributes {delay, FDV, CLR} are 'Met'. If any of the attribute objectives are 'Missed', then the overall performance objective for Δt (5 minutes) is determined to be 'Missed'.

5.4.2.2 Per-Month/Quarter/Annual Conformance

For the month, quarter or annual period, a determination is made as to the percentage of hours that the overall performance objective is 'Met'. Thus, for a given month, the monthly performance guarantee is 'Met' if the % of Δt 'Met' for the month does meet or exceed the monthly objective.

5.4.2.3 Credits

For a given Class of Service on a given EVC, an agreed upon percentage (%) of the monthly recurring charge (MRC) will be credited back to the customer if the provider fails to meet the performance objective.

5.5 ACES Enhancements

As discussed previously, the ACES simulation model examines the trunk-link bandwidth utilization per service class per queue. The model maximizes the link and network efficiency. Also, ACES looks for engineering practices for measurements to engineer the network for average delay, frame loss ratio and availability. In this section, we examine how to enhance the ACES model by looking at the key performance metrics, feedback from the probes deployed in the network, and the new developed algorithm (to be

discussed later) that consults a table of choices available for ACES to meet the end-to-end service level guarantees along a selected path, as shown in Figure 46.

With this proposed ACES enhancement model, we ensure that the Ethernet architecture provides the needed QoS resource reservation, EVC admission control, network bandwidth tracking, frame scheduling and buffer management.

Admission control function of ACES limits the load on the queuing system by determining if an incoming request for new service can be met without disrupting the service guarantees to the already established Ethernet Virtual Circuits (EVCs). So, when a request arrives, the admission control will determine whether to accept, reject, or place the request in a holding (wait) status, pending results of adding bandwidth in the optical network or path link utilization feedback from real-time measurements from the network collection probes. To this end, this section will discuss key performance measurement architecture and related tools. In addition, the admission control may ensure certain behavior on admitted EVC to ensure network efficiency and congestion management techniques, see Figure 51.

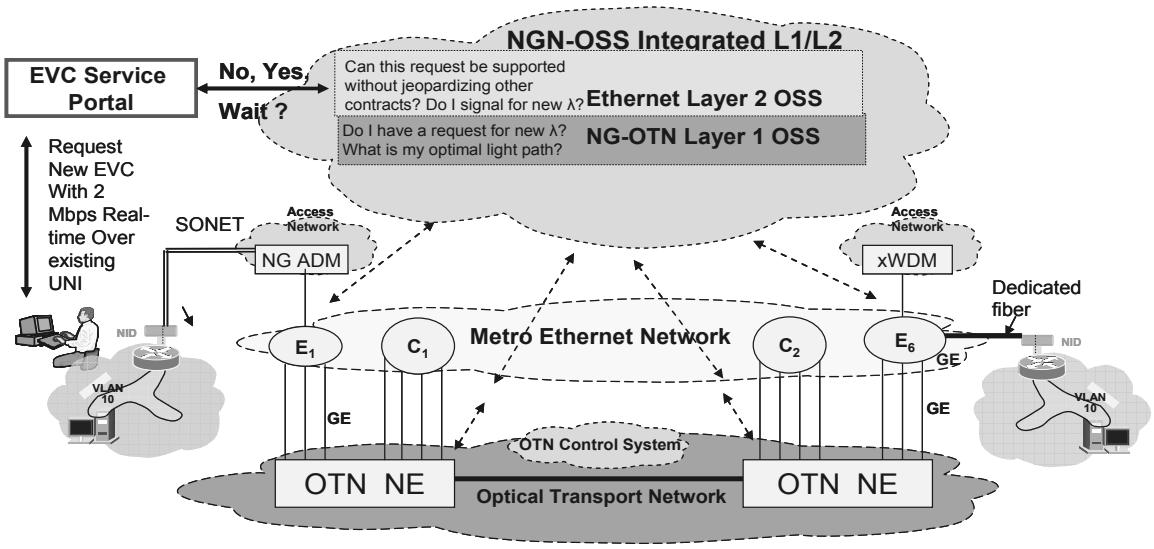

Figure 51: Admission Control for Ethernet Services

5.5.1 EVC Access Control

This algorithm will shape the EVC data flows at the ingress to network such as a customer router or network interface device. In addition, this algorithm helps ensure that the overall EVC conformance to the bandwidth profile is met at other specific points within the network. At each point, the network traffic descriptor is enforced and excess customer traffic may be dropped, tagged low priority or delayed.

5.5.2 Packet Scheduler

Packet scheduling defines the queue service discipline at a node in the network and enforces a set of rules in sharing the trunk-link bandwidth. Packet scheduling prioritizes user traffic as *delay priority* for real-time traffic or *loss priority* for data-type traffic. In addition to packet scheduler, packet fair queuing is applied to ensure fairness across all users.

5.5.3 Buffer Management

Sharing a common pool of buffers is a major concern in switches where high speed capacity links such as 1GigE and 10 GigE are employed. Buffer management defines the sharing policy and decides which packet should be discarded when the buffer overflows.

5.5.4 Congestion control

In Ethernet networks, the UNI/NNI load may be larger than the mapped SONET channel of STS-3 or STS-12 (for GigE) or STS-48 (for 10GigE) circuits. The same is true for a GigE circuit over an ATM OC-12. Hence, if no measures are taken to restrict the entrance of traffic, queue sizes at bottleneck links will grow and packet delays will increase, which will eventually cause the buffer space to be exhausted and some of the incoming frames to be discarded. Congestion control regulates the frame population within the network and helps eliminate speed mismatch issues created by different UNI speeds in an EVC.

5.5.5 Bandwidth Profile Rate Enforcement

In Figure 52, One bucket, referred to as the ‘Committed’ or ‘C’ bucket, is used to determine CIR-conformant, ‘in-profile’ service frames while a second bucket, referred to as the ‘Excess’, or ‘E’ bucket, is used to determine EIR-conformant, excess Service Frames. Each token bucket consists of a bucket of bytes referred to as ‘tokens’. Initially, each token bucket is full of tokens. As service frames enter the provider’s network, the ‘two rate three color marker’ (trTCM) algorithm, which can be implemented via two token buckets, decrements the number of tokens in the C bucket (green tokens) by the number of bytes received from the service frame. If green tokens still remain, then the service frame is CIR-conformant, it is colored green and then allowed into the provider’s network. However, if no green tokens remain, then a second E bucket is checked to

determine whether any E bucket tokens (yellow tokens) remain or not. If yellow tokens are available, then the service frame is colored yellow and allowed into the provider's network. Otherwise, no yellow tokens are available, and the service frame is declared red and discarded. MEF Technical committee has defined an additional, optional capability of the (tcTCM) algorithm, whereby, unused green tokens from the C bucket may be added to the E bucket as yellow tokens when checking EIR-conformance. If this capability is enabled, then when operating in color-aware mode, more yellow service frames are allowed into the service provider's network. MEF currently has no quantitative data describing the implications of this approach. This capability is not expected to be specified in an SLA to the subscriber.

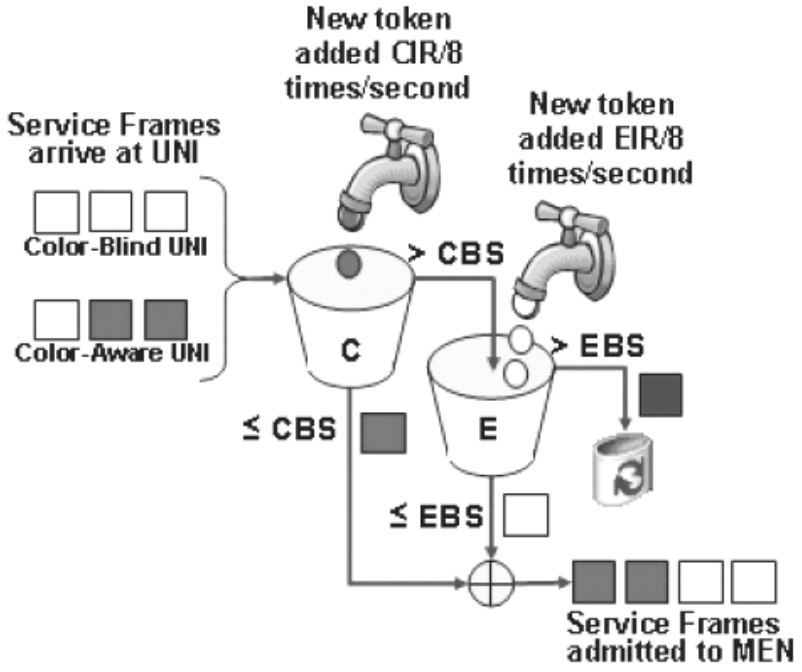

Figure 52: Ethernet Dual Token Buckets

Assume that an Ethernet frame is marked yellow by the subscriber's edge device based on a yellow DSCP value mapped to the CE-VLAN CoS value. The token bucket

algorithm (TBA), a popular rule-based traffic descriptor, would then only check the service frame's conformance with the E bucket and bypass the C bucket all together. As with the color blind UNI, if no yellow tokens are available, then the service frame is declared red and discarded. However, if yellow tokens are available, then the service frame is admitted into the provider's network.

Therefore, the use of rate-based throughput system does not control the number of outstanding frames in transit, as in windowing system, but controls the explicit rate at which the frames are transmitted (i.e frames per second). The rate-based system is proposed here as a mechanism to reduce the probability of overload in Ethernet networks by spreading out the Ethernet frame arrivals into a congested network resource, thus reducing the queuing delay (see Figure 52) and the number of frames in the congested resource buffer. It is proposed to use the leaky bucket rate-based throughput mechanism, which is used in a similar manner in frame relay service [34].

$$E(W) = [U/(1-U)] * Ts * \exp [-(1-U)*K/S],$$

Where U is the system load, S is the system average bucket size; K is a measure of the size of the token buffer (same units used to measure average frame size). The mean queuing delay expression is derived based on several assumption that include: Poisson arrivals, exponential service times, and infinitesimal token sizes. If the token buffer size K=0, then the mean waiting time becomes that of M/M/1 expression. Deterministic traffic descriptors have advantages over statistical descriptors for specifying and monitoring traffic. The dual leaky bucket descriptor is a popular traffic descriptor and the CAC for rule based sources raises the following characterization issues, assuming that each source submits its worst-case traffic to the network, they are:

- Worst case traffic compliant with the traffic descriptor
- Combination of sources that can be supported without violating the QoS requirements

Figure 52 shows the Ethernet dual token buckets. This model does not introduce delay into the Ethernet traffic, unlike dual leaky bucket, which performs pacing and shaping and thus adds delay.

5.6 Ethernet QoS Switching

Current Ethernet switching mechanisms are made without QoS awareness or verification of whether resource availability and requirements are met. This means that an EVC may be switched over a path that is unable to support the service requirements, while an alternate path with sufficient resources exits that supports its requirements. This will potentially lead to performance issues and service degradation. Strict switching algorithms are needed to identify a feasible path that has sufficient residual bandwidth resources that can satisfy the QoS constraints for a given EVC connection.

Ethernet frame-switching services provide for statistical multiplexing. In addition, Ethernet best effort service can be oversubscribed. Therefore, Ethernet will allow different services to share a link capacity that is less than the sum of their service requirements. However, since the resources are shared, then it is up to an admission control algorithm to determine the proper resource allocation for an admitted EVC, and to track it to ensure its service commitments are met. The call admission control (CAC) may consider the resource allocation at different intervals: EVC CoS Level, UNI burst level, or frame rate level. The material presented in the following sections have been influenced

by reading the following references [1], [4], [10], [24] and definitions of metro Ethernet services.

5.6.1 Deterministic and Probabilistic Bounds

Ethernet traffic is bursty in nature and can be statistically multiplexed. Statistical multiplexing is the interleaving of frames from different sources where the instantaneous degree of multiplexing is determined by the statistical characteristics of the sources. Networks that support statistical multiplexing can achieve a higher level of utilization without sacrificing much on QoS.

5.6.1.1 Deterministic

A deterministic guaranteed service provides for the worst-case requirements for EVCs. These requirements are based on measured service parameters performance on a similar or on the same path within the metro area. Hence, the call admission control must determine whether admitting a new EVC request to the network will cause the network to violate any of its key SLA parameters. Deterministic traffic descriptors offer many advantages over statistical descriptors for specifying and monitoring traffic.

To characterize uniformly worst-case behavior would provide valuable information for the CAC and simplify its rule-based system. Hence, a CAC may require sources to provide peak rate characterization of their traffic. The CAC algorithm will then check whether the sum of all peak rates is less than the link bandwidth available. The network will ensure that all admitted flows' token rates is less than the link bandwidth and the sum of all of the token bucket depths is less than the available buffer space.

Using deterministic algorithm for Ethernet bursty traffic will result in low utilization. Hence, shaping the traffic to meet network resources may improve the network utilization but it will not meet the user end-to-end service requirements. Therefore, instead of imposing traffic shaping at the setup time for the entire EVC, perhaps, negotiating different segment characterization of a real time stream prior to the transmission of each segment as in video on demand service where the entire stream is available for a *priori* characterization prior to the transmission.

5.6.1.2 Probabilistic Bounds

Probabilistic guaranteed service exploits the statistical multiplexing and does not provide for the worst-case scenario. Instead, it uses the statistical characteristics of the current and incoming traffic and guarantees a bound on the probability of lost frames:

$$\Pr\{(\text{aggregate traffic} - \text{available bandwidth})*\tau > \text{buffer}\} \leq \epsilon,$$

Where τ is the time interval and ϵ is the desired loss rate. The effective bandwidth is equal to the aggregate traffic of the statistically multiplexed sources. Amongst the key distribution algorithms used include: *Bernoulli* and *Binomial, Gaussian,* and *Poisson.*

5.6.1.3 Effective Bandwidth model

Effective bandwidth is calculated based on the source traffic characteristics and the service requirements. In a network, modeled as Markovian traffic sources, it is possible to assign an effective bandwidth to each source that is explicitly identified [24]. Each source is modeled as either 'on' or 'off', (see Figure 53). The 'on' state represents an exponentially distributed length of time with a mean length of $1/\beta$ and emits data with a peak rate of λ_p. The 'off' state represents an exponentially distributed length of time other than the 'on' state with a mean length of $1/\alpha$ and emits no data.

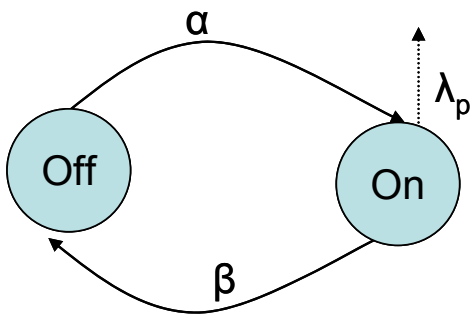

Figure 53: On-Off source model

Consider a departure process from a dual leaky buckets with parameters (λ_p, λ_s, and B_s) modeled as an external periodic on-off process with 'on' and 'off' periods as $T_{on} = B_s / \lambda_p$ and $T_{off} = T_{on} . (\lambda_p - \lambda_s) / \lambda_s$. The probability P that the source is on is given by P_{on}:

$$P_{on} = T_{on} / (T_{on} + T_{off}) = \lambda_s / \lambda_p.$$

The above process has been proven to be the worst-case source model for a bufferless multiplexer but not always in the buffer case. However, the model may still be considered as the worst-case source model. Let us look at a number of similar sources that share a buffer size B, which is serviced by a channel of variable capacity c. If the acceptable frame loss ratio (FLR) is P, then we define $\eta = \log (P/B)$ and the effective bandwidth for source (representing the bandwidth of all of the sources) is ϵ given by:

$$\epsilon = MRE (\Lambda - M * 1/ \eta)$$

Where MRE ( ) represent the maximum real eigen-value of a matrix

$$\Lambda = \begin{bmatrix} 0 & 0 \\ 0 & \lambda p \end{bmatrix},$$

$$M = \begin{bmatrix} -\alpha & \beta \\ \alpha & -\beta \end{bmatrix}$$

Based on the effective bandwidth, the admission criteria is satisfied if $\epsilon < c$ or violated if $\epsilon > c$.

Data traffic was observed to be self similar and traffic streams can be correlated with themselves over some period of time. Self similarity can be characterized by the long-range dependence of some abstract portion of a network statistics with the same abstract portion of the network statistics magnified over a period of time [51]. This type of traffic is of concern in meeting the service guarantees; however, to alleviate this problem, performance measurement probes are placed in the network for better monitoring, data collection for correlation and auto-correlation analysis, and better understanding of the traffic behavior generated at the source. This should help in building the appropriate control points across the network to reduce the impact of the traffic burstiness and to provide a fair distribution amongst the different users and their associated service classes. This is not an easy task, however, if we can characterize the source traffic flow based on historical behavior and performance, then we can at least ensure delay sensitive and priority based traffic are minimally impacted. Also, this should help us better understand the switch configuration requirements with the network. It is very difficult to calculate or predict the behavior of a source transmitting data by stochastic means [51]. In the next section, we will look at the performance architecture and data collection to enable us to build a better, efficient and more scalable network topology.

5.7 Performance monitoring architecture

Generally, performance monitoring (PM) is done in-band or out-of band. However, in our model, performance monitoring for a UNI-to-UNI is done in-band with test probes,

located at the customer demarc or at the ingress switch, generating test traffic that follow the same EVC logical path as the customer traffic. The architecture consists of:

- A probe at the customer premises co-located with a router or as part of the aggregation switch
- The aggregation switches (edge of a service provider network)
- Tandem switches (provide switching functionality in a metro or WAN, if any)

5.7.1 Performance monitoring test probe traffic definition

The PM test probe is configured to use the same VLAN ID, p-bit and DSCP values as those being monitored for the customer. The tag header is normally included in each of the 802.3 Ethernet MAC frames. The inclusion of the tag header in each packet serves the following purposes:

- Each frame can carry user-priority information
- Each frame can carry a VLAN identifier (ID)
- Each frame can indicate the type of payload carried in MAC user data
- VLANs can be supported across different payload types

A separate test profile is normally created for each Class of Service being monitored. All the frames in a given test profile have the same size, but the frame size can vary between test profiles, e.g., smaller frame sizes can be used to monitor performance in the real-time class of service, which is designed for packet voice applications. The header format for the VLAN tagged frame is the same as the customer traffic with the payload format specified in the probe section. The payload is a standard ICMP ping packet, with vendor-specific fields for adding up to four time stamps.

5.7.2 Measurement Methodology

This section describes how the PM will be implemented. It also provides a way to estimate the additional bandwidth requirement.

5.7.2.1 Test Methodology

The performance of the service will be approximated through the performance of the synthetic PM traffic injected into the data stream at the ends of the monitored network segment. The PM traffic is injected periodically into the network in a series of packet streams, called test probes. These test probes are generated into each CoS instance of each EVC that requires monitoring (i.e., each service instance that carries performance guarantees) at a given probe.

The measurement traffic has all the characteristics (VLAN ID, VLAN CoS) of the customer traffic and is treated as customer traffic by the network elements.

The PM traffic has several characteristics that have to be specified for the tests. They are:

- Number of frames in the probe
- Size of each frame (typically in Bytes)
- Time interval over which the PM test probes are spread

Note: Specific traffic parameter recommendations for the PM test probes are described later in this document.

5.7.2.2 Bandwidth Requirement for PM Test Probe

Based on the type of service and service guarantees, the test probe may inject s string of frames that represent the type of anticipated class of service. The frame length may vary

between 64 and 1500 bytes over a time interval. The bandwidth requirement for a given PM test probe is calculated from the following formula:

$$B = \frac{(L \times N \times 8)}{\Delta t \times 1000}$$

Where,

- B is the bandwidth, in Kbps, required for the PM test probe traffic
- L is the length of each frame, in Bytes
- N is the number of frames in a PM test probe
- Δt is the time interval, in seconds, over which the PM test probe traffic is spread

For example, let's assume that we are monitoring a priority data CoS instance on a given EVC, using the following test profile parameters:

- L = 512 Bytes (average data frame size)
- N = 100 frames (enough frames to closely monitor performance)
- Δt = 2 sec (selected so as to give an inter-frame gap of ~ 20 ms)

For this example, the bandwidth requirement is given by the following formula:

$$B = \frac{(L \times N \times 8)}{\Delta t \times 1000} = \frac{(512 \times 100 \times 8)}{2 \times 1000} = 205 Kbps$$

5.7.2.3 PM Test Probe Scheduling

In a customer's service domain, one or more PM probes are selected to generate the test string. A given master probe could have multiple services (EVCs and CoS instances) that need to be monitored. Scheduling control is required at each master probe to ensure that PM test probes for those multiple services are not scheduled at the same time.

The master probe will be able to schedule each test separately throughout the test period T, typically ~ 5-10 minutes, such that only one PM test probe will be performed in a given test interval Δt, typically ~1-2 seconds.

5.7.2.4 Accuracy of measurements

The accuracy of the measurements will be determined by the accuracy of the testing equipment. The test probe uses hardware time stamping, and so it is expected to provide accuracy on the order of 1ms, or better.

5.7.2.5 Scaling Issues

This architecture has to address scaling issues inherent in the proposed design. The following subsections elaborate on these.

5.7.2.6 Backbone Bandwidth Impact

As the number of services in a given network increase over time, the amount of PM test probe traffic also goes up. There is not, however, a direct 1:1 relationship between aggregate PM test probe bandwidth and impact on the backbone network. The backbone bandwidth requirement in any given time interval, Δt, is a function of the bandwidth requirement for each PM test probe (i.e., number and size of frames spread over the time interval), and the aggregate number of PM test probes firing simultaneously on a given trunk link, which in turn is dependent on network topology and communities of interest.

For the analysis, we assumed that all test probes would use a worst-case test profile that generates 400Kb of traffic on each of the 5000 test probes spread over a 2-second test cycle. This test profile corresponds to our recommendations for ERS-Silver. By comparison, other test profiles generate 80-160 Kb of traffic.

A backbone impact analysis tool was built to then study the impact of PM traffic on the backbone (see appendix C). The tool was used to simulate 5000 probes randomly starting up and then injecting test probes every 5 minutes. As expected, the distribution of these tests within a five-minute test cycle follows a normal probability curve, with less than 1% chance that more than 28 probes would burst together in the same second interval. Conservatively, we assumed 33 test probes firing concurrently in our impact analysis. The next step in assessing backbone impact was to determine the number of probes that would appear simultaneously on a given trunk link. For this part of the analysis, we assumed a three-tier network topology consisting of 20 edge switches, four aggregation switches and two core switches. It is estimated that out of a total of 5000 services, only 20% would coincide on the worst-case trunk link in the network. Taking these factors into account, the maximum impact of PM traffic on backbone bandwidth is on the order of 2 Mbps for a network with 20 edge switches and 5000 services.

5.7.2.7 Data Collection

PM test probe data is collected periodically by a data collection system that is centrally located – one data collection system can collect data from probes in many LATAs. As the test probe data increases in the network over time, the amount of data that needs to be collected by the data collection server increases. In general, each PM test probe cycle is expected to generate ~ 400 bytes file that is transferred to the data collection server over the network management VLAN. The server needs to be able to support very large storage capacity and significant file processing speeds. The following elaborates on each of the three key data collection components:

- Bandwidth impact on network management VLAN: Assume that each test cycle requires Ethernet frame size of approximately 500 bytes to carry each 400 bytes file. Furthermore, assume that 5000 test files need to be uploaded from the probes to the data collection server every test period (5 minutes), which are randomly distributed within each test period, then the bandwidth impact on the network management VLAN would be in the order of 120 Kbps.
- Server processing speed: Assume 5000 test probe files are needed to be uploaded every test period (5 minutes) and these are randomly distributed over that interval, then the server would need to process approximately 30 files per second.
- Server storage capacity: Assume that each test probe data needs to be stored for longer periods of time; we would require ~2GB of storage per month to accommodate the 5000 test probes running continuously over each 5-minute interval. Therefore, the annual storage requirement is approx. 24GB.

5.8 Latency Considerations

Latency is typically defined as the time elapsed from the first bit of a given service frame entering the network until the transmission of the first bit of the service frame at the egress of the network. The key components of latency for a given metro Ethernet service are: propagation delay, serialization delay and queuing delay. These are elaborated in the following subsections.

5.8.1 Propagation Delay

The speed of light in Single Mode Fiber (SMF) cable is assumed to be 0.65c, a conservative estimate for a variety of SMF deployed in the network. Propagation delay

(D_P) is defined as the one-way delay dependent solely on the actual diameter of the network, such that

$$D_p = .65c * L,$$

where D_p is the propagation delay (ms) and L is the diameter of the network (miles). Propagation delay could be significant for larger LATAs, some of which have network diameters approaching 200 miles.

Note: A rule of thumb is that $D_p \approx 1$ ms, for every 120 miles (200 Km).

5.8.2 Serialization Delay

Serialization delay D_s is defined as the time it takes to insert a given service frame into a given link. As such, D_s is dependent on the size of the frame and the speed of the link. It is calculated as follows:

$$D_s = \frac{F \times 8}{S},$$

where D_s is the serialization delay (ms), F is the frame size (Bytes) and S is the link speed (bps). For an example of calculating D_s in the SES network, let's assume that F is 1522B (maximum frame size allowed in SES) and the link speed is 1 Gbps (typical trunk link speed). For this case, D_s = 0.012 ms.

It is important to note that there is a serialization delay component for each link traversed across the network, and that slower speed links introduce higher serialization delays.

5.8.3 Queuing Delay

Queuing delay D_q is defined as the time elapsed from the moment a given service frame is put into a queue until the moment the frame is sent out of the queue on a given link. D_q is dependent on the amount of buffer space allocated to each queue, and the queue scheduling algorithm used to de-queue frames and the relative level of congestion on a given link. The maximum D_q can be calculated for queues using Strict Priority (SP) scheduling. For example, the queue we use for EVPL-Gold CoS is 120 KB and the SP queue always gets serviced first, hence, the maximum queuing delay for EVPL-Gold is in the order of 1 millisecond (ms) for a GigE link. However, calculations of the maximum D_q for the other classes of services (EVPL-Silver, E-LAN-Gold and Bronze) are not so simple. Each of these CoS is mapped into three different Deficit Weighted Round Robin (DWRR) queues, with weights assigned per CoS.

Each trunk link uses a combination of one SP and up to three DWRR queues. Backbone CAC is performed in the provisioning process, limiting EVPL-Gold to ~20% of link capacity. However, the CAC algorithm is based on long-term aggregate CIRs. Depending on the load at any given interval in time (say 10 ms), the short-term EVPL-Gold load can exceed 50% or more of the link capacity. Therefore, the available link capacity (i.e., drain rate) for the other (DWRR) queues becomes variable, thus making precise D_q calculation very difficult. Upper bounds can be determined based on statistics analysis, but these are not generally of practical use.

Practically speaking, a combination of Data Traffic Engineering (DTE) (used to monitor link utilization) and backbone CAC are used to indicate where and when to augment

backbone trunk link capacity to ensure that both loss and queuing delay performance objectives are met.

5.8.4 End-to-end Delay

End-to-end delay of a given service frame can be approximated with summing up the propagation delay, link serialization delays and queuing delays along the path. This calculation does not account for switch processing delays, which are typically minimal (in the order of 20 μs per switch).

$$D_t = \sum_{i=1}^{N} P_{gd} + \sum_{i=1}^{N} T_{ds} + \sum_{i=1}^{Q} Q_{dq}$$

where D_t is the total delay (one-way), N is the number of links along the path and Q is the number of queuing points along the path. Experimental measurements are presented in Appendix A.

5.9 CAC reference model

An admission control agent is required at each switch to ensure the configuration and inventory of available resources are up to date and to help determine whether a new EVC request should be admitted/granted without impacting earlier guarantees. An Ethernet frame scheduler algorithm needs to be implemented at each switch to enforce service commitments, such as delay and loss rate. In addition, a frame classifier is required to distinguish frames of different EVCs based on their specific information such as source and destination MAC addresses and VLAN tag. Hence, as frames arrive at a switch (ingress port), data is inputted into a queue, classified, admitted, and then forwarded into

the appropriate interface managed and scheduled and then transmitted over the Egress link.

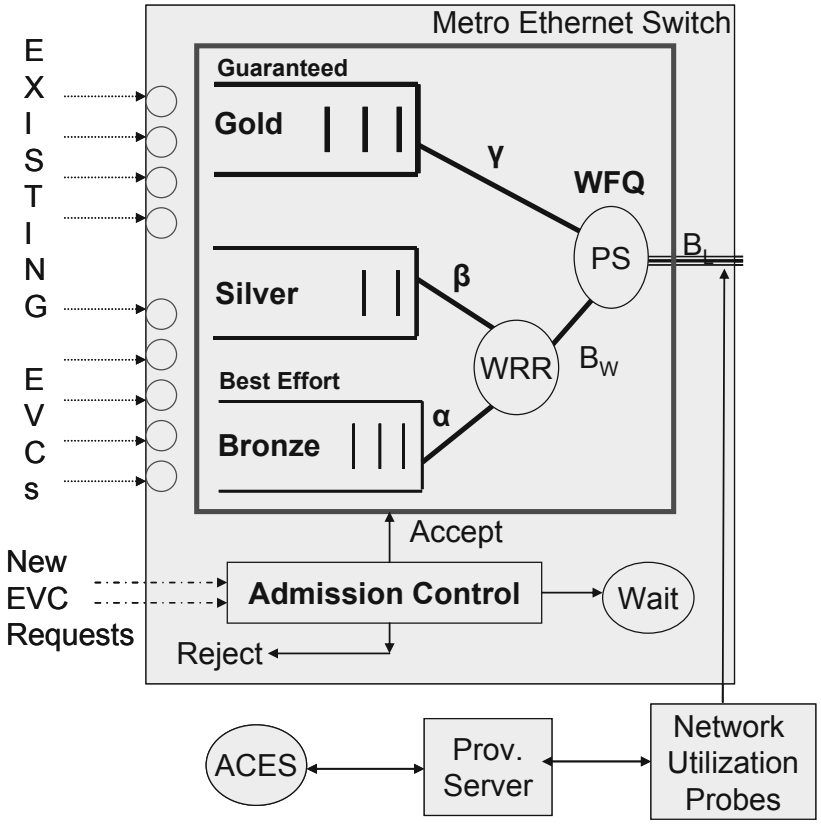

Figure 54: Switch Implementation Framework

In Figure 54, the admission control considers new EVC requests and determines whether to accept, reject or put into a holding wait bucket. The wait status depends on the time it takes the optical transport network to add capacity or the feedback from the utilization server is received allowing for the increase of the oversubscription factors for the service classes. The utilization server information contains data from performance measurement probes collected over a period of time.

In this chapter, the proposed ACES enhancement model was discussed; however, further research and work on appropriate signaling mechanism such as RSVP coupled with

Operation And Maintenance (OAM) input will validate and finalize our proposed model. This work is left for future research.

5.10 Summary

Proposals made by Bob Grow [22], David Martin [23], and Manoj Wadekar [58] to the IEEE congestion management group, May 2005, clearly suggest that the current 802.3 congestion management techniques are not sufficient to address QoS even for Ethernet frames that are colored ‘green’ and are conformant to the bandwidth service profile. However, with the implementation of suggested work, ACES can then signal the respective network element(s) to locally manage and control the traffic and prevent further congestion into the network and allow only high priority data as part of the EVC service bandwidth profile.

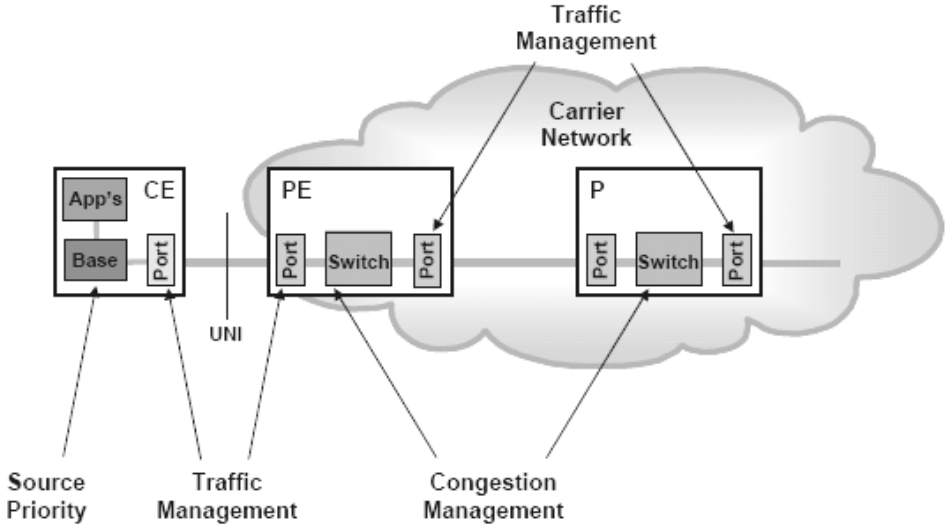

Figure 55: E2E Traffic and Congestion Management reference model

In Figure 55, the customer equipment (CE) is typically a router that is managed or un-managed by a service provider. The CE can be made to behave as a local CAC, which

limits the source traffic (since all traffic with priority gets queued at one transmit buffer) and at the port where clearly the MAC control sub-layer can benefit from seeing 802.1p priority and larger buffer size. In addition, ACES functionality can include traffic management rules and utilize feedback from real time network probes with congestion management rules. Thus ACES, with the collaboration of other network tools, can have the end to end view of an Ethernet connection and ensure meeting its service guarantees.

6. Long-term Vision

6.1 Overview

The main objective of this chapter is to present our long term vision on how to scale the Ethernet service into a global multi-service infrastructure and retain the simple characteristics of Ethernet. The advancement in fiber optics technology coupled with the new demand for metro Ethernet services have presented us with a unique opportunity to explore new networking designs. In addition, it is motivating a new networking paradigm based on simple layer-2 connectionless global communication that supports high data rate, converges multi-services, and supports bandwidth on demand provisioning. The ultimate converged network element will support packet and frame level services based on the service needs and characteristics specified.

Our long-term vision stage is based on Ethernet over DWDM with access, aggregation and core node layers utilize integrated L1/L2 management and control planes and GMPLS-like routing algorithm. We envision this model to scale a metro area to hundreds of miles and tie its core node to a national backbone network spanning thousands of miles. In addition, this model can support L2 only services or L2/L3 services. The proposed vision takes advantage of recent advances in optical communications to increase the capacity of a single fiber, by more than ten fold, utilizing wavelength-division multiplexing (WDM), switching speeds, and buffer management techniques.

The proposed model includes a simplified network model that:

- Integrates L1/L2 control planes with a CAC function for bandwidth on demand (BoD) service offerings

- Optimizes L1/L2 switches for CO deployment to maximize service availability and facility utilization
- Utilizes a model that exploits the huge bandwidth availability in the fiber by splitting it up into multiple non-overlapping wavelength channels
- Addresses redundancy, scalability, and network topology changes that ensure QoS service metrics are within the customer service SLA.

Specifically, this work proposes a truly native end-to-end layer-2 MAC frame-based "Optical Ethernet" infrastructure seamlessly stretching from enterprise LAN to metro to global networks. The proposed "Ethernet-over-WDM" model is truly a two-layer networking architecture model, where native Ethernet frames are mapped directly over WDM. It offers advantages over existing Layer-2 and MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other layer-2 protocols.

The key for realizing such an ambitious initiative rests on resolving the following five critical issues:

1. How to replace the legacy layer-3 switching (routing) and hierarchal IP addressing scheme with layer-2 switching and flat (non-hierarchal) MAC/VLAN-based addressing scheme?
2. How to provide comprehensive end-to-end Operations, Administration, Maintenance, and Provisioning (OAM&P) in a unified Ethernet-Optical environment?
3. How to realize a converged Layer 2/1 network model in terms of control plane and management functionality?

4. How to totally eliminate the reliance on ST/RST/MST routing and redundancy functionality?
5. How to reliably transport native Ethernet frames that have no overhead capability to perform network OAM&P across the WAN?

This proposal is implemented in a phased approach in support of our ultimate vision to implement a truly native end-to-end Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. The vision includes drivers that are: based on carrier grade Ethernet services, which are cost effective, simple to implement and efficient. Hence, the final architecture will lead to a scalable multi-service infrastructure that benefits a service provider in its deployment and service offerings. The primary rational behind our vision is decreased cost and complexity in the network. Simplifying network design and reducing costs by utilizing Ethernet as an end-to-end LAN/MAN/WAN protocol is the key for Ethernet to win the MAN and WAN. The outcome of this research will bridge the gap that exists between Optical Networking and Ethernet research communities. The hope is that this work will generate new interest in the research community to expand on the proposed model and further explore this important technology innovation.

The research outcome is a simplified network model that integrates L1/L2 control planes with a CAC function for BoD service offerings; thus enabling an optimized L1/L2 switching in CO deployment, which will maximize the service availability and facility utilization; in addition, it will utilize a model that addresses redundancy, scalability, and network topology changes that ensure QoS service metrics are within the customer service SLAs.

6.2 Ethernet over WDM model

The main characteristics of the proposed Ethernet-over-WDM model are:

1. A unified GMPLS control plane
2. A unified management plane

Conventional Ethernet MAC frames and/or jumbo Ethernet frames must be transported natively (translation into some other protocol is not allowed) end to end from the access network through the metro and core networks to access network.

Only pure layer-2, switching at the packet/frame granularity, is allowed throughout the entire network including access, MAN and WAN.

Unlike layer-2 MPLS VPNs (point-to-point and multipoint Virtual Private LAN Services (VPLS)), where a full mesh of static label switched paths (LSPs) must be set up between all L2 VPN sites, the proposed optical Ethernet is dynamically reconfigurable network that supports real-time additions/deletions of all customer connections (EVCs). It can also support a fully automated optical networking service (layer 1) at any bandwidth granularity.

Supports an IP/GMPLS-based unified control plane that offers a tighter integration between layer-1 (optical transport layer) and layer-2 (Ethernet layer), leading to the collapse of the two layers into a single integrated layer managed and traffic engineered in a unified manner. The unified control plane supports real-time provisioning and restoration of both full lambda and EVCs by running a single instance of an integrated routing and signaling protocols (use of ST, RST, and MST routing are totally eliminated).

Native Ethernet frames are routed across the MAN/WAN using only layer-2 addressing scheme (MAC and/or VLAN ID).

6.3 Implementation Strategy

It is important to emphasize from the outset that there are strong analogies between the IP-over-WDM interconnection models and the proposed Ethernet-over-WDM model. Anyone who has followed the development of IP-over-WDM interconnection models (the overlay and peer models) throughout 1990s can easily observe that most of the initial problems encountered in the development process were mainly due the optical network-Internet (IP) gap. A partitioning between the optical networking and the IP/MPLS research communities caused this gap. It took the industry, the standards bodies, and the two research communities nearly ten years of extensive continuous hard work and collaboration to narrow this gap. Likewise, we strongly believe that the main problem that will hinder the viability of implementing the vision of a global optical Ethernet infrastructure is the wide gap that exists between the optical networking research communities and the Ethernet communities. It is our expectations that the proposed research program would be an important starting point to bring the two communities together, leading eventually to bridging this gap and the realization of the proposed vision.

Now we have the opportunity to reapply a lot of this technology, suitably modified to Ethernet technology. Our strategy is first to take full advantage of the knowledge and developments gained during the past ten years course of developing the IP-over-WDM interconnection models. The key to a successful strategy rests on taking the best features from both the overlay and peer models while avoiding their limitations. Specifically, it is

imperative, when devising the proposed integrated L2-L1 control plane that will manage both GigE and optical switches, to avoid the major limitations of the peer model's integrated control plane, particularly the scalability problem. Previous attempts to address the practical feasibility of implementing the peer model and its integrated control plane failed in large part due to the complexity of the model and the edge router scalability problem.

6.4 Lessons Learned From IP-Over-WDM Interconnection Models

Several industrial organizations Including the Optical Interworking Forum (OIF) and the Internet Engineering Task Force (IETF) have already proposed several architectural options on how IP routers must interact with the optical layer to achieve end-to-end connectivity, including overlay, augmented, and peer-to-peer models (interconnection models) [97], [98], [59]-[62]. The simplest is to treat the optical layer as completely separate from the IP layer. In this "overlay" model, the optical layer provides point-to-point connections (lightpaths) to the IP domain. The client routers request high-bandwidth connections from the optical network, via some User-to-Network Interface (UNI), and are provided with no knowledge of the optical network topology or resources. A more sophisticated model that offers a tighter integration between IP and optical layers (peer model) collapses the two layers into a single integrated layer managed and traffic engineered in a unified manner.

The overlay model is the most practical for near-term deployment because it is appropriate for the current telecommunications infrastructure that consists of multiple administrative domains, where there is clearly a need to maintain topology and control isolation between the optical transport and the service layers since they are most likely to

be under different administrative controls and policies. However, the simplicity of the overlay model comes at the expense of the inefficient use of network resources due to information hiding at the domain boundaries.

Unlike the overlay model, the peer model supports an integrated IP/MPLS-based control plane that manages both routers and optical switches [62].The peer model does, however, present a scalability problem due to the amount of state and control information to be handled by any network element within an administrative domain. This means that a significant amount of state and control information must flow between the IP and optical layer, making the development of this model more time consuming and complex. Compounding the problem is the fact that it is highly unlikely that service provider who owns the optical transport network (OTN) would ever want give a client full access to the topology and resources of the optical network. The peer model may well take hold in the future, but it is likely to be a rather distant future.

It is clear from the above discussion that the viability of the proposed unified control plane of the Ethernet-over-WDM model rests entirely on avoiding the two major obstacles that hindered the practical implementation of the integrated control plane of the IP-over-WDM interconnection peer model. Specifically, the following are the two lessons that will guide the process of devising a unified L1-L2 control plane:

The edge IP/LSR router of the integrated control plane of the peer model is the device that manages all network resources including both the physical and layer-3 logical resources. Thus, the edge routers are choked by constant barrage of network state updates and optical network topology and resources, leading to a major scalability problem.
Lesson 1: when devising an integrated control plane that manages both GigE and optical

switches, never delegate this task to the GigE switches. If the smart IP/LSR router could not make it; how do you expect that GigE switch along with its primitive ST to make it?

Since the boundaries between the transport network and data network are impenetrable and since it is highly unlikely that routers or GigE switches (especially those owned by a customer) would have the ability to "see" the topology and resources of the optical network and make changes. Lesson 2: topology isolation between the optical transport and the service layers (layer-2) must be maintained. Devise an optical layer-based unified control plane that manages both GigE and optical switches (analogous to the peer model), while still retaining the client/server relationship with the network (the customer has no network visibility and depends on network intelligence) and the simplicity of the optical UNI of the overlay architecture.

6.5 Implementation Approach

Current Switched Ethernet Services are based on a single Spanning Tree topology and a single homing core switch that also provides aggregation for all interlata EVC traffic. Hence, there are several short-term and longer term strategies that can address a provider's network immediate service concerns and scalabilities.

To simplify the implementation approach, we decompose the overall problem into three phases aimed at solving the issues and questions presented in section 2.

In our proposed model, the first phase addresses issues 1-3, the second phase addresses issue # 4, and the third and last phase will specifically address issue 5.

6.6 The First Phase

The main objective of this section is to address most of the known shortcomings and concerns associated with transporting native Ethernet frames across the optical layer. The notion of supporting native Ethernet directly over optics ”Ethernet-over-WDM”, which is a truly two-layer model, is driven by the promise that elimination of unnecessary network layers will lead to a vast reduction in the cost and complexity of the network. However, the important functionality provided by the intermediate layers (traffic engineering in ATM, routing/restoration in IP/MPLS, and multiplexing and fast restoration in SONET) must be retained in the proposed “Ethernet-over-WDM” two-layer model. Specifically, since there is no IP/MPLS layer to provision/restore sub-lambda connection requests (LSPs) at the logical layer; and that Ethernet alternative, ST/RST, does not scale and represents a major bottleneck for real-time provisioning/restoration of EVCs.

Rather than pursuing the conventional approaches that have always addressed existing layer-2 limitations by introduce further Enhancements to existing ST/RST/MST protocols, our strategy is to deviate from the conventional approaches by alleviating the Ethernet layer from its current routing and restoration functionality, i. e., to do away with STP and its derivatives MSTP and RSTP altogether. In this case, and since there is no any intermediate layers (IP/MPLS/ATM/ SONET), these functionality, along with most of all other networking functionalities and intelligence (that would have been performed by the presumably absent intermediate layers), must now be passed on to the optical layer (the only available option). This would require three optical networking innovations: 1) a fully intelligent and agile optical transport layer; 2) a novel hybrid optical node architecture; and 3) an integrated control plane that manages both layers (later-1 and

layer-2), which must be owned by the optical layer rather than by the Gig E switches (or by the IP/MPLS routers as it is the case in the peer model).

6.6.1 The Envisioned Optical Ethernet Architecture

In the proposed networking architecture, see Figure 56, the customer GigE switches (might be owned by the service provider, as part of a managed service, who also owns the OTN, e.g. provider edge (PE) switches) are attached to a fully intelligent optical core network. The optical network consists of multiple hybrid optical nodes interconnected via WDM links in a general mesh topology. The edge GigE switches are clients of the optical network and are connected to their peers over dynamically switched lightpaths spanning potentially multiple optical nodes.

6.6.2 A Hybrid Optical Node Architecture

In order to efficiently utilize the capacity of each wavelength channel (lightpath), the traffic of several independent lower-speed EVCs must be multiplexed into a single lightpath. The process of combining low-rate traffic streams onto high-capacity optical channels (lightpaths) is known in the literature as “traffic grooming [111]-[113]. To support traffic grooming, the cross-connect fabric of each optical node should have the capability of switching traffic at the wavelength granularity as well as at finer granularities. Therefore, a hybrid switching solution that capitalizes on existing electronic switch fabrics (GigE switches) as well as all-optical switch fabrics, by making use of each switch’s functionality and capability, appears to be the most appropriate for building the optical nodes of the proposed Optical Ethernet.

The node architecture of the proposed model is composed of four key modules:

1. **All-optical switch fabric (Optical Cross-Connect (OXC)):** Performs pure optical switching without wavelength conversion capabilities where the granularity of switching is the entire wavelength.
2. **Backbone electronic-switch fabric (GigE switch):** Capable of multiplexing, de-multiplexing, and switching low-speed traffic streams (EVCs) onto the wavelength capacity. The backbone GigE switch is attached to the optical switch fabric through an array of transceiver and can generate and terminate the traffic to or from a lightpath. The number of wavelength channels that can be terminated or generated into or from the electronic switch is a function of the transceiver array size.
3. **OXC-Controller:** A non traffic-bearing IP/MPLS-based intelligent control plane module managing both optical and logical domains.
4. **Transponders:** an array of transceiver ports connecting the GigE switch to the OXC, limiting the number of wavelength channels sourced and sinked at the optical node.

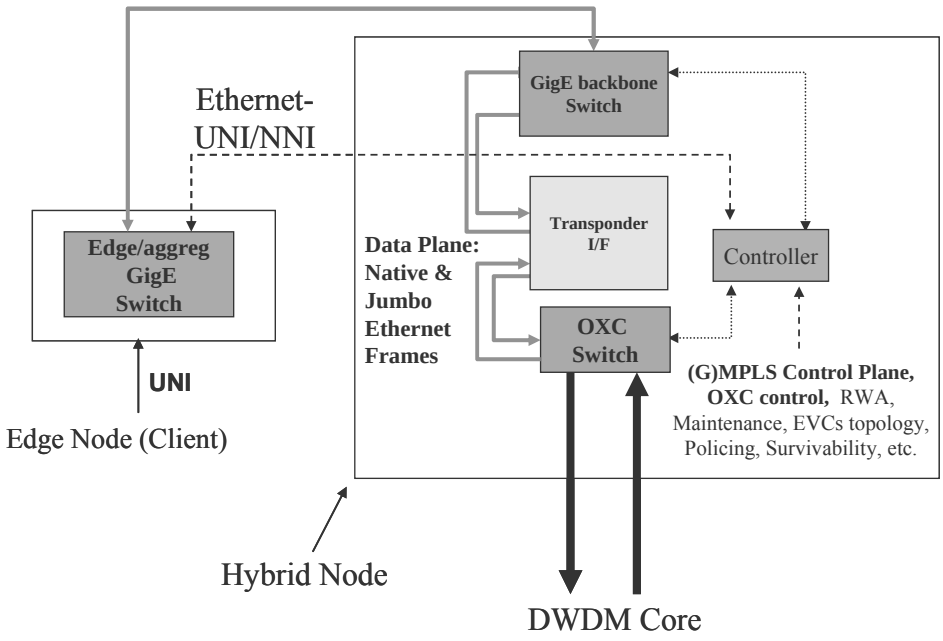

Figure 56: Proposed hybrid node model

6.6.3 A Fully Intelligent Agile Optical Networking Layer

To realize the "ultimate vision" of an agile, fully intelligent optical networking layer capable of supporting integrated routing and signaling algorithms for real-time provisioning/restoration of EVCs at any bandwidth granularity (on a per-call basis including both full-lambda and sub-lambda traffic flows), the following two salient features must be implemented [114]:

1. Most of the networking functionalities and intelligence must be migrated down to the optical layer including switching, protection, traffic engineering, OAM&P, provisioning of both full lambda and sub-lambda connection requests, and selective restoration (differentiated resilience for different classes of service), all supported entirely on the optical layer's terms

2. The optical layer must own and manage both the physical connectivity and resources (layer-1 optical resources) and logical connectivity and resources (layer-2 Ethernet resources). Thus, both the logical and physical topologies now belong to a single administrative domain managed and controlled by the optical layer, leading to the creation of a unified control plane with the optical layer running a single integrated routing/signaling protocol instance.

6.6.4 A Unified Control Plane GigE and Optical Switches

To implement the proposed optical layer-based unified control plane, layer-2 control plane functionalities must now be shifted to the non-traffic bearing IP/MPLS-based OXC controller modules located within the optical domain. Under this scenario, the OXC controller, in addition to its conventional functionality of managing the optical layer resources including full wavelengths and fibers on physical links, must now support databasing and provisioning of finer granularity sub-lambda connection requests (EVCs) as well. This is achieved by simply augmenting the OXC controller with an additional resource usage database for sub-lambda connection requests. Thus, each OXC controller is assumed to maintain a single topology and resource usage database for both layer-1 and layer-2.

The OXC controller is now responsible for creating, maintaining and updating both the physical and logical connectivity tables. Thus, the OXC controller performs all complex functions, including addressing, routing and global topology discovery (both physical and logical topologies), and is responsible for network optimality including traffic engineering and QoS. The responsibility of the edge GigE switch (owned by client or

provider) is then simply to request a service from the OTN (both full-lambdas and sub-lambdas) and the latter is responsible for providing this service.

Each node maintains a representation of the state of each link in the network. The link state includes the total number of active channels, the number of allocated channels and the remaining bandwidth capacity of each. Once the local inventory is constructed, the node engages in a routing protocol to distribute and maintain the topology and resource information at both the physical and logical (layer-2) layers. Standard IP routing protocols, such as Open Shortest Path Forwarding (OSPF) or Intermediate System-Intermediate System (IS-IS), may be extended to include additional information regarding the optical link state.

Each client/provider edge GigE switch registers its unique VLAN ID with the corresponding attached OXC-controller. The OXC controllers then run an IP routing protocol amongst themselves to determine all VLAN IDs destinations (of all registered edge switches attached to the core network) reachable over the optical network.

The OXC controllers are assumed to communicate with each other over a data communications network (DCN). Using either out-of-band or in band signaling, the DCN will serve as the packet transport network for all the signaling messages required for connection set-up and tear-down in the optical transport network. Thus, data plane is native Ethernet *without* encapsulation and is run between backbone GigE switches, while the control plane is (G)MPLS-based and is run out-of-band between controllers. Note that the two planes are *completely* segregated.

Two comments are in order here. First, note that layer-2 is neither aware of, nor does it maintain any topology and/or resource usage information about the optical layer or its

own layer. Furthermore, there is no exchange of information between the boundaries of the two layers except for that of the simple UNI. Thus, shifting all the networking functionality and intelligence down to the OXC controllers eliminates the major two problems that hindered the practical implementation of the peer model's integrated control plane.

Second, it is important to emphasize that the immediate/near-term implementation of the proposed optical layer-based unified control plane is highly feasible. The main reason is that many carriers are migrating away from their current dominant ring-based core transport network architecture to an electronic cross-connect (EXC)-based mesh architecture, where each of these EXCs is equipped with point-and-click provisioning capability that allows the OTN carriers to provide dynamic optical connections to the clients attached to their core network. This means that EXC vendors are currently supplying each EXC with an intelligent controller. Each controller must maintain the topology and resource usage information database of the optical layer. Given the "existence proof" provided by these EXC networks, it is rather a simple logical transition for the OXC controllers of the proposed model, in addition to their lambda-level management, to support provisioning and databasing of finer granularity sub-lambda connection requests as well, since the incremental cost for doing so is likely to be very low.

6.6.5 Frame Size Limitations

Extending Ethernet's frame size from the traditional 1518-bytes to a 9000-bytes jumbo frame has been a topic of hot debate [114], [115]. With the explosive growth of Gigabit Ethernet, the impact of this decision is critically important and will significantly affect

next-generation optical Ethernet performance. The choice of 9000 bytes for the Jumbo Frame payload length provides a good compromise between frame efficiency, frame check sequence effectiveness and host protocol stack efficiency. The case for Jumbo Frames is based upon a simple premise- larger packets mean lower frame rates and, therefore, better performance. Independent tests have verified that the use of Jumbo Frames can deliver a 50% increase in throughput combined with a 50% reduction in host CPU processing [115]. Most of the debate about jumbo frames has focused on LAN performance and the impact that frame has on host processing requirements, interface cards, memory, etc., [97], [98]. The impact that frame size has on MAN/WAN performance has not received any attention. With Ethernet operating at an aggregated high bit rate (multi-Gigabit) in the WAN environment, a small frame size might be an issue. It has been shown that maximum TCP throughput is directly proportional to the Maximum Segment Size, which is frame size minus TCP/IP headers. We will examine the impact of utilizing traditional Ethernet frames versus jumbo frames on the overall end-to-end performance of the proposed Optical Ethernet. Performance metrics include network throughput, end-to-end delay, jitter, for each of at least three different CoSs.

6.7 Second Phase: E2E OAM in a Unified Ethernet-Optical Environment

6.7.1 Overview of Ethernet OAM Mechanisms

Currently, native Ethernet has no carrier grade management capabilities that would allow Ethernet to report network behavior at Layer 2. Ethernet OAM provides tools to monitor and troubleshoot an Ethernet network and quickly detect failures. For native Ethernet, the

only alternative, in the absence of Ethernet OAM, is expensive and time-consuming diagnostics by technicians sent into the field.

Ethernet links and networks can be managed via higher-level management protocols such as SNMP. SNMP is an application-layer protocol that allows the exchange of management information between network devices. A Network Management System (NMS) executes applications that monitor and control managed devices. While SNMP mechanisms are adequate to determine the status of a link or equipment, SNMP cannot verify the end-to-end connectivity of the service delivered to customers or an Ethernet Virtual Connection (EVC). Any NMS-based troubleshooting can poll only the individual nodes and report the status of the EVC. This polling is slow, and it would take a lot of work to correlate the defects at different layers. Therefore, end-to-end network management for EVCs depends on the presence of OAM tools in the data plane and the control plane to verify continuity, connectivity, and performance of the EVC.

In general, OAM has two components: Connectivity Fault Management (CFM) and Performance Monitoring. Performance Monitoring is the ability of the network to predict EVCs performance metrics such as delay, loss, and jitter that reflects the status of customer's SLAs. End-to-end CFM is the ability of the network to monitor the health of an end-to-end service delivered to customers as opposed to just links or individual switches/bridges. Any Ethernet OAM mechanism must:

1. Monitor the health of the links
2. Check continuity and connectivity of EVCs and detect fabric failures and mis-configurations

3. In a centralized architecture, communicate with the NMS, so that the NMS can take corrective action in the event of a failure
4. In a fully distributed architecture, as the one proposed here, report the failure to the source node, so that it can take corrective action.

ITU-T SG13 and IEEE 802.1ag are currently standardizing Ethernet OAM under the names of "OAM Functions and Mechanisms for Ethernet based networks" (draft Rec. Y.17ethoam) and "Ethernet Connectivity Fault Management (CFM)" (draft IEEE 802.1ag v3.0 April 2005) for provider bridged networks [115], [116]. Mechanisms supported by 802.1ag include Connectivity Check (CC), Loopback, Link trace and Alarm Indication Signal (AIS). CFM allows for end-to-end fault management that is generally reactive (through Loopback and Link trace messages) and connectivity verification that is proactive (through Connectivity Check messages). Additional mechanisms supported in Y.17ethoam and G.ethps include Remote Defect Indication (RDI), Loss Measurement (LM) and Automatic Protection Switching (APS).

6.7.2 Proposed End-To-End Ethernet OAM Mechanisms

Since the proposed Optical Ethernet supports an integrated control plane that manages both the optical (layer-1) and Ethernet (lyer-2) layers, control coordination and fault handling among network elements with different technologies are now simplified. Furthermore, since a single layer "the optical layer" owns both the physical and logical connectivity and network resources, OAM functionality can now be implemented solely at the optical layer. We will examine two different approaches to implement in-band and out-of-band OAM functionality.

In the first in-band approach, we will take advantage of the fact that the framing overhead (header) size of Ethernet Jumbo frames have not been standardized yet and still an open issue. Specifically, we will investigate how some SONET-like OAM functionality can be embedded within the jumbo frame overhead bytes, while taking into account that unlike SONET frames, Ethernet frames are variable size frames. Note that in contrast to SONET OAM functionality that serves only the physical layer, OAM functionality embedded within the jumbo frames will serve both the physical (layer-1) and logical (layer-2) layers, because the optical layer owns both. This is a challenging task that requires close interaction with vendors and service providers in forum like OIF.

In the second out-of-band approach, we will examine how to integrate OAM Ethernet mechanisms supported by IEEE 802.1ag into our proposed architecture. Specifically, we will assess the pros and cones of implementing CFM in bridges through a dedicated ETH CFM ethertype versus implementing it in OXCs-controllers. Ethernet CFM enables a provider to detect and/or accurately locate a link fault, a port failure or a fabric failure/misconfiguration. Fault monitoring and quick identification of fault location is done by sending Ethernet CFM frames along the data path of the Ethernet service provided to customers. In particular, loss of specific periodic frames (Connectivity Check) that ride along the data path would indicate a connectivity failure. Note that the optical layer of the proposed Optical Ethernet, as described above in section 5.1.1 above, can detect and restore link, fabric/switch, and port failures. This will simplifies the implementation of an end-to-end OAM mechanism and functionality.

6.8 Third Phase: Scalable Global layer-2 MAC/VLAN Address Structure

This task will look at Ethernet as a converged network with large address space to serve the global needs with a new hierarchal address plan that is unique and scalable. This implies that each frame will have a source and destination address in its format that are unique. Address resolution is based on ARP, and address resolution servers are distributed across the network to support address resolution. This section examines two different approaches. The first approach utilizes existing technology and IP capability, while the second approach proposes and utilizes new technology independent of IP capability.

6.8.1 Global MAC/VLAN-based Addressing Structure Utilizing IP Capability

Ethernet Switching includes a basic concept called MAC address learning. This concept allows a switch port to learn the MAC addresses of the hosts or devices connected to it so that the switch can make a decision as to which port it needs to send the traffic onto. In order to accomplish this, a switch maintains association tables: a MAC/port or in the event VLAN is enabled, a VLAN/MAC/Port, as shown in Table 18 below.

Table 18: Switch VLAN/MAC/Port association

VLAN	MAC	Port
VLAN 100	1	X
VLAN 100	2	Y

There are three types of Ethernet services: point to point (P-P), point to multi-point (P-MP) and multipoint to multipoint (MP-MP). For P-P, no MAC address learning is required since VLAN switching is used. For P-MP (E-Tree) and MP-MP (E-LAN), MAC

address learning is required for unicast and multicast switching. Further research is needed in covering E-Tree and E-LAN services.

6.8.1.1 Switching Domains

6.8.1.1.1 VLAN Domain

Devices attached within one VLAN domain can communicate with each other. Currently, traffic between VLANs is only routed using IP routers and is not switched. For switching within a single domain, the switch looks up the destination MAC address of the packet and makes a forwarding decision as to which port to send the frame on within the domain. However, if the destination MAC address is a unicast (i.e. unknown address), or a multicast, or a broadcast address, then the Ethernet frame is flooded on all the ports within the VLAN.

6.8.1.1.2 VLAN-VLAN Domain

Switching between multiple domains is currently done in IP and MPLS routers. Switching between domains may require MAC layer re-write of the destination MAC address. Hence, an originating system with source address SA will send its packet to a destination address DA of the default switch MAC address. The switch, in turn, will receive the packet and re-write the source address to be its own MAC address and destination to be the destination MAC address of the server or the application host.

6.8.1.2 Current Ethernet Addressing Plan Scalability Issues

The current Ethernet address space scalability issues include two key factors: MAC and VLAN address spaces. 1) MAC address is a unique (48 bits long) but is flat. This means that any MAC address can be anywhere in the world (with some exceptions). This results in its inability to be summarized thereby putting pressure on the content addressable

memory (CAM) table size in a switch and forwarding wire-speed. This inefficiency requires routers and switches to maintain a large table lookup due to unique source and destination MAC addresses. 2) The current VLAN address space provides a good segmentation of traffic within a large campus LAN but is limited to approximately four thousand addresses (4096). This VLAN address space is not sufficient for a service provider with a large network.

In addition, the scalability issues expand beyond the MAC and VLAN address spaces to the actual service offered within a service provider and include the service profile and bandwidth profiles across customers' applications and needs. The multi service provider interoperability networks (MSPIN) include: availability, interoperability, bandwidth and traffic profiles across providers. These service related issues are resolved within a provider when adapting and implementing a standard UNI and NNI service definitions and profiles such as those defined by the metro Ethernet forum (MEF) specifications. These definitions need to be interoperable across islands in a service provider environment or across multiple service providers serving customer needs over a large geographic areas and include:

1. Service profile(s) interoperability within and across service providers
2. Bandwidth profile(s) interoperability within and across service providers

6.8.2 Global MAC/VLAN-based Addressing Structure Utilizing New Technology Independent of IP Capability

6.8.2.1 Global Ethernet Address Plan

There is a great need to further explore through additional research the end-to-end layer-2 addressing plan and to develop an Ethernet Label Switching (ELS) mechanism that will

scale globally. This addressing plan must be unique and hierarchal and provide switching per VLAN ID uniqueness per site. In addition, the addressing plan needs to support existing GMPLS control plane and Ethernet services. In particular, the solution must be interoperable with the IEEE 802.3 frame structure. The new addressing plan can incorporate key features in link state routing protocols such as intermediate state to intermediate state (IS-IS), open shortest path first (OSPF), and generic multi-protocol label switching (GMPLS) to provide the ability for the network to learn about the network elements and associated link characteristics and to build the network topology database. Specifically, the addressing plan must:

- Scale and handle large MAC address table size. Switches need to learn the MAC address of each device node attached to it and associate it to a VLAN for switching.
- Distinguish between different enterprise customers internal VLAN address spaces. This is partially alleviated by use of stacked VLAN address space but the use of VLAN stacking does not resolve the MAC address problem in multipoint Ethernet services such as E-Tree and E-LAN.

6.8.2.2 Proposed Options

The proposed options based on existing technology do not require any changes in the network elements or Ethernet protocols. The proposed options include the use of a cache within the switch with a pre-configured VLAN value or the use of a lookup server for a dynamic lookup– similar to that of a DNS hierarchal structure. One option, which is inefficient, requires address management via a central master server per customer site and supports the following:

- The use of structured stacked VLAN Tags to provide up to 64 billion addresses
- The ability to broadcast within a VLAN tag address space – This is more complex since it requires a new address plan similar to how NPA-NXX works globally. This option is tightly coupled with switched virtual circuit (SVC), that can setup and tear down a connection dynamically, which minimizes the complexity in keeping states in the network
- The ability to configure each site with the VLAN addresses of the entire customer destination sites on the gateway router either statically or dynamically, which is very complex to manage. This requires vendor community support and is not scalable. This can be implemented dynamically but initially is static
- A single administrative domain (the optical layer), which is responsible to keep both logical and physical topologies; however, integration can result in better resource management
- Optical Signaling to overcome challenges in provisioning fine (sub-lambda) granularities along with whole wavelength channels

6.8.2.3 Other Proposed Options

Further research is needed to refine the truly global addressing plan and make it more efficient. The research will ensure that the addressing structure is widely accepted. Among the MAC and VLAN addressing schemes, the suggested options include:

1. MAC Address space

    - The Universal Ethernet Telecommunications Service (UETS) [119] proposes to use the U/L value of one to specify a local LAN MAC address and a U/L value of zero to specify a network address used with a switching netmask indicator

    - MAC in MAC

      Utilize MAC in MAC (VPLS) to encapsulate customer MAC addresses at the provider edge switch to minimize the number of MAC addresses learned by the network

2. VLAN Address space

    - Utilize IEEE 802.1q with multiple VLAN tags (stacked or cascaded tags: double, triple, etc) to increase the VLAN address space

    - Increase the IEEE 802.1q VLAN ID field

In addition, STP needs to be enhanced or eliminated as the default Ethernet routing algorithm. The proposed options include:

- ACES system short term and long term options discussed in this thesis

- Ethernet Label Switching with IS-IS (neighbor discovery, track cost per CoS, compute paths using SPF, and GMPLS to signal and to map Ethernet onto an optical bandwidth channel)

- RBridges within a customer LAN environment in coordination with ACES

6.9 Routing and Switching of Ethernet Frames Across MAN/WAN

In routing and switching of Ethernet frames across the MAN/WAN, the following assumptions are made:

- Uses of cascaded VLAN tags are allowed for both clients and service providers. To distinguish the subscriber's VLAN tag from the provider VLAN tag, the MEF has defined the term CE-VLAN ID (Customer Edge VLAN ID) to represent the subscriber's VLAN ID
- Client tag (CE-VLAN ID): IEE802.1q VLAN tag(s) to identify the access ports to the optical network (i.e. addressing). Use of two cascaded tags or more for the client will pose no scalability issues to the number of users served by the carrier
- Each client site that is attached to the OTN must have a unique VLAN ID

Service Provider tag: IEE802.1q VLAN tag(s) to identify the next lightpath (i.e. optical layer label). Each provider VLAN tag corresponds to a color/lightpath.

Frames routing and switching across the core network are performed using the following key steps:

- Customers VLAN tagged frames are presented to the ingress nodes of the provider's core network. The core's ingress nodes insert a stack of labels (provider VLAN tags) into each frame and are stripped off at the egress nodes before the customer's frames are handed over to the appropriate customer equipment (CE). The label stack is organized in a last-in/first-out order; forwarding decisions are based solely on the top label in the stack

- Each inserted provider VLAN ID corresponds to a color/lightpath, and the numbers of cascaded VLAN IDs are equal to the number of lightpaths (colors) transporting the frame. These fields have local significance within the carrier's backbone domain

For instance, when an ingress node receives a request to provision an EVC, it first determines the route and allocates required resources along the chosen route. Then, assuming that the EVC will be transported over a newly created lightpath (say red lambda), the ingress node's OXC-controller assigns a unique VLAN ID to the lightpath (whose ID corresponds to the red lambda), which it communicates to the ingress and egress nodes. The ingress node then attaches the VLAN tag to each jumbo frame in the EVC. At the egress node, this VLAN tag is stripped and since there are no other tags in the stack, the frame is dropped to the local destination of the customer VLAN site.

The VLAN tags (electronic labels) have meaning only to the GigE switch that performs frame-by-frame switching based on it. Since the OXC controllers carry out the entire control plane, it is their responsibility to perform label assignment and label distribution by executing the appropriate protocols.

Based on the color of the outermost label (provider VLAN ID) in the stack encapsulated within each frame, the GigE switch performs layer 2 switching of transit and/or generated/terminated frames as follows:

- A single provider tag means, "Drop to the local attached Client VLAN site".
- Two tags mean, "Transport those frames on a second lightpath (based on the color of the outermost label in the stack)".

A signal is either switched from an OXC input port to an output port without undergoing OEO conversion (express), or terminated at the GigE switch and converted into an

electrical signal. If it is converted into an electrical signal it can be either dropped, if this is its final destination (assuming there is a single tag in each frame), or statistically multiplexed with another traffic stream onto a second lightpath (based on the color of the outermost label in the stack and assuming that there is more than one label in the stack) and sent out at the corresponding output port.

6.10 Optical Layer-Based EVCs Restoration

In today's Ethernet-over optical metro networks, protection/restoration may be provided at layer-2 and/or the optical layer. To provide restoration at layer-2, STP is used to provide loop free connectivity while providing an alternate path if a link, port or switch fails. Classical STP implementation detects failures using a keep-alive mechanism based on hello packets and recovers network connectivity within a minute. We can run one STP instance per each VLAN in order to utilize all the fiber in the Metro network. However, as the Metro networks grow and scale to accommodate more customers, the number of STP instances and span of the STP instance can become a bottleneck. Rapid Spanning Tree protocol (RSTP) (IEEE 802.1w), builds upon the original 802.1D STP standard, and was introduced to achieve faster recovery time. However, most implementers remain convinced that RSTP recovery is rarely much below the one second mark—far slower than the carrier standard.

With optical layer restoration, a significant number of failure scenarios including fiber cable breaks and transmission equipment outages can be detected and recovered much faster than ST rerouting (e.g. 50 ms compared with tens of seconds). Optical layer restoration requires additional optical layer mechanisms and policies, and leads to conflicts with fast restoration schemes at higher layers (i.e. layer-2 or IP/MPLS),

requiring that recovery mechanisms in different layers be carefully coordinated. Complex escalation schemes [117] coordinating the activities of the different recovery schemes have been proposed to address this problem.

Two major obstacles would hinder the implementation of current optical layer-based restoration schemes to future Optical Ethernet networks. First, different classes of emerging Ethernet services need varying degrees of resilience requirements. However, the optical layer can only provide coarse granularity restoration at the wavelength level. Thus, the coarse granularity of protection/restoration would lead to costly, inflexible and inefficient resiliency. Second, since the optical layer operates independently and has no awareness of a switch/router failure, theses failures and certain other failures cannot be recovered by optical protection/restoration. This means that protection against GigE switch/router failures must be provided directly at the Ethernet/IP layer.

Our proposed Ethernet-over-WDM networking paradigm addresses the major shortcomings of current optical layer-based restoration schemes by utilizing optical layer-based unified control plane architecture. The proposed integrated IP/MPLS-based control plane simplifies control coordination and fault handling among network elements with different technologies and allows a single layer “the optical layer” to mange network resources and topology information of both layers (Ethernet and optical). In this architecture, each IP/(G)MPLS-based OXC controller keeps a single updated topology and resource usage database for both the physical and logical layers. Thus, the optical layer can now independently restore all disrupted EVCs on a per-call basis at any bandwidth granularity (including both full-lambda and/or sub-lambda EVCs). The proposed all optical restorations strategy avoids interoperability problems caused by

vendor specific solutions. Utilizing a single layer (optical layer) to manage and control all network resources, the proposed all optical restoration strategy avoids the complexity of having to coordinate restoration policies between different layers.

6.10.1 GigE Switches and Physical & Logical Link Failures

The all-optical resilience strategy proposed here is capable of protecting against link failure (both physical and logical) as well as GigE switch failures and is based on fast restoration techniques where the recovery path is established dynamically after the detection of a failure.

6.10.2 Backbone/Edge (PE) GigE Switch Failure:

In the case of a backbone/edge switch failure, the physical layer can independently restore all the disrupted traffic streams originating/terminating at the failed GigE switch (assuming dual-switch architecture), as well as transit traffic. In this case, all lightpaths originating or terminating at the failed switch will be immediately released so that the resources can be made available for future connections. The OXC controller attached to the failed switch directly detects the failure, and then floods the network with a message identifying the failed switch. The affected calls (EVCs) are then grouped based on their corresponding (S-D) pairs. The different (S-D) pairs are stored in order of descending bandwidth or according to their Class of Service (CoS) priorities.

The algorithm then tries to restore the EVCs one-by-one starting from the ones that belong to the (S-D) pair with the highest bandwidth/priority using the path selected by the integrated routing algorithm. Depending on available network resources, the algorithm may choose to restore or not to restore best efforts EVCs. Note that the path selection for a given call is based on global optimization of network topology and resources. Note also

that the selected path may use an existing lightpath(s) over layer-2, create a new lightpath (RWA) over layer-1, or use a mixture of existing and a newly created lightpaths (hybrid approach) in a way that optimizes resource usage.

6.10.3 Physical Link Failure (Trunk Cut)

In the case of a trunk cut, the affected calls are grouped and recovered similar to the case of a switch failure described above.

In the signaling scheme used here, the OXC-controller detecting a failure sends a failure indication (or alarm) message to the source OXC of each of the failed connections. Upon receiving the alarm, the source OXC initiates failure recovery by sending a request message towards the destination OXC along the restoration path. Most current restoration signaling proposals can be characterized as “per-connection” in nature since each failed connection is restored using a separate set of signaling messages. As the number of connections affected by a failure increases, the number of restoration signaling messages will increase and hence most likely the queuing delays experienced by these messages at particular nodes along restoration paths (switch configuration waiting times). These queuing delays impact the overall failure recovery times, which become unacceptably high for moderately large number of connections in the network. To cope with this, we will use the concept of alarm and signaling aggregation [118], where connections with the same restoration path (end-to-end) can be restored using a single aggregated alarm, request, and response message.

7. Conclusion

This thesis has examined the technological requirements and assessed the performance analysis and feasibility for implementing a truly native end-to-end Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. Specifically, we have proposed and devised both short and long-terms innovative, graceful networking transition scenarios to evolve Ethernet into a next generation networking technology that can truly support carrier-class Ethernet services.

The short-term solution has addressed the immediate need for legacy systems to support evolving Ethernet services. It utilizes current Layer 2 technology and expands it into the MAN, WAN and long haul networks over DWDM or MPLS Core infrastructures. It has also addressed the current Ethernet shortcomings such as spanning tree protocol and the lack of QoS support required by most applications. Specifically, we have introduced a new admission control mechanism to address key challenges within an end-to-end service architecture along with a novel E-UNI and E-NNI with QoS support. The CoS service performance models proposed by this research is vital for expanding Metro Ethernet into the Core and Long Haul as well as provide the opportunity to standardize Ethernet services (voice, data and video) in the metro, wide, national and global networks.

We have also presented a long-term ambitious vision to implement a truly native end-to-end layer-2 MAC frame-based Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. It has been shown that by combining the simplicity and cost effectiveness of Ethernet technology with the ultimate intelligence of WDM-based optical transport layer, Optical Ethernet (Ethernet-over-WDM) could evolve as a

next generation networking paradigm that provides a seamless global transport infrastructure for end-to-end transmission of native Ethernet frames.

The proposed Optical Ethernet networking architecture is a true two-layer model, realizing the significant goal of Ethernet-over-WDM, where native Ethernet frames are mapped directly over WDM. It offers significant advantages over existing Layer-2 and MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other layer-2 protocols.

The primary rational behind our vision is decreased cost and complexity in the network. This is how Ethernet won the LAN years ago; it was not necessarily the best technology, it was the most cost-effective and easily implemented. Simplifying network design and reducing costs by utilizing Ethernet as an end-to-end LAN/MAN/WAN protocol is the key for Ethernet to win the MAN and WAN.

Appendix A Switch Processing Time Measurements

Latency values used here are based on tests performed in lab testing. The lab testing determined the actual delays on a switch in a given metro network. Tests were performed over GigE links that were adjusted for utilization (10%, 50%, and 100%). Ethernet Frame sizes were varied between 64 bytes, 512 bytes and 1500 bytes. Average latency (µs) was then calculated for all frames over the duration of the test. The results are provided in Table 19.

Table 19: Latency measurement over a GigE with adjusted utilization

Frame Size	*10% Load*	*50% Load*	*100% Load*
64 bytes	0.7 µs	0.70 µs	1.07 µs
512 bytes	0.71 µs	0.71 µs	1.07 µs
1500 bytes	0.71 µs	0.71 µs	1.07 µs

In Table 20, the average latency, in µs, is presented for Agilent tester connected to GigE UNIs on a Sup 2/PFC2 using Sup GE ports.

Table 20: Average latency using Agilent tester and GigE SUP-2 ports

Frame Size	*10% Load*	*50% Load*	*100% Load*
64 bytes	5.69 µs	5.76 µs	6.95 µs
512 bytes	9.82 µs	10.05 µs	11.35 µs
1500 bytes	19.25 µs	19.26 µs	20.55 µs

In Table 21, Average latency, in µs, is presented for Agilent tester connected to GigE UNIs on a Sup 720/PFC3bXL using WS-X6724-GE-SPF ports.

Table 21: Average latency using Agilent tester and GigE-SPF ports

Frame Size	*10% Load*	*50% Load*	*100% Load*
64 bytes	9.70 µs	9.70 µs	10.20 µs
512 bytes	13.96 µs	13.96 µs	14.45 µs
1500 bytes	23.25 µs	23.25 µs	23.85 µs

In summary, it is estimated that in a metro Ethernet network, the Ethernet switch should add no more than 24 µs of delay for a given frame from the incoming port (ingress) to the outgoing port (egress), under normal operating conditions with no congestion.

References

[1] Technical Specification MEF 1.0, OBSOLETE; replaced by MEF 10

[2] Technical Specification MEF 2.0, “Requirements and Framework for Ethernet”, February 2004

[3] Technical Specification MEF 3.0, “Circuit Emulation Service Definitions, Framework and Requirements in Metro Ethernet Networks“, April 2004

[4] Technical Specification MEF 4.0, “Metro Ethernet Network Architecture Framework - Part 1: Generic Framework”, May 2004

[5] Technical Specification MEF 5.0, OBSOLETE; replaced by MEF 10

[6] Technical Specification MEF 6.0, “Ethernet Services Definitions - Phase I”, June 2004

[7] Technical Specifications MEF 7.0, “Element Management System (EMS) – Network Management System (NMS) Information Model”, October 2004

[8] Technical Specification MEF 8.0, “Implementation Agreement for the Emulation of PDH Circuits over Metro Ethernet Networks”, October 2004

[9] Technical Specification MEF 9.0, “Test Procedure - Abstract Test Suite for Ethernet Services at the UNI”

[10] Technical Specification MEF 10, “Ethernet Services Attributes Phase 1”, November 2004

[11] Haidar Chamas, William Bjorkman, Mohammed Ali, “Verizon Experience with Next Generation Ethernet Services: Evolution To A Converged Layer 1, 2 Network”, pp S18-25, IEEE Optical Communications Design, Technologies,

and Applications, August 2005, Vol. 3, No. 3; IEEE Communications Magazine, Vol. 43, No. 8, August 2005.

[12] Andrew Moore, Laura James, Adrian Wonfor, Ian White, and Richard Penty, "Chasing Errors Through The Network Stack: A Testbed For Investigating Errors In Real Traffic On Optical Networks", pp S34-39, IEEE Optical Communications Design, Technologies, and Applications, August 2005, Vol. 3, No. 3; IEEE Communications Magazine, Vol. 43, No. 8, August 2005.

[13] Suresh Bazaj, "The Ethernet Story" presentation, Bazaj Management Consultant, March 13, 2004

[14] Tim Szigeti, Christina Hattingh, 'End-to-End QoS Network Design', Cisco Press, Hardcover, November 2004

[15] Malik Khan, 'Quality of service can also deliver performance monitoring', Server World Magazine, May 200o Issue

[16] Quality of Service Networking, Cisco Systems, Web article, Thu Feb 20 15:38:15 PST 2003

[17] Cisco IOS QoS, http://www.cisco.com/warp/public/732/Tech/quality.shtml

[18] Cisco Systems. Cisco IOS 12.0 Quality of Service. Indianapolis: Cisco Press, 1999.

[19] Ferguson, Paul, and Huston, Geoff, "Quality of Service: Delivering QoS on the Internet and in Corporate Networks", New York: John Wiley & Sons, 1998

[20] Haidar Chamas, William Bjorkman, and Mohamed Ali, 'A Novel Admission Control Scheme for Ethernet Services', IEEE ICC05, Seoul, Korea, May 2005.

[21] Huei-Wen Ferng,"Modeling of Split Traffic Under Probabilistic Routing", IEEE COMMUNICATIONS LETTERS, VOL. 8, NO. 7, JULY 2004

[22] Bob Grow, "The Structure for Congestion Management", IEEE 802.3 Congestion Management Study Group, May 2004

[23] David Martin, "A Survey Of Standards Efforts On Traffic & Congestion Management In Ethernet Networks", IEEE 802.3 Congestion Management Study Group May 24-25, 2004

[24] H. Jonathan Chao and Xiaolei Guo, "Quality of Serivce Control in High-Speed Networks, January 2000, pp13-18

[25] Ralph Santitoro, "Bandwidth Profiles for Ethernet Services", MEF 2004

[26] N. McKeown, et al., "Achieving 100% throughput in an input queued switch", IEEE Transactions on Communications, Aug. 1999.

[27] Stevens, "TCP/IP Illustrated", Volumes 1&2

[28] Peterson and Davie, "Computer Networks"

[29] Walrand and Varaiya, High performance communication networks

[30] Haidar Chamas, William Bjorkman, and Mohamed Ali, "A Novel Admission Control System for Bandwidth on Demand Ethernet Services over Optical Transport Networks", OFC/NFOC March 2005

[31] F. Tobagi, C. Fraleigh, W. Noureddine, "Congestion Control in Local Area Networks", IEEE 802.3 plenary,

[32] Sam Halabi, 'Metro Ethernet', pp23-50, 168-191, Cisco Pres 2003

[33] Mark Dickie, ‘Routing In Today’s Internetworks’, pp2-7, 56-64 Van Nostrand Reinhold 1994

[34] Robert Cole, Ravi Ramaswany, “Wide Area Data Network Performance Engineering”, pp 55-67, 82-89 101-129, Artech House Publishers, 2000

[35] Grenville Armitage, Technology Series Quality of Service In IP Network s”, Macmillan Technical Publishing, 2000

[36] John Spragins with Joseph Hammond and Krzysztof Pawlikpwski, “Telcommunications Protocols and Designs, Addison Wesley, 1991

[37] James Martin with Kathleen Chapman,‘Local Area Networks Architecture and Implementations’, pp 81-91, 163-183, Prentice Hall, 1989

[38] Christian Huitema, ‘Routing in the Internet’, pp135-153, Prentice Hall, 1995

[39] Christian Huitema, ‘Routing in the Internet, 2nd Edition’, pp283-285,333-361, Prentice Hall, 2000

[40] Radia Perlman, Donald Eastlake, “RBridges: Transparent Routing”, IEEE 802.11-05/241r0, March 200

[41] Jose Morales Barroso, “ From ‘Computer Networks’ to the ‘computer on the Net’ The Convergence of Internet, Broadband, and Telephone Neworks in the IEEE 802 standards”, pp2-4, IEEE Global Communications Newsletter, IEEE Communications Magazine, Vol. 43, No.10, October 2005

[42] James Jones, Lyndon Ong, Monica Lazer, “Interoperability Update: Dynamic Ethernet Services Via Intelligent Optical Networks”, ppS4 S10, Standards Report, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[43] Ingvild Sorteberg, ivind Kure, “The Use of Service Level Agreements in Tactical Military Coalition Force Networks”, pp107-114, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[44] Aref Meddeb, “Why Ethernet WAN Transport?” pp136-141, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[45] Mike McFarland, Samer Salam, and Ripin Checker, “Ethernet OAM: Key enabler for carrier class Metro Ethernet Services”, pp152-157, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[46] Ronald van Haalen, Richa Malhotra, and Arjan de Heer, “Optimized Routing for Providing Ethernet LAN Services”, pp158-164, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[47] Anush Elangovan, “Efficient Multicasting and Broadcasting in Layer 2 Provider Backbone Networks”, pp166-170, IEEE Communications Magazine, Vol. 43, No.11, November 2005

[48] Li Xinwan, et al, ‘An Experimentation Study Of An Optical Burst Switching Network Based On Wavelength-Selective Optical Switches’, ppS3-S10, IEEE Optical Communications Design, Technologies, and Applications, Vol. 3, No. 2, May 2005

[49] Mallik Tatipamula,et al, ‘Implementation of IPv6 Services over a GMPLS-Based IP/Optical Network, pp114-122, IEEE Communications, Vol. 43, No. 5, May 2005

[50] James Durkin, ‘Voice-Enabling the Data network: H.323, MGCP, SIP, QoS, SLAs, and Security’, pp83-94, 101-103Cisco Press 2003

[51] Ashwin Gumaste and Tony Antony, ‘DWDM Network Designs and Engineering Solutions’, pp210-215, 260-265, Cisco Press 2003

[52] Dimitri Bertsekas and Robert Gallager, ‘Data Networks’, pp312-323, 423-439, Prentice Hall 1987

[53] Janet Kreiling ‘Metro Ethernet Coming your way’, pp69-71, PACKET Cisco Systems Users Magazine, Vol. 17, No. 1, First Quarter 2005

[54] Chiara Regale ‘Planting New Spanning Trees, Implementing IEEE 802.1w and IEEE 802.1s’, pp23-26, PACKET Cisco Systems Users Magazine, Third Quarter 2003

[55] Janet Kreiling ‘High Availability Networking’, pp53-58, PACKET Cisco Systems Users Magazine, Third Quarter 2003

[56] Alessandro Barbieri ‘10GbE and its X Factors’, pp25-28, PACKET Cisco Systems Users Magazine, Third Quarter 2005

[57] Ferit Yegenoglu and Erick Sherk, ‘Network Characterization Using Constraint-Based Definitions of Capacity, Utilization, an Efficiency’, pp 132-138, IEEE Communications magazine, Vol. 43, No. 9, September 2005

[58] Manoj Wadekar, et al, ‘Proposal for 802.3 Enhancements for Congestion Management’, IEEE Congestion Management Group, May 2004

[59] N. Ghani et al., “on IP-over-WDM Integration,” IEEE Communications magazine, March 2000.

[60] D. Awduche et al, "Multiprotocol Lambda Switching," Internet Draft, work in progress, November 1999.

[61] "IP over Optical Networks: A Framework," draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[62] E. Mannie *et al.*, "Generalized Multi-Protocol Label Switching (GMPLS) architecture," IETF Internet draft, Mar. 2002…

[63] IEEE Std 802.1w-2001, [Amendment to IEEE Std 802.1D, 1998 Edition, (ISO/IEC 15802-3:1998), and IEEE Std 802.1t-2001

[64] IEEE Std 802.1s™-2002, (Amendment to IEEE Std 802.1Q™, 1998 Edition)

[65] IEEE 802.1s "Multiple Spanning Tree Protocol"

[66] IEEE 802.3 Congestion Management Study Group, Portland, Oregon, July 2004

[67] IEEE 802 Standard, "802.3" web link

[68] IEEE Std 802.3-2002

[69] IEEE 802.3ae, "10 Gigabit Ethernet"

[70] IEEE 802.1D-2004 "Spanning Tree Protocol revision of IEEE std 802.1D-1998"

[71] Y. L. Takahashi-Iturriaga, J. Martínez, V. Alarcón, 'A Multiservice Architecture for Dynamic Bandwidth Allocation and Traffic Engineering Applications', Proceedings of the 14th International Conference on Electronics, Communications and Computers (CONIELECOMP 2004)

[72] Francesco De Pellegrini, David Starobinski, Mark G. Karpovsky, and Lev B. Levitin, 'Scalable Cycle-Breaking Algorithms for Gigabit Ethernet Backbones', IEEE INFOCOM 2004

[73] Jeff Hayes, 'Spanning Tree Rapidly branches out', Network World, 02.10.2003

[74] Srikant Sharma, Kartik Gopalan, Susanta Nanda, and Tzi-cker Chiueh,'Viking: A Multi-Spanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks', IEEE INFOCOM 2004

[75] Yankee Group Report - U.S. Ethernet Services Market (May 2004)

[76] "PON & FTTx Update", Light Reading - PON_FTTH.htm, August 2005,

[77] "Ethernet in Access Networks", Light Reading, June 2005

[78] "Next-Gen Sonet", Light Reading, May 2002

[79] "Metro Ethernet", Light Reading, July 2002

[80] "Optical Add/Drop Muxes", July 2002

[81] RFC 2386, "A Framework for QoS-Based Routing in the Internet."

[82] RFC 2474, "Definition of the Differentiated Service Field (DS Field) in the IPv4 and IPv6 Headers",

[83] RFC 2597, "Assured Forwarding PHB Group", June 1999

[84] RFC 2598, "An Expedited Forwarding PHB", June 1999

[85] RFC 2697, "A Single Rate Three Color Marker [srTCM]", September 1999

[86] RFC 2698, "A Two Rate Three Color Marker [trTCM]", September 1999

[87] RFC 2859, "A Time Sliding Window Three Color Marker", June 2000

[88] RFC 2859 , "A Time Sliding Window Three Color Marker, June 2000

[89] RFC 3246, "An Expedited Forwarding PHB",

[90] Tom Sheldon, "Differentiated Services", Linktionary networking defined and hyperlinked, 2005

[91] William Stallings, 'LAN QoS', The Internet Journal, Vol. 4, No. 1, March 2001

[92] Technical Specification MEF 11.0, “User Network Interface (UNI) Requirements and Framework, November 2004

[93] Technical Specification MEF 12.0, “Metro Ethernet Network Architecture Framework; Part 2: Ethernet Services Layer”, April 2005

[94] Peter Tomsu, “High-Speed IP and Optical Networking-Theory and Praxis’, Network World + Interop, May 2004

[95] Chris Olsen, “CCNP Switching Study Guide”, pp.172-185; 226-228, Mc-Graw Hill 2001.

[96] Simon Stanley, “10-Gigabit Ethernet”, Light reading, March 2003ITU-T G.7041, Generic Framing Procedure, October 2001

[97] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects,” IEEE Communications Magazine, pp. 94-102, September 2000

[98] M.A. Ali, A. Shami, C. Assi, Y. Ye, R Kurtz, “Architecture Options for Next-Generation Networking Paradigm: Is Optical Internet the Answer,” Journal of Photonic Network Communications, vol. 3, no. 1/2, Jan-Jun 2001.

[99] ITU-T G.7041, Generic Framing Procedure, October 2001

[100] ITU-T G.707, Network Node Interface for the Synchronous Digital Hierarchy (SDH), October 2000

[101] ITUT G.7042, Link Capacity Adjustment Scheme (LCAS) for Virtual Concatenated Signals

[102] ITU-T G.8080/Y1304, Architecture of the Automatically Switched Optical Network (ASON), 2001

[103] ITU-T G.7713/Y1704, Distributed Connection Management, 2001

[104] ITU-T G.7715 /Y1706, Architecture and Requirements for Routing in the ASON, 2002

[105] OIF2003.248.05, User Network Interface (UNI) 1.0 Signaling Specification, Release 2: Common Part, January 21, 2004

[106] OIF2003.249.03, Extensions for User Network Interface (UNI) 1.0 Signaling, Release 2, July, 26, 2003

[107] OIF2003.293.03, Draft UNI 2.0 Specification, Draft, October 15, 2004

[108] IETF RFC 3471, Generalized MPLS Signaling Functional Description

[109] IETF RFC 3473, Generalized MPLS Signaling: RSVP-TE Extensions

[110] ITU-T, G.7718/Y1709, Framework for ASON Management, draft v0.5, 2004

[111] K. Zhu and B. Mukherjee, “Traffic Grooming in an Optical WDM Mesh Network”, IEEE Journal on Selected Areas in Communications, Special Issue on "WDM-Based Network Architectures", vol. 20, no. 1, pp. 122-133, Jan. 2002

[112] E. Modiano and P. J. Lin, “Traffic Grooming in WDM Networks,” IEEE Communications Magazine, vol. 39, no. 7, pp. 124-129, July 2001.

[113] K. Zhu, H. Zhu, and B. Mukherjee, “Traffic Engineering in Multigranularity Heterogeneous Optical WDM Mesh Networks through Dynamic Traffic Grooming”, IEEE Network Magazine, PP. 8-16, March/April 2003.

[114] M. A. Ali, Keren Bergman, and G. Ellinas, “ Transportation & Switching of native Ethernet frames across MPLS/GMPLS Managed and Controlled Optical

data networks,“ **(INVITED),** Proceedings of the 17th IEEE/LEOS Annual meeting on Optical Networks and Systems, Puerto Rico, Nov 7-11 2004.

[115] Phil Dykstra, “Gigabit Ethernet Jumbo Frames,” http://sd.wareonearth.com/ phil/ jumbo.html.

[116] Shara Evans, “Jumbo Frames,” http://telsyte.webboy.net/standardswatch /jumbo_a.htm.

[117] ITU-T Question 5 Study group 13: Draft Recommendation Y.17ethoam – OAM Functions and Mechanisms for Ethernet based networks, December 2004.

[118] “IEEE 802.1ag/D3.0”, Draft Standard for Local and Metropolitan Area Networks, April 2004.

[119] Jose Morales, “From Computer Networks to the computer on Net”, IEEE Communications Magazine/Global Communication Newsletter, October 2005