Traffic Engineering for Inter-domain Optical Networks

By

Mohamad Alassaf

A Thesis submitted to the Faculty of Graduate Studies and Postdoctoral Studies in partial fulfillment of requirement for the degree of

Master of Applied Science
in
Computer Science

Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada

Université d'Ottawa
University of Ottawa

Abstract

Because of increasing demands for network bandwidth and the advances in WDM technologies, it is expected that IP (Internet Protocol) over Wave Division Multiplexing (WDM) will be the infrastructure for the next generation Internet by carrying IP packets over WDM. As this infrastructure matures, a number of important control issues are surfacing. One of them is traffic engineering, which concerns network performance optimization.

This thesis presents a new approach to the congestion avoidance problem for wavelength-routed inter-domain networking under dynamic traffic demands where IP traffic is directly routed over WDM. The main idea is to adapt the underlying optical connectivity by measuring the actual traffic load over the existing lightpaths continuously and reacting to the load changes by adding or tearing down lightpaths. We avoid load instability directly by setting up a lightpath when congestion occurs or by tearing down a lightpath when the congestion is relieved. For the decision of setting up or tearing down lightpath, we propose a traffic engineering algorithm. The algorithm is characterized by two system parameters, high threshold (H) and low threshold (L).

A simulation is conducted with different network topologies. The simulation results show the improvement we achieved in term of packets loss, resource utilization, blocking probability and delay.

Acknowledgment

I am deeply grateful to my supervisor Prof. G. V. Bochmann for his knowledgeable academic guidance, his valuable advices and his helpful suggestions. I also appreciate his financial support during my study. As a member of his research group, professor Bochmann never hesitated in bringing our group together as a family working together and helping one another.

I want to give my great thanks to my fried Abdelilah Maach for his help and knowledge, to the university administration staff who never hesitated in giving me excellent services and advices, and to all friends that I met at the university with whom I had a wonderful friendship.

Finally, I want to contribute all my accomplishment to my parents who brightened my future in this life through their love and hope that they gave me. Thanks to all my brothers who always pushed me to work hard for my education.

Table of Contents

Abstract **1**
Acknowledgment **3**
Table of Contents **4**
List of Figures **6**

Chapter 1. Introduction **8**
1.1 Context of the Thesis 10
1.2 Objectives of the thesis 11
1.3 Thesis Organization 12

Chapter 2. IP over Optical Networks **13**
2.1 Optical Transport Networks Layered Module 14
2.2 Logical Control Structure 15
2.3 IP and Optical Control Planes 16
2.4 Optical Internetworking Models 17
2.4.1 Overlay Model 18
2.4.2 Peer Model 20
2.5 Packet Switching in the Optical Domain 20
2.5.1 Optical Packet Switching 21
2.5.2 Photonic Slot Routing 21
2.5.3 Optical Burst Switching 22

Chapter 3. Quality of Service in IP Networks **23**
3.1 Quality of Service Attributes 23
3.1.1 Throughput 24
3.1.2 Delay 24
3.1.3 Loss 25
3.1.4 Jitter 25
3.2 Integrated Service Module 26
3.3 Differentiated Services 27
3.3.1 Service Level Agreement and Traffic Conditioning Agreement 27
3.3.2 Basic Architecture for Differentiated Services 29
3.3.3 Packet Classifier and Traffic Conditioner 30
3.3.4 Per-hop Behaviors 31
3.3.5 Discussion 32

Chapter 4. IP and Wavelength Routing Networks **33**
4.1 Routing in Datagram Networks. 33
4.1.1 Inter-domain Routing 34
4.1.2 Border Gateway Protocol 35
4.2 Wavelength Routing Networks 37
4.2.1 Optical Border Gateway Protocol 37
4.2.2 Routing and Wavelength Assignment. 38

Chapter 5. Traffic Engineering 40
5.1 Definition of TE 40
5.2 Activities Required for TE 41
5.3 Criteria for Selecting the Best Traffic Route 42
5.4 Congestion Control 42
5.4.1 Categories of Congestion Control 42
5.4.2 Control Policies 44
5.5 MPLS, and GMPLS Traffic Engineering 45
5.5.1 MPLS Traffic Engineering Architecture 46
5.5.2 MPλS/GMPLS Control Plane 48
5.6 Related Works 49
5.6.1 BGP and MPLS, GMPLS Technology 50
5.6.2 Third Party Agent 50

Chapter 6. Traffic Engineering Algorithm 53
6.1 Network Architecture 55
6.2 TE-DPM, Router and OXC Interaction 56
6.3 TE Decision Process Module 58
6.3.1 System Parameters 58
6.3.2 Oscillation Interval 59
6.3.3 Decisions Taken by the DPM 59
6.4 An Example Network 63
6.4.1 First scenario (traffic load larger than H) 64
6.4.2 A second scenario (traffic to a new destination) 66

Chapter 7. Simulations Based performance 69
7.1 Performance Metrics 69
7.2 Simulation Results 70
7.2.1 NSFNET Network 70
7.2.2 Simple Network 77
7.2.3 Simulations to study the effect of oscillations 83

Chapter 8. Conclusions of the Thesis 85
8.1 Contributions of the Thesis 86
8.2 Topics for Future Study 86

References 88
Appendices 95
A-List of Acronyms and Abbreviations **95**
B-Terminology and Concepts **97**

List of Figures

Figure 1: Optical network layered architecture. 14
Figure 2: Optical network logical control structure. 15
Figure 3: IP and optical control and data plane. 17
Figure 4: Overlay model. 19
Figure 5: Peer model. 19
Figure 6: Smart Router Simple Optics architecture. 20
Figure 7: Differentiated service domain. 30
Figure 8: DiffServ field in the 8-bit ToS. 30
Figure 9: Packet classifier and traffic conditioner. 31
Figure 10: Current internet architecture. 35
Figure 11: MPLS-TE functional components. 47
Figure 12: GMPLS label-stacking hierarchy. 49
Figure 13: Decision process model. 54
Figure 14: Network architecture. 55
Figure 15: Network node architecture. 56
Figure 16: TE-DPM, IP Router and OXC interaction. 57
Figure 17: Oscillation interval. 59
Figure 18: DPM algorithm. 62
Figure 19: Network example. 64
Figure 20: Buffers for AS100. 67
Figure 21: NSFNET network 70
Figure 22: Fraction of packets lost with different TE techniques and no traffic forwarding (NSFNET network). 72
Figure 23: Average lightpaths established with different TE techniques and no traffic forwarding (NSFNET network). 72
Figure 24: Blocking rate with different TE techniques and not traffic forwarding (NSFNET network). 73
Figure 25: Average delay with different TE techniques and no traffic forwarding (NSFNET network). 73
Figure 26: Packet loss rate (NSFNET network). 75
Figure 27: Average number of lightpaths established (NSFNET network). 75
Figure 28: Blocking rate (NSFNET network). 76
Figure 29: Average delay (NSFNET network). 77
Figure 30: Simple Network. 78
Figure 31: Packet loss rate (simple network one). 79
Figure 32: Packet loss rate (simple network two). 79
Figure 33: Packet loss rate (s 79
Figure 34: Average number of lightpaths established (simple network one). 80
Figure 35: Average number of lightpaths established (simple network two). 80
Figure 36: Average number of lightpaths established (star network). 80
Figure 37: Blocking rate (simple network one). 81
Figure 38: Blocking rate (simple network two). 81
Figure 39: Blocking rate (star network). 82

Figure 40: Average delay (simple network one). 82
Figure 41: Average delay (Simple network two). 83
Figure 42: Average delay (Star network). 83
Figure 43: Oscillation interval example. 84
Figure 44: Oscillation factor. 84

Chapter 1

Introduction

One of the critical issues related to the automatic control and provisioning in optical networks is how to provide effective lightpath provisioning to improve network performance (e.g. blocking probability, delay). Transmission of client traffic (e.g. Internet protocol traffic) on the optical network is achieved through a logical topology in the form of switched lightpaths. Each lightpath is used to provide the link connectivity for a pair of client nodes. Therefore, in WDM optical network, a control mechanism is needed to establish and tear down a lightpath based on client traffic characteristics.

Simplifying the network architecture by reducing the number of layers is needed to reduce both the deployment costs and the operating expenses. There are three possible approaches for IP over WDM. The first approach is to transport IP over ATM, over SONET/SDH and WDM. The advantages of this approach is the ability to carry different type of traffic onto the same fiber with different QoS requirement. Another advantage is the ATM traffic engineering (TE) capability and its flexibility in network provisioning. However, the drawback of this approach is its management complexity. The second approach is IP/MPLS over SONET/SDH over WDM. Packets over SONET/SDH technology is used for transporting IP data directly over SONET/SDH infrastructure to eliminate the ATM transmission layer. The advantages of this architecture are the SONET optical signal multiplexing hierarchy, where low–speed signals can be multiplexed into high-speed signals, and the SONET network protection/restoration capability. The third approach is IP/MPLS over WDM, where SONET/SDH and ATM layers are eliminated, thus building only a two-layer network using IP and WDM layers. This approach requires that the IP layer looks after path protection and restoration unless it is provided by the optical layer. This approach is considered to be the most efficient solution among these approaches.

One of most potent technical advancement in the telecommunication technology has been the development of Wave Division Multiplexing (WDM). WDM has increased the

bandwidth of the optical fiber; it is based on a concept called frequency division multiplexing (FDM). WDM technology was born from the idea to simultaneously inject into the same optical fiber several digital signals, each one with at distinct wavelength. International standard ITU-T G 692 (optical Interfaces for systems with optical amplifiers) defines a series of recommended wavelengths in the transmission window 1530 to1565 nm. It standardizes spacing in nanometers (nm) or Gig hertz (GHz) between two recommended wavelengths in the transmission window: 200GHz or 1,6 nm and 100 GHz or 0,8 nm.

WDM technology is said to be Dense (DWDM) when spacing used is equal or less than 100 GHz. Systems at 50 GHz (0,4 nm) and at 25 GHz (0,2 nm) spacing have already been tested. WDM / DWDM on the market today have 4, 8, 16, 32 or even 80 optical channels, which allows transmission rates of 10, 20, 40, 80 even 200 Gb/s with 2,5 Gb/s per channel, and 4 times more at 10 Gb/s per channel. To be more concrete, a system with 16 x 2,5 Gb/s channels makes it possible to transmit 500 000 simultaneous phone calls on one fiber.

As the carriers build global and national networks, the network architecture became more complex and difficult to manage which led to the need to partition the network into sub-networks (domains). A domain is defined as a set of routers or networks that are technically administrated by a single organization such as corporate network, a campus network, or an Internet Service provider (ISP). Routing protocols running inside these domains are called Interior Gateway protocols (IGP) (e.g. OSPF, IS-IS). While routing protocols running between these domains are called Exterior Gateway Protocol (EGP). The Border Gateway Protocol (BGP) is the current inter-domain routing protocol running over the Internet. The control boundary between these domains and the extent of information shared across them became an important issue, since they influence path selection that is very important for traffic management. This kind of interaction between domains in optical networks led the researchers to work on what is called Inter-domain optical Routing.

Finally, this thesis focuses on TE for inter-domain networks, where IP packets are directly routed over lightpaths. We introduce an algorithm that controls the traffic load at an edge node of a given domain by measuring the actual traffic over each established lightpath.

Based on this measurement, a decision of establishing or tearing down lightpaths between this domain and other domains will be taken. The next two sections will describe in more details the focus and the objective of this thesis.

1.1 Context of the Thesis

Most recent research in the field of TE focused on intra–domain TE issues using Intra-domain technologies such as MPLS [AMAO99] and GMPLS [BANE01][RVCA01] where the network is managed by a single administration. In this architecture, the network topology and its resources are controlled by a single network administration. Fewer research focused on TE for Inter-domain routing where the network resources are shared between different Autonomous Systems (ASs). In this study, there is no intra-domain TE consideration, we focus on inter-domain TE where each AS (e.g. enterprise network and Internet Service Provider), uses BGP [RELE95] as its routing protocol for the IP layer.

Measurements are an important component for TE and it is important for the optimization function of the network [ACEW02]. IETF [BMRU99] provided a framework for describing network traffic flow and architecture for traffic flow measurement. The work of this thesis is based on applying dynamic measurement of the actual traffic load over the lightpath, where the measurement system monitors the traffic load continuously based on the measurement period. A mechanism is proposed to follow the changes in traffic load, where a decision of establishing and tearing down lightpath is taken based on these traffic loads.

The current Internet architecture is partitioned in different domains, also called an ASs. An AS is defined as a set of network nodes that are technically managed by a single networks administration. Inter-domain TE is more difficult to deal with than intra-domain TE for some reasons: First, in intra-domain architecture, the routing protocol is aware of the network topology (e.g. in MPLS, a link-state database provides a topological view of the whole network), this awareness facilitates the network management. In the inter-domain architecture, there is no clear view of the network architecture and connectivity (e.g. BGP is the current inter-domain routing protocol). BGP is called a reachability protocol which means that BGP knows how to reach a destination but does not know about the network topology.

This makes network management more difficult. Second, from the resources management point of view, the network administrator of the domain has control over the network resources internally but not externally (between domains), where the network resources are shared between different domains. This leads to a competition between domains over the network resources which makes resource management more complex.

1.2 Objectives of the thesis

Our objective is to define a TE mechanism that improves the inter-domain network performance. This can be achieved by reducing resources utilization (reduce the competition over network resources, reduce the blocking probability), by reducing the network delay, and by reducing the packets losses.

One of the key performance metrics in wavelength routed optical networks is the blocking probability (e.g. the probability that a connection request for a lightpath will be denied because of unavailable resources). A connection request is blocked when a route can not be found from the source to destination. Furthermore, without wavelength converters a lightpath has to use the same wavelength on each hop along its path. If such a wavelength can not be found, the connection request will be blocked even if there is free capacity on each hop of the path. To reduce the blocking probability and to make the network workable and cost-effective, network resources (wavelengths) must be utilized wisely. We achieve this objective by tearing down under-utilized wavelengths. If the traffic load (between source and destination) over the established lightpath reaches the assigned low threshold (L), the lightpath will be torn down and the residual traffic will be forwarded to an intermediate network node. Then the intermediate network node forwards the residual traffic to its final destination (Chapter 6 describe this process in ore details).

In the IP network, delay and packet loss are important metrics for the network performance. Packets can be lost because of network outages, networks re-routing or congested routers. The approach to minimize delays and losses is to engineer the network to minimize the situations that may cause delay. To achieve this objective, we define a high threshold to detect congestion over the established lightpaths. As the traffic load reaches the

defined threshold, a new lightpath will be established. As a result, packet loss and delay will be minimized.

1.3 Thesis Organization

The thesis is organized as the follows. *Chapter 2* gives a brief description of optical transport network layer architecture, and then we introduce the IP over WDM architecture in term of logical control relation, control plane and their management interface models. *Chapter 3* intends to give an introduction about quality of service in the internet and its attributes, including the integrated service module. *Chapter 4* describes an IP routing, inter-domain routing and BGP routing protocol. Then we give a brief description of BGP. Finally we give an overview of wavelength routing in optical networks. *Chapter 5* focuses on TE related issues. We start by giving a definition of TE and its requirements. Then we introduce the parameters that can be considered when selecting the best traffic route. As congestion is considered to be an important issue in TE, we explain different mechanisms for congestion avoidance. In the last section, we give general description of MPLS-TE and GMPLE -TE.
Chapter 6: We describe our TE algorithm and its working mechanism. Then a network example is presented to explain the proposed algorithm. *Chapter 7* presents the simulation results of the proposed TE algorithm are shown and explained. *Chapter 8* explains the main contributions of the thesis and point out directions for future work.

Chapter 2

IP over Optical Networks

The widespread of the Internet and its related applications, based upon Internet Protocol (IP), made IP the dominant protocol in telecommunications. It is expected that the common traffic convergence layer in communication networks will be IP. It is well known that voice traffic, data traffic and IP-based applications are causing an increasing demand for bandwidth. In parallel with these demands, developments in wave-division multiplexing (WDM) technology have brought a very high capacity on fiber optic links. As expected, the rapid growth of IP technology and the large bandwidth increases provided by WDM technology have brought the researchers to work on bringing these two technologies closer together.

Two technologies are used in switching networks: circuit switching and packet switching. In circuit-switched networks, network resources are reserved all the way from source to destination before the beginning of the data transfer. The resources are dedicated to the circuit during the transfer time. Control signaling and payload data transfer are separated in circuit-switched networks. Processing of control information and control signaling such as routing is performed mainly at circuit setup and termination. Consequently, the transfer of payload data within the circuit does not contain any overhead in the form of headers or the like. An example of circuit switching is the traditional voice telephone service. An advantage of circuit-switched networks is that they allow for large amounts of data to be transferred with guaranteed transmission capacity, thus providing support for real-time traffic.

Packet switching was developed to cope more effectively with the data-transmission limitations of the circuit-switched networks during bursts of random traffic. In packet switching, a data stream is divided into standardized packets. Each contains address, size, sequence, and error-checking information, in addition to the payload data. The packets are then sent through the network, where specific packet switches or routers sort and direct each single packet.

Packet-switched networks are based either on connectionless or connection-oriented technology. In connectionless technology, such as the Internet, packets are treated independently of each other inside the network, because complete information concerning the packet destination is contained in each packet. This means that packet order is not always preserved, because packets destined for the same receiver may take different paths through the network. In connection-oriented technology such as asynchronous transfer mode (ATM), a path through the network often referred to as a logical channel or virtual circuit is established when data transfer begins. Each packet header then contains a channel identifier that is used at the nodes to guide each packet to the correct destination.

2.1 Optical Transport Networks Layered Module

Like most communication systems in place today, optical networks are described by a layered model. The ITU-T [G.872] has published a layered model for next generation optical networks, as shown in Figure 1. The generic functional capabilities of the optical transport layer can be decomposed into three independent logical transport sub-layers:

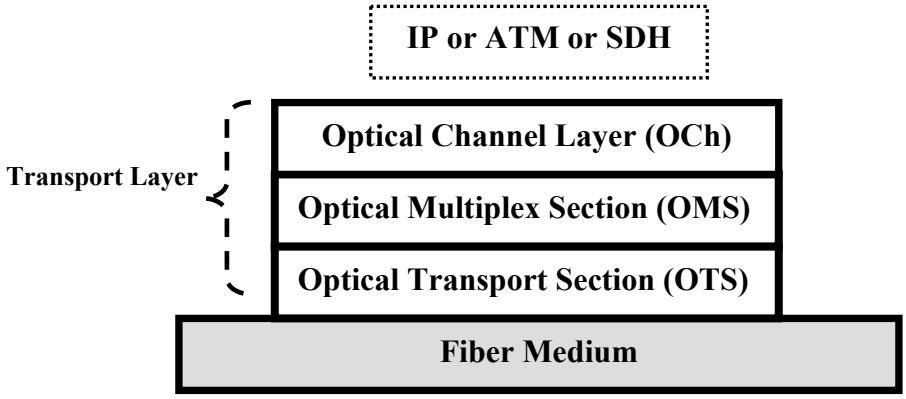


Figure 1: Optical network layered architecture.

- *Optical Transport Section (OTS)* – Provides the functionality for transmission of optical signals on various types of optical media.
- *Optical Multiplex Section (OMS)* - provides the transport of multi-wavelength optical signals, including the insertion of the multiplex section overhead to ensure the integrity of the signal. It provides multiplex section survivability.

- *Optical Channel Layer (OCh)* - provides end-to-end networking of optical lightpaths for transparently carrying various client signals (e.g. IP, ATM. SDH). OCh also prepares and inserts an overhead for channel configuration information such as wavelength tag, port connectivity, payload label and other.

2.2 Logical Control Structure

In optical networks, we can define three logical interfaces that present the relation between nodes, as shown in Figure 2 [RLAW03]. These interfaces are: User–Network Interface (UNI), Internal- Node-to-Node-Interface (INNI), and External-Node-to-Node-Interface (ENNI), respectively. The difference between these interfaces is based on the type and amount of control information that flows across them.

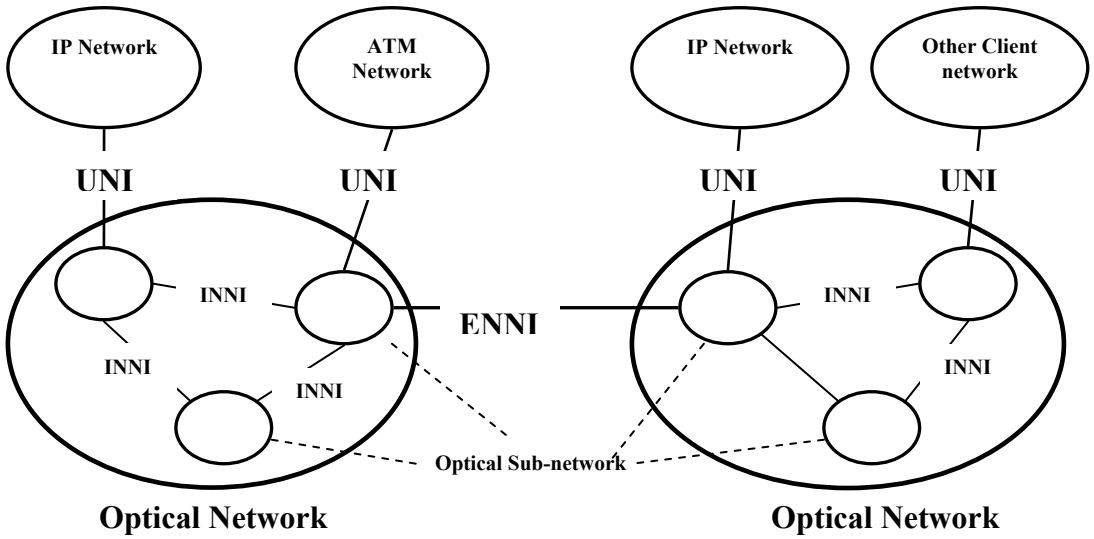


Figure 2: Optical network logical control structure.

- UNI represents a service boundary between the client (e.g. IP router) and the optical network, where the client requests a service connection from the optical network (Server) and the server establishes the connection to fulfill the client request.
- INNI represents the interface within a given network under a single technical administration.
- ENNI indicates an interface at the administration boundary between networks.

We can notice that INNI may differ from ENNI in policies that restrict information flow between nodes.

2.3 IP and Optical Control Planes

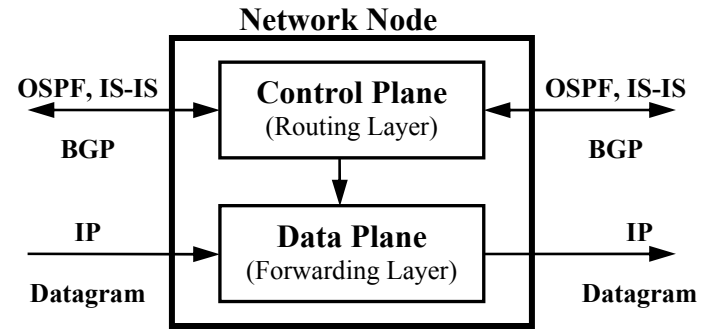
A control plane is a set of software and/or hardware modules in a network node that is used to control several vital operations of the network (e.g. bandwidth allocation, route discovery) [BLAC02].

The Internet control and data planes have usually been named the routing layer (for the control plane) and the forwarding layer (for the data plane), as shown in Figure 3-a. The control plane is comprised of the routing protocols (e.g. OSBF, IS-IS, and BGP). The data plane is comprised of the forwarding protocol (IP).

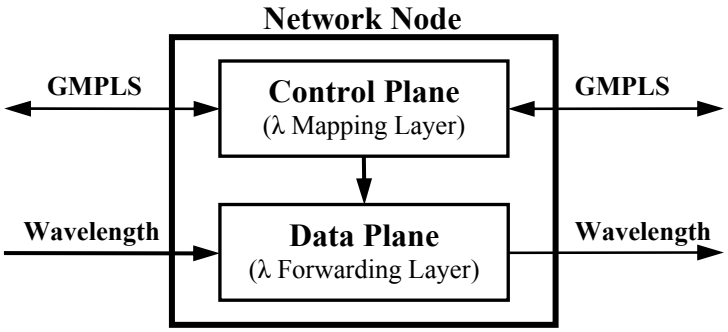
The optical control and data plane could be named the λ mapping layer (for the control plane) and the λ forwarding layer (for the data plane) [BLAC02], as shown in Figure 3-b. The control plane can be executed with GMPLS.
For optical networks, some of the tasks that the control plane performs are:

- Setting up and tearing down optical lightpaths.
- Providing timing messages to keep nodes' clock in sync with each other.
- Building a forwarding (cross-connect) table to allow the data plane to relay traffic from its input link (input port) to its output link (output port).

Control plane messages can be exchanged on separate physical fibers, on the same fiber link as the link used for data traffic, or on a separate wavelength on the same fiber that is transporting user traffic on the other wavelengths of the fiber.



a) The Internet control and data planes



b) Optical control and data planes

Figure 3: IP and optical control and data plane.

2.4 Optical Internetworking Models

Traditionally, the optical layer of the network has been viewed as a simple transport medium. Its primary function has been to provide a connection between electrical layer devices (e.g. IP routers) or asynchronous transfer mode (ATM) switches, in the form of static optical links. In this traditional model, internetworking between the electrical layer and optical layer has proven to be slow, complex, and error prone.

A new boundary between the electrical layer service networks and the high-speed optical switched core is needed for faster provisioning bandwidth. Two models have been proposed for internetworking between the electrical layer and the optical layer: the overlay model and the peer model [TOSC02].

2.4.1 Overlay Model

In the overlay model, electrical devices (e.g. IP routers) operate more or less independently of the optical devices (e.g. OXC) and have no visibility of the optical network. This model is commonly referred as user-network interface (UNI), as shown in Figure 4. General characteristics of the overlay model include:

- User-client devices (e. g. IP routers) signal the optical network through the UNI to request an optical circuit.
- Routing protocol, topology distribution, signalling protocols, and control plane are independent for the IP and optical layers.
- UNI is used only at the edge of the optical network; optical layer devices use an internal signalling protocol between optical network switches.
- The model allows multiple services to be provided over the optical networks (e.g. IP, ATM, and SONET/SDH).
- The network hides optical layer topology and management from the client devices.

There are separate and different control planes in the IP layer and the optical layer. In other words, the IP layer address scheme, routing protocol, and signalling scheme run independently from the addressing scheme, routing protocol, and signalling protocol deployed in the optical layer. In this model, the optical layer provides point-to-point connections to the client devices (e. g. IP routers).

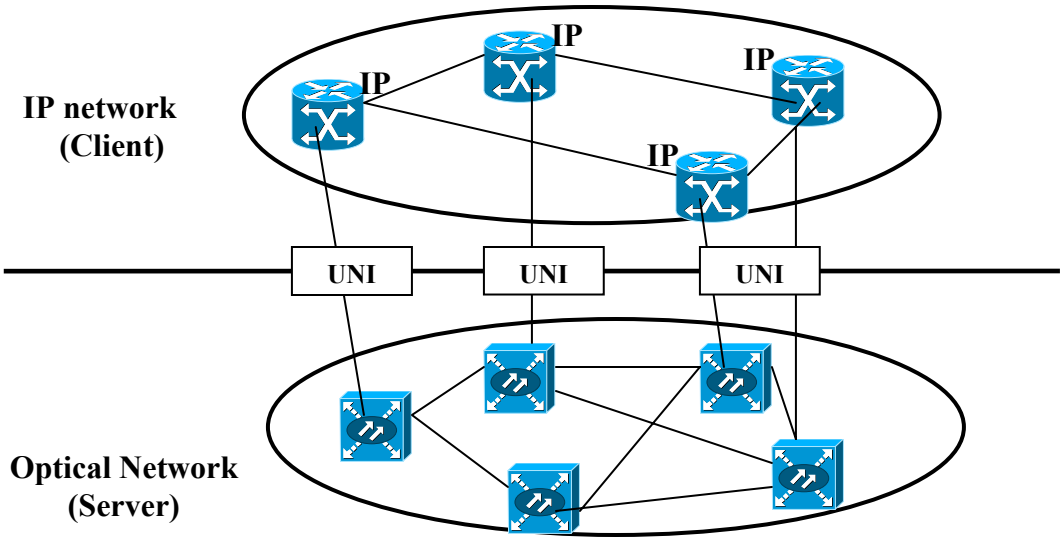


Figure 4: Overlay model.

The IP over WDM network architecture has been studied widely [GIDI00]. An architecture called "Smart Router Simple Optics "(SRSO) [GRHJ00], as shown in Figure 6, represents an architecture for IP over WDM. In this architecture, each node consists of an IP router and an optical cross connects. The IP layer operates directly over the optical resources where the IP router owns the optical resources and controls their use. The importance of SRSO is its ability to reconfigure and manage optical resources to improve the network performance.

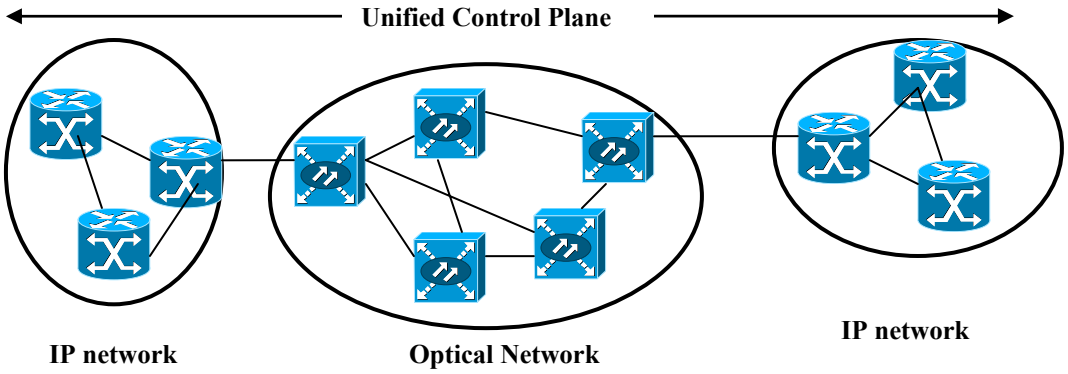


Figure 5: Peer model.

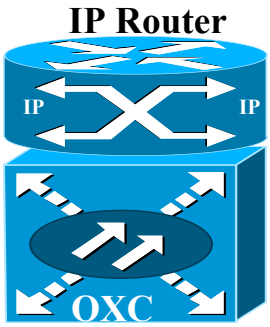


Figure 6: Smart Router Simple Optics architecture.

2.4.2 Peer Model

In the peer model, the electrical layer systems (e.g. IP routers) and optical layer systems act as peers, as shown in Figure 5. In this model, there is a single control plane that spans the client layer and the optical layer. In other word, there is a uniform control plane for both the optical and electrical layers [SWAL99]. The peer model has the following characteristics [TOSC02]:

- Single and common signaling and routing protocol scheme.
- Common addressing scheme used for optical and IP networks.
- Routers and optical switches act as peers and share a common view of the entire network topology.
- Single administration domain.

2.5 Packet Switching in the Optical Domain

Switching is used to achieve connectivity among users while sharing network links. In switched networks, we can establish higher capacity connection paths among users with fewer links and lower cost. In optical switching, a number of research group have proposed various optical switching techniques, such as optical packet switching, photonic slot routing and optical burst switching. In this section we give an overview of these switching techniques. We need to mention that none of these techniques were used in our research work.

2.5.1 Optical Packet Switching

In optical packet switching [YXMB02], incoming IP packets are switched all-optically without being converted to electrical signals. The biggest challenge that packets face in an optical switch is the lack of large buffers for time of contention. When packets are being switched, contention occurs whenever two or more packets are trying to leave the same switch through the same output port, on the same wavelength, at the same time. In electrical packet-switched networks, contention is resolved using the store-and-forward technique, which requires the packets in contention to be stored in memory and be sent out at a later time when the desired output port becomes available.

To overcome the lack of buffers in optical packet switching, WDM networks provide three mechanisms for contention-resolution, wavelength conversion, optical delay lines and the space deflection approach.

- *Wavelength conversion* offers effective contention resolution without relying on buffer memory [ERLI00][HNCA99]. Wavelength converters can convert wavelengths of packets that are contending for the same wavelength of the same output port.
- *Optical delay lines* [YMHA99] are a close imitation of the random-access memory (RAM) in electrical routers, although these offer a fixed and finite amount of delay. However, since optical delay lines depend on the propagation delay of the optical signal to buffer the packet in time, they have more limitation than electrical RAM. Moreover, buffering techniques based on fiber delay lines can accommodate at most a few tens of packets. With such small buffers, the packet drop rate of an optical switch is quite high even for moderate loads.
- *The space deflection approach* [HLHU02] is a multipath routing technique. Packets that have contention are routed to nodes other than their preferred next-hop nodes, with the expectation that they will be routed to their destination.

2.5.2 Photonic Slot Routing

In photonic slot routing [WEDZ02], time is divided into time units, called slots. All the wavelengths are slotted synchronously. Each photonic slot, across all wavelengths, has a pre-

assigned destination, therefore, all the packets transmitted in a photonic slot on different wavelengths are routed together the same path. By requiring the packets to have the same destination, the photonic slot is routed as a single entity, without the need for demultiplexing individual wavelengths at the intermediate switch, resulting in less complexity and faster routing.

2.5.3 Optical Burst Switching

Optical burst switching (OBS) [QIYO99] [TURN99] is an approach that attempts to shift the computation and control complexity from the optical domain to the electrical domain, from the core to the edge of the optical network. In an IP over WDM content, a burst formatted at the network edge contains a number of IP packets in OBS, a control packet, separate from the data burst that carries the payload, is sent first over a separate control wavelength. The switch along the path will be configured according to the information carried in the control packet as it propagates along. Shortly after the control packet is sent, the optical burst leaves the source node and will go through the configured nodes.

OBS uses electrical buffers at the ingress to aggregate regular IP packets destined to the same egress node into burst. The aggregation reduces the number of forwarding decisions that have to be done by the OBS so that they can be done electronically. The trade-off for this is that OBS requires more buffering at the ingress of the optical backbone than the optical circuit switching solutions because IP packets in OBS have to wait until the next burst departs.

Chapter 3

Quality of Service in IP Networks

IP networks have traditionally supported only public Internet service. At the beginning, Internet applications (e.g. web access, e-mail) were not considered mission-critical and did not have specific performance requirements for throughput, delay, jitter, and packet loss. As a result, best-effort class of service was adequate to support the Internet applications. Best-effort means that IP makes a reasonable effort to deliver each datagram to its required destination, but there is no guarantee that a packet will be duplicated, corrupted, lost, or reordered. There is no guarantee that a traffic stream will not experience low throughput, delay, jitter, or loss.

As the Internet grows, and as a various applications increase, there is an immediate need to provide different service classes to different traffic flows. Supporting different service classes for specific applications or customers is concerned with treating packets that belonging to certain data streams differently from packets that are belonging to other data streams. Differentiated service levels are supported by manipulating the attributes (throughput, delay, jitter, loss) of certain streams to change the customer's perception of the quality of service that the network is delivering. These attributes are described in the next section.

3.1 Quality of Service Attributes

Quality of service (QoS) is a measure of how well the network behaves and attempts to define the characteristics and properties of specific services [FEHU98]. Packet transmission services have the following attributes [SEST01].

3.1.1 Throughput

Throughput is a generic term used to describe the capacity of a system to transfer data. It is easy to measure the throughput for a time division multiplexing (TDM) service, because the throughput is the bandwidth of the transmission channels. However, for TCP/IP, throughput is measured in different ways, including:

- The packet or byte rate of host-to-host aggregated flows,
- The packet or byte rate across the connection,
- The packet or byte rate of a specific application flow, or
- The packet or byte rate of a network-to-network aggregated flow.

In best-effort service, the router does not control the amount of bandwidth assigned to different classes. Instead, during a congestion period, all packets are placed into a single first-in, first-out (FIFO) queue. The FIFO is a queuing discipline in which entities in the queue leave the queue in the same order in which they arrive. When congestion occurs, User Datagram Protocol (UDP) flows continue to transmit at the same rate, but Transmission Control Protocol (TCP) flows detect and then react to packet loss by reducing their transmission rate. As a result, UDP flows end up consuming relatively more bandwidth in the congested node.

3.1.2 Delay

Delay is defined as the amount of time that it takes for a packet to be transmitted from the ingress node in the network to the egress node. There are many factors that contribute to the amount of delay experienced by the packet as it travels to its destination. These components are: forwarding delay, queuing delay, propagation delay and serialization delay. End-to-end delay can be calculated as the sum of the forwarding, queuing, serialization, and propagation delay.

- **Forwarding delay.** Forwarding delay is the amount of time that it takes a router to receive a packet, make a forwarding decision, and then transmit the packets through an uncongested output port. The forwarding delay is typically of the order of tens or hundreds of microseconds in high-speed routers.

- **Queuing delay.** Queuing delay is the amount of time that a packet has to wait in a queue while other packets are serviced before it can be transmitted over the output port.
- **Propagation delay**. Propagation delay is the amount of time that it takes data to traverse a physical link. The propagation delay is based on the speed of light and measured in milliseconds.

In addition to the previously mentioned factors, there are many other components that must be considered for end-to-end delay such as operating system scheduling delays, physical layer framing delay and the stability of routing in the network.

3.1.3 Loss

There are three reasons for packet loss in IP networks: line failures, data corruption, and buffer overflow. Breaks in physical links could occur, that prevents the transmission of a packet. In the modern physical layer technology, the chance of data corruption is statically insignificant, so it can be ignored. The primary reason for packet loss in an IP network is buffer overflow. Buffer overflow occurs when congestion occurs. The amount of packet loss is typically expressed in terms of the probability that a given packet will be discarded by the network.

The amount of packet loss, in any network, should be limited, significant packet loss affects the operation of the network. High packet loss can create many problems for the network (e.g. routing instability as a result of disruption in the exchange of routing information, new TCP sessions cannot be established, or the network is no longer able to absorb busts of traffic).

3.1.4 Jitter

Jitter is the variation in delay over time experienced by consecutive packets that are part of the same flow. Jitter is measured by using a number of different techniques, including the mean, maximum, or minimum of interpacket arrival times for consecutive packets in a given flow.

TDM systems can cause jitter, but the variation in delay is so small that it can be ignored. With voice or video applications, jitter can result in jerky or uneven quality to the sound or image. However, for other applications (e.g. applications that run TCP/IP), jitter is not a problem.

3.2 Integrated Service Module

In 1981[DARBA81], Internet Protocol was standardized and the second byte of the IP header was reserved as the type of service (ToS) field. The bits of the ToS byte were defined. The first three bits (called precedence bits) could be set by a node to select the relative priority or precedence of the packets. The next three bits could be set to specify normal or low delay, normal or high throughput, and normal or high reliability. The final two bits of the ToS were reserved for the future use. However, very little architecture was provided to support the delivery of differentiated service classes in IP networks using this capability. Then, IETF developed a mechanism that would allow IP to support more than a single best-effort class of service. This mechanism is called Integrated Services (IntServ). The IntServ architecture is based on per-flow resource reservation and defines a reference module that specifies a number of different components. These components are [GRLR00]:

- The resource reservation setup protocol (RSVP) [EBBB97]. This protocol allows individual applications to request resources from routers and then install per-flow state along the path of the packet flow.
- *Two service modules* guaranteed service and controlled load service.
    - Guaranteed service provides firm assurances (through strict admission control, bandwidth allocation, and fair queuing) for applications that require guaranteed bandwidth and delay.
    - The controlled load service does not provide guaranteed bounds on bandwidth or delay, and emulates a lightly loaded, best-effort network.
- *A flow specification* provides a syntax that allows applications to specify their specific resource requirements.
- *A packet classification process* examines incoming packets and decides which of the various classes of service should be applied to each packet.

- *An admission control process* determines whether a requested reservation can be supported, based on the availability of both local and network resources.
- *A policing and shaping process* monitors each flow to ensure that it confirms to its traffic profile. Traffic profile specifies the temporal properties of a traffic stream
- *A packet scheduling process* distributes network resources (buffer and bandwidth) among the different flows.

The IntServ requires that source and destination nodes exchange RSVP signaling message to establish packet classification and forwarding state at each node along the path between them.

The IntServ was not a suitable mechanism to support the delivery of differentiated service classes in large IP networks for many reasons. First, IntServ is not scalable, because of the significant amounts of per-flow state that needs to be monitored at each node in the packet forwarding path. Second, IntServ requires that applications running on end systems support the RSVP signaling protocol. In the past, there were very few operating systems that supported an RSVP API that application developers could access. Third, IntServ requires that all nodes in the network path support the IntServ model. This includes the ability to map IntServ service classes to link-layer technologies.

3.3 Differentiated Services

As discussed in the previous section, the failure of the IntServ model was due to signaling explosion and the amount of per-flow state that needed to be maintained at each node in the packet-forwarding path. As a result, the IETF created a new service called Differentiated Services (DiffServ or DS). DiffServ is a set of technologies that allows network service providers to offer services with different kinds of network QoS objectives to different customers and their traffic streams [CHGU02] [BBCD98].

3.3.1 Service Level Agreement and Traffic Conditioning Agreement

A service level agreement (SLA) is a formal definition of the relationship that exists between two organizations, usually a supplier of services and its customer. The SLA is a contract that

specifies the forwarding service a customer should receive. It can be dynamic or static. Static SLAs are negotiated on a regular basis (e.g. monthly or yearly). Dynamic SLAs use a signaling protocol to negotiate the service on demand. The SLA typically contains:

- The type and nature of service to be provided, which includes the description of the service to be provided.
- The expected performance level of service (e.g. service availability).
- The process for reporting problems with the service (e.g. it could include information about the party who is responsible for problem resolution).
- The time frame for response and problem resolution, which specifies a time limit by which someone would start investigating a problem that was reported, etc.
- The process for monitoring and reporting the service level (e.g. how performance levels are monitored and reported).
- The credit, charges, or other consequences for the service provider on not meeting its obligation (e.g. failing to provide the agreed-upon service level).
- Escape clause and constraints, including the consequences if the customer does not meet his obligations.

An SLA also specifies the traffic conditioning agreement (TCA) that contains the rules used to realize the service. TCA specifies classifier rules and any corresponding traffic profiles and metering, marking, discarding, and/or shaping rules that are to apply to the traffic streams selected by the classifier. A TCA includes all the traffic conditioning rules explicitly specified within the SLA along with all of the rules implicit in the service level agreements and/or the service provisioning policy of the network domain.

A traffic profile specifies the temporal properties of a traffic stream selected by a classifier. It provides rules for determining whether a particular packet is in profile or out of profile. As packets enter the domain (DiffServ domain), they will be classified into a traffic aggregate by the specified filter at the domain ingress interface of the border router. The filter must be associated with a traffic profile that specifies the committed information rate (CIR) and a description on how it is to be measured. CIR is the minimum amount of bandwidth "guaranteed" for the service.

3.3.2 Basic Architecture for Differentiated Services

DiffServ divides a network into several domains, as shown in Figure 8. A DS domain [BBCD98] is a contiguous set of nodes that operate with common sets of service provisioning policies and per-hop behavior (PHP) group definition. The PHP is an externally observable forwarding treatment applied to all packets that belong to the same service class or behavior aggregate (BA). A DS domain has a well-defined boundary, and there are two types of nodes associated with a DS domain: boundary nodes and interior nodes.

Boundary nodes connect the DS cloud with other DS clouds, as shown in Figure 7. Interior nodes are connected to other interior nodes or boundary nodes, but they must be within the same DS domain. DS boundary nodes function as both DS ingress and egress nodes for different directions of traffic flows, as shown in Figure 7.

- The DS ingress node is responsible for classification, marking, and possibly conditioning of ingress traffic, it classifies each packet, based on the examination of the packet header, and then write the Differentiated Service Code Point (DSCP) to indicate one of the PHB groups supported with the DS domain.
- The DS egress node may be required to perform traffic conditioning functions on traffic forwarded to a directly connected peering domain.

Both ingress nodes and egress nodes enforce the TCAs between their own DS domain and the other domains they connect to. DS interior nodes select the forwarding behavior applied to each packet, based on the examination of the packet's PHB. They map DSCP to one of the PHP groups supported by all of the DS interior nodes within the DS domain.

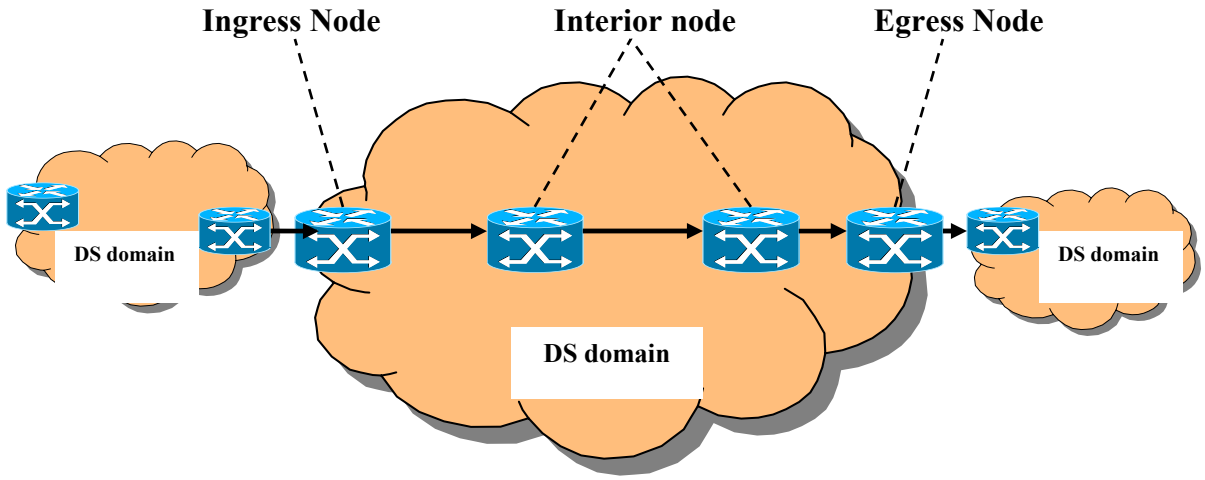


Figure 7: Differentiated service domain.

In DiffServ, the IPv4 ToS octet was renamed the DS byte and was given a new meaning for each of its bit. The DS byte is divided into two subfields, as shown in Figure 8:

- The first six bits are called the Differentiated Service codepoint (DSCP). The DSCP is used by a node to select the per-hop behavior (PHB) that a packet experiences at each hop within the DS domain.
- The other two bits are currently unused (reserved for future use).

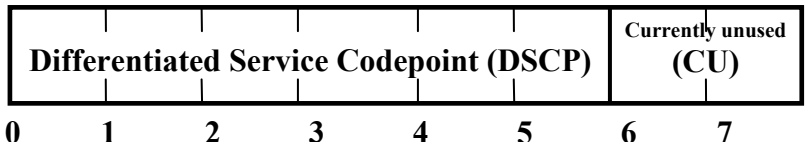


Figure 8: DiffServ field in the 8-bit ToS.

3.3.3 Packet Classifier and Traffic Conditioner

A packet classifier selects packets in a traffic stream based on the content of the packet header. There are two types of packet classifiers [BBCD98]. A behavior aggregate (BA) classifier which selects packets based on the value of the DSCP. A multifield classifier (MF) which selects packets based on a combination of the value of one or more header fields. These fields can include the source address, destination address, DS Field, protocol ID,

source port, destination port, or other information (e.g. incoming interface). After the classifier identifies packets that match some rules, the packet is sent to a traffic conditioner. A traffic conditioner consists of a meter, a marker, and a shaper/dropper [BBCD98], as show in Figure 9.

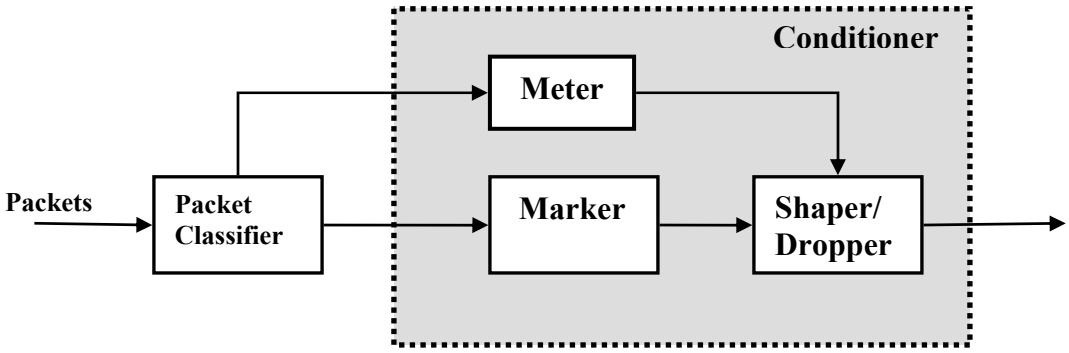


Figure 9: Packet classifier and traffic conditioner.

- *A meter* measures a traffic stream to determine whether a particular packet from the stream is in-profile or out-of-profile. The meter passes the in-profile or out-of-profile state information to conditioning elements so that different conditioning actions can be applied to these packets.
- *A marker* writes the DS Field of a packet header to a specific DSCP value, so that the packet is assigned to a particular DS behavior aggregate.
- *A shaper* delays some or all packets in a traffic stream to bring the stream into conformance with its profile. A *dropper* polices the incoming stream by dropping packets that are out of profile.

3.3.4 Per-hop Behaviors

A per-hop behavior (PHB) is a description of the externally observable forwarding behavior of a mode applied to a particular behavior aggregate. The PHB is the means by which a DS node allocates its resources to different behavior aggregates. PHBs are defined in terms of the behavior characteristics that are relevant to a provider's service provisioning policies. A specific PHP may be defined in term of the amount of resources allocated to the PHP (buffer size and link bandwidth), the relative priority of the PHB compared with other PHBs, or the observable traffic characteristics (delay, jitter, and loss). DS has defined two PHBs,

Expedited Forwarding (AF) PHP and Assured Forwarding (AF) PHB. EF [JNPO99] provides low cost, low delay, low jitter, assured bandwidth, and end-to-end service. AF [HBWW99] is a group of PHBs designed to ensure that packets are forwarded with a high probability of delivery, as long as the aggregate traffic in the forwarding class does not exceed the subscribed information rate. If ingress traffic exceeds its subscribed information rate, then out-of-profile traffic is not delivered with as high probability as traffic that is in-profile.

3.3.5 Discussion

The main idea behind the DiffServ model is that the deployment of multiple queues on a port allows a router to service certain traffic before other types of traffic, and thus isolates congestion to a subset of a router's queues. The queue-servicing algorithm arbitrates a queue's access to the link. The deployment of many queues on a port allows some queues to experience congestion while other queues do not. So, the DiffServ model can be described as a mechanism that tries to insure that the important traffic has access to network resources and does not experience congestion. This approach is based on the assumption that it is acceptable for less important traffic not to have access to network resources during periods of congestion.

In our proposed TE algorithm, we did not take DiffServ into account. All traffic has the same service level. Moreover, as our work focuses on inter-domain traffic, the SLA may be different among domains based on their respective policies and business interests. Applying DiffServ in our TE scheme needs more investigation.

One possible way to introduce DiffServ into our scheme would be to give each lightpath a different service level. So, each established lightpath to a specific destination may provide a different service level. In this case, the AS can establish a lightpath based on the service level agreement with the destination AS.

Chapter 4

IP and Wavelength Routing Networks

The Internet Protocol (IP) is a part of the Internet transfer control protocol (TCP)/IP suite [DAVI97]. It has been treated as the foundation of the Internet. Moreover, IP defines the basic unit that can be sent through an Internet. Routing protocols are protocols that exchange routing information between the different network nodes. Routers forward traffic based on network layer logical destination addresses. In this chapter we will give a general description of routing in the Internet and we will focus on inter-domain routing.

As wavelength routing constitutes one of the most fundamental elements of the control and management of the optical network, this chapter also provides an introduction to the routing and wavelength assignment (RWA) problem in optical networks.

4.1 Routing in Datagram Networks.

Routing is the process of finding a path from a source to a destination in the network. The two basic categories or routing algorithm are static routing and adaptive routing. In a *static routing* algorithm, the routing table is initialized during system start-up. Static routing is not suitable for practical use in data networks because it is incapable of adapting to changes in the network. Adaptive routing algorithms are able to adopt their routing information dynamically to current network characteristics. Adaptive routing algorithm can be categorized into either centralized or distributed. In centralized algorithms, routing decisions are made by a central node in the network and then distributed to other nodes. In distributed routing, each node in the network (router) makes an independent routing decision based on the information available to it. Distributed routing can be categorized to link-state algorithm and distance vector algorithm.

In a link-state approach (e.g. used by the OSPF routing protocol), every node in the network must maintain complete network topology information. Then, each node finds a

route for a connection request, based on the local information. When the network state changes, all the nodes must be informed by broadcasting an update messages to all nodes in the network.

In the distance vector approach (e.g. used by the BGP routing protocol), the algorithm makes the assumption that each network node is aware of distance between the node itself and all other nodes in the network. The routing tables of the distance vector algorithm contain an entry for each possible destination node in the network. The routing table will initially use the value infinity for all non-neighboring systems. This indicates that the distance to a particular system is not known and that it therefore cannot be reached. When a router receives a routing message (e.g. Next-hop, AS-path), it compare the information received with the available information in the routing table. If the new information is different, the router replaces the old information with the new one.

4.1.1 Inter-domain Routing

The current Internet is partitioned in sub-networks or domains, also called ASs, as shown in Figure 10. An AS is defined as a set of routers or networks that are technically managed by a single networks administration (e.g. corporate network, campus network or Internet Service Provider). The routing protocols running inside an AS are called Interior Gateway Protocols (IGP) (e.g. IS-IS, OSPF). The ASs exchange routing information using Exterior Gateway Protocols (EGP). BGP is the current routing protocol running between ASs.

Each AS is identified by a number, a globally unique AS number that is represented by a 16-bit integer and thus limited to 65,000 numbers. This number is obtained from the American Registry for Internet Number (ARIN). These numbers are in the ranging from 1-65535. Numbers 64512–65535 are reserved for private use and should never appear in an Internet routing table.

The AS can be categorized into three categories based on their connection to teach other, and the traffic they accept. They are categorized as following:

- *A Stub AS (or single–homed AS)* has only a single connection to the outside world or another AS.
- *A Multihomed AS has* multiple connections to the outside world but refuses to carry transit traffic. It carries only local traffic.
- *A Transit AS has* multiple connections to the outside world and can carry transit and local traffic.

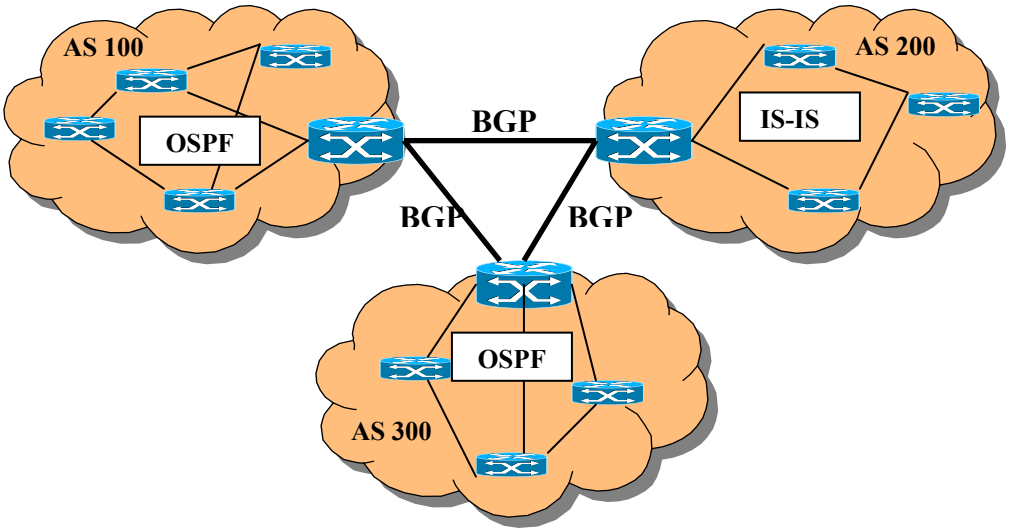


Figure 10: Current internet architecture.

4.1.2 Border Gateway Protocol

Border Gateway Protocol (BGP) is the current inter-autonomous-systems routing protocol of the Internet. BGP is described as a reachability routing protocol, because it exchanges network reachability information with other BGP speakers (AS). It does not convey any information about the optimal topology, quality of service or bandwidth of a particular route.

BGP uses TCP [POST81] as its transport protocol. Two BGP routers form a TCP connection between one another (peer routers) and exchange messages to open and confirm

the connection parameters between them. BGP routers exchange network reachability information, they exchange the full paths (BGP AS-Numbers) from source to destination that each router should take in order to reach the required destination.

Any two routers that have formed a TCP connection and opened a BGP session are called peers. BGP peers initially exchange their full BGP routing tables. After this exchange, incremental updates are sent as the respective routing tables change. BGP keeps a version number of the BGP table, which should be the same for all its peers. The version number changes whenever BGP updates the table due to routing information changes. KEEPALIVE messages are sent to ensure that the connection is alive between the BGP peers while NOTIFICATION messages are sent in response to errors or special conditions.

Routes in BGP are stored in the Routing Information Bases (RIBs). The BGP-RIB has three routing tables, Adj-RIBs-In, Loc-RIB, and Adj-RIBs-Out. Routes that will be advertised to other BGP speakers must be present in the Adj-RIB-Out; routes that will be used by the local BGP speaker must be present in the Loc-RIB, and routes that are received from other BGP speakers are present in the Adj-RIBs-In. BGP uses the following criteria, to select a path for a destination:

1. If the path received specifies a next hop that is inaccessible, drop the update (do not update RIB).
2. Prefer the path with the largest weight. *Weight* is an attribute that is local to a router. The weight attribute is not advertised to neighboring routers. If the router learns about more than one route to the same destination, the route with the highest weight will be preferred.
3. If the weights are the same, prefer the path with the largest local preference. The *local preference* attribute is used to prefer an exit point from the local AS. Unlike the weight attribute, the local preference attribute is propagated throughout the local AS. If there are multiple exit points from the AS, the local preference attribute is used to select the exit point for a specific route.
4. If the local preferences are the same, prefer the path that was originated by BGP running on this router (the path that was originally advertised by this node).

5. If no route was originated, prefer the route that has the shortest AS_path.
6. If all paths have the same AS_path length, prefer the path with the lowest origin type. The *origin attribute* indicates how BGP learned about a particular route. The origin attribute can have one of three possible values:
    - IGP - The route is interior to the originating AS.
    - EGP - The route is learned via the Exterior-BGP (EBGP).
    - Incomplete - The origin of the route is unknown or learned in some other way.

7. If the origin codes are the same, prefer the path with the lowest MED attribute.
8. If the paths have the same MED, prefer the external path over the internal path.
9. If the paths are still the same, prefer the path through the closest IGP neighbor.
10. Prefer the path with the lowest IP address, as specified by the BGP router ID.

As we are proposing TE for inter-domain optical networks, BGP can be used to exchange resource availabilities (wavelengths) between ASs and to exchange lightpaths status (e.g. congested lightpath) between ASs. In [KHIR04], a modification over BGP was proposed to make the protocol capable of exchanging resource availability in the optical domain.

4.2 Wavelength Routing Networks

In optical transport networks, routing protocols must be extended to distribute information about the network topology, resource availability, and network status. For IGP, The IETF proposed various extensions for OSPF and IS-IS routing in optical networks [KRBD02] [KORE03]. For EGP, different architectures [FPHL02] [AHHC00] have been proposed for BGP routing in optical networks.

4.2.1 Optical Border Gateway Protocol

For BGP, to work as a routing protocol in optical networks, a modification has to be made. BGP does not have any functionality and attributes to maintain information on optical resources across the network. The essential requirements for BGP to exchange this information are:

- It must provide a set of attributes to exchange the availability of wavelengths.
- It must provide a mechanism to propagate up to date information on the resource status and availability in the network.

Different extensions of BGP have been proposed, called Optical Border Gateway Protocol (OBGP) [AHHC00] [FPHL02] [KHIR04]. They provide routing, resource management and signaling of wavelengths between ASs. The protocol controls the OXC to establish and tear down wavelengths. As BGP does not have any messages or attributes to maintain information on lightpaths across networks, modifications to BGP to carry this information were proposed. The advantages of making BGP act as a routing and signaling protocol are [FSPH01]: When combining routing and signaling into one protocol, the signaling function can gain some features from the existing routing functions, (e.g. TCP connection can be used for both signaling and routing messages), and BGP has AS-PATH attributes that gives end-to-end view, this attribute is required for setting up a lightpath.

Different architectures [FPHL02] [AHHC00] have been proposed for BGP to work as a routing and a signaling protocol. The main difference between these architectures is that, one proposes a new separate message to be added to BGP messages. This message will be responsible of the exchange of resource availability and the establishment of lightpaths. Architecture proposes a modification of the BGP UPDATE message to be capable of exchange optical resource information and setup lightpath. However, all the proposed architectures have the same objective to make BGPs peer capable of exchanging resource availabilities (Lambdas) and to make BGP work as a signaling protocol.

4.2.2 Routing and Wavelength Assignment.

One of the requirements in the control plane of a wavelength-routed optical network (WRON) is to setup and teardown optical connections that are established by lightpaths between source and destination nodes. A lightpath is an all-optical path connecting a source and a destination such that each link of the path uses the same wavelength. This occurs when the OXCs do not have wavelength conversion capability. But, when the lightpath allocation is done link by link so that the end-to-end optical path is a chain of lightpaths of different

wavelengths, it is called a virtual wavelength path [CFZH96]. In this case, the OXCs provide a wavelength conversion capability.

Depending on the nature of the traffic in the network, a lightpath can be classified as static or dynamic. When the nature of the traffic pattern is static, a set of lightpaths is established all at once and remain in the network for a long period of time. For the dynamic traffic scenario where the traffic pattern changes rapidly, the network has to respond to traffic demands quickly and economically. For that, a lightpath is setup for each connection request as it arrives, and the lightpath is torn down after some amount of time. This is called dynamic lightpath establishment.

Routing and wavelength assignment (RWA) [ZJMU00] in the optical domain is a key process of provisioning lightpaths in response to connection request. Given a constraint on the number of wavelengths, it is necessary to determine both the routes over which these lightpaths are to be established and the wavelengths that are to be assigned to these routes. This problem is referred to as the RWA problem. The RWA process can be classified as dynamic or static.

The process of RWA is said to be static if it is designed on the premise that the traffic distribution in a network will not be changed while the network is in operation. In this case, a traffic matrix is given, which specifies the bandwidth demand between any node pair in the network. Then lightpaths are allocated within the network according to the traffic matrix.
The dynamic RWA is aimed at satisfying lightpath setup requests one at a time with a goal of maximizing the probability of successful allocation for subsequent demands. The most commonly used performance index in the dynamic network is the blocking probability under specific potential traffic load for each source-destination pair.

Chapter 5

Traffic Engineering

5.1 Definition of TE

TE [ACEW02] is defined as that aspect of Internet network engineering that deals with the issue of performance evaluation and performance optimization of operational IP networks. The main objective of TE is to enhance the performance of a network at the traffic and resource level. This can be achieved by addressing traffic-oriented performance requirements, while utilizing network resources economically and reliably. A major goal of TE is to provide efficient and reliable network operations. This reliability can be achieved by minimizing unexpected service outages arising from errors, faults, and failures occurring within the network infrastructure.

TE performance optimization can be achieved by capacity management and traffic management. Capacity management includes capacity planning, routing control and resource management (e.g. link bandwidth, buffer space). While, traffic management includes network nodes traffic control functions, such as queue management, scheduling, policing, and traffic flow control through the network.

Measurement can be achieved by running a measurement program on a host attached to the network, and it may also require the use of special hardware such as network interface cards with the ability to measure packet arrival times very accurately. These methods can allow us to measure the link behavior or traffic flows.
TE Performance evaluation can be achieved by using different techniques such as analytical methods and simulation-based methods measurements. For analytical and simulation methods, network nodes and links can be modeled to depict dynamic and behavioral traffic characteristics and network performance evaluation.

5.2 Activities Required for TE

A TE process model [ACEW02] describes a set of actions that the system of TE must follow to optimize the network performance [AMAO99] [Awdu99]. This model has the following components:

Measurement: Measurement is an important component of the TE function. The network performance can only be determined through measurement. Traffic measurement is an essential tool to guide the network administrator of large IP networks in detecting and diagnosing performance problems, and evaluating potential control actions. The data measurement is analyzed and a decision based on the analysis is taken for network performance optimization. Measurement is needed to determine the quality of services and to evaluate TE policies.

Modeling, Analysis, and Simulation: Modeling and analysis are important aspects for TE. A network model is an abstract representation of the network that captures the network features, attributes and characteristics (e.g. link and nodal attributes). A network model can facilitate analysis or simulation, and thus can be useful to predict the network performance.

Network modeling can be classified as structural or behavioral module. Structural modules focus on the organization of the network and its components. Behavioral modules focus on the dynamics of the networks and its traffic workload. As stated in [ACEW02], because of the complexity of realistic quantitative analysis of network behavior, certain aspects of network performance studies can only be conducted effectively using simulation.

Optimization: Network performance optimization can be called corrective when a solution to a problem is made, or perfective, where an improvement to the network performance is made, even when there is no problem [ACEW02]. Many actions could be taken such as adding additional links, increasing link capacity or adding additional hardware. Planning for future improvement in the network (e.g. network design, network capacity or network architecture) is considered as a part of network optimization.

5.3 Criteria for Selecting the Best Traffic Route

Traditionally, there have been three parameters that describe the quality of a connection: bandwidth, delay, and packet loss. A connection with high bandwidth, low delay, and low packet loss is considered to be better than one with low bandwidth, high delay, and high packet loss. The following parameters can be considered when selecting the best traffic route [BEIJ02]:

Congestion: Congestion decreases the available bandwidth and increases delay and packet loss. It is important to avoid routes over congested paths.

Distance: Two routes may have different paths. Some networks interconnect only at relatively few locations, so they may have to transport traffic over long distances to get it to its destination. Others have better interconnection, so the traffic does not have to take a detour. There may be reasons not to prefer the more direct route, such as lower bandwidth or congestion, but generally a shorter geographic path is better.

Hops: The number of hops (e.g. routers) that shows up on the path to the destination increases the delay. Each hop potentially adds additional delay, because packets have to wait in a queue before they are transmitted, and the extra equipment in a path means that a failure somewhere along the way is more likely. So, paths with fewer hops are better.

5.4 Congestion Control

Congestion in a packet switching network is a state in which the performance of the network degrades because of the saturation of network resources. Congestion could result in degradation of service quality to users. To avoid congestion, certain mechanisms have to be provided; such mechanisms are usually called congestion control.

5.4.1 Categories of Congestion Control

Congestion control policies can be categorized differently based on the objective of the policy, the time period of applying the policy, and the action taken to avoid congestion. In the following we will explain some of these policies.

5.4.1.1 Response Time Scale

Response time scale can be categorized as one of the following: long, medium and short. *In the long time scale*, expansion of the network capacity is considered. This expansion is based on estimates of future traffic demands, and traffic distribution. Because the network elements are expensive, upgrades take place in a long time scale between week to month or years. *In the medium time scale*, network control policies are considered (e.g. adjusting the routing protocol parameters to reroute traffic from a congested network node). These policies are mostly based on a measurement, and the actions are applied during a period of minutes to days. *In the short time scale*, packet level processing and buffer management functions in routers are considered (e.g. active queue control schemes in TCP traffic using Random Early Detection RED) [FLJA93].

5.4.1.2 Reactive Versus Preventive

In reactive congestion control, congestion recovery takes place to restore the operation of a network to its normal state after congestion has occurred. Control policies react to existing congestion problems to remove or reduce them. In preventive congestion control, keeping the operation of a network at or near the point of maximum power is the main objective, so congestion will never occur. Control policies applied to prevent congestion are based on estimates and predictions of possible congestion appearance.

5.4.1.3 Supply Side versus Demand Side

Increasing the capacity in the network is called a supply side congestion control. Supply side control is achieved by increasing the network capacity or balancing the traffic, (e.g. capacity planning to estimate traffic workload). For demand side control, policies are applied to regulate the offered traffic to avoid congestion (e. g. traffic shaping mechanism is used to regulate the offered load).

5.4.2 Control Policies

Different congestion control policies have been proposed to deal with congestion in networks [MARA91]. Generally speaking, these policies differ in the use of control messages. The following will describe some of them.

Source Quench: Source Quench is the current method of congestion control in the Internet. When a network node responds to congestion by dropping packets, it could send an Internet Control Message Protocol Source Quench message (ICMP) [POST81] to the source**,** informing it of packet drop. The drawback of this policy is that it is a family of varied policies. The major gateway manufacturers have implemented various source quench methods. This variation makes the end-system user, on receiving a Source Quench**,** uncertain of the cause in which the message was issued (e.g. heavy congestion, approaching congestion, burst causing massive overload).

Random Drop: Random Drop is a congestion control policy intended to give feedback to users whose traffic congests the gateway by dropping packets. In this policy, randomly selected packets for a particular user, from incoming traffic, will be dropped. A user generating much traffic will have much more packets drop than the user who generate little traffic. The selection of packets drop in this policy is completely uniform. Random Drop can be categorized as Congestion recovery [MANK90] or congestion avoidance [JACO89]. Congestion recovery tries to restore an operating state, when demand has exceeded capacity. Congestion avoidance is preventive in nature. It tries to keep the demand on the network at or near the point of maximum power, so that the congestion never occurs.

Congestion Indication: The so-called Congestion Indication policy [RAJA87] uses a similar technique as the Source Quench policy to inform the source gateway of congestion. The information is communicated in a single bit. The Congestion Experienced Bit (CEB) is set in the network header of the packets already being forwarded by a gateway. Based on the value of this bit, the end-system user should make an adjustment to the sending window.

The Congestion Indication policy works based upon the total demand on the gateway. For fairness the total number of users causing the congestion is not considered. Only users who

are sending more than their fair share (allowed bandwidth) should be asked to reduce their load, while others could attempt to increase their load where possible.

Fair Queuing: Fair queuing is a congestion control policy where separate gateway output queues are maintained for individual end-systems on a source-destination-pair basis. When congestion occurs, packets are dropped from the longest queue. At the gateway, the processing and link resources are distributed to the end-systems on a round-robin basis [NAGL 85]. (round-robin is an arrangement of choosing all elements in a group equally in a circular). Equal allocations of resources are provided to each source-destination pair.

A Bit-Round Fair Queuing algorithm [DKSH89] was an improvement over the fair queuing described in [NAGL 85]. It computes the order of service to packets using their lengths, by using a technique that emulates a bit-by-bit round-robin discipline. In this case, long packets do not get an advantage over short packets. Otherwise the round-robin would be unfair.

Stochastic Fairness Queuing (SFQ) [MCKE91] is a similar mechanism to Fair Queuing. SFQ looks up the source-destination address pair in the incoming packets and locates the appropriate queue that packet will have to be placed in. It uses a simple hash function to map from the source-destination address pair to a fixed set of queues. The price paid to implement SFQ is that it requires a potentially large number of queues.

5.5 MPLS, and GMPLS Traffic Engineering

Network control (NC) can be classified as centralized or distributed [XYDQ01]. In centralized network control, the route control and route computation commands are implemented and issued from one place. Each node in the network communicates with a central controller and it is the controller's responsibility to perform routing and signaling on behalf of all other nodes. In a distributed network control, each node maintains partial or full information about the network state and existing connections. Each node is responsible to perform routing and signaling. Therefore, coordination between nodes is required to alleviate the problem of contention.

Since its birth, the Internet (IP network) has employed a distributed NC paradigm. The Internet NC consists of many protocols. The functionality of resource discovery and management, topology discovery, and path computation and selection are the responsibility of routing protocols. Multiprotocol label switching (MPLS) [RVCA01] has been proposed by IETF, to enhance classic IP with virtual circuit-switching technology in the form of label-switched path (LSP). MPLS is well known for its TE capability and its flexible control plane.

Then, IETF proposed an extension to the MPLS-TE control plane to support the optical layer in optical networks, this extension is called the Multiprotocol Lambda Switching (MPλS) control plane [AWDU01]. Another extension to MPLS was proposed to support various types of switching technologies. This extension is called Generalized Multi-Protocol Label Switching (GMPLS) [BANE01]. GMPLS has been proposed in the Control and Measurement Plane working group in the IETF as a way to extend MPLS to incorporate circuit switching in the time, frequency and space domains. Next we will describe MPLS and the GMPLS architecture in more details.

5.5.1 MPLS Traffic Engineering Architecture

Multiprotocol label switching (MPLS) [RVCA01] is a hybrid technology that provides very fast forwarding at the cores and conventional routing at the edges. MPLS working mechanism is based on assigning labels to packets based on forwarding equivalent classes (FEC) as they enter the network. A FEC identifies of a group of packets that share the same requirements for their transport. All packets in such a group are provided the same treatment en route to the destination.

Packets that belong to the same FEC at a given node follow the same path and the same forwarding decision is applied to all packets. Then packets are switched through the MPLS domain using simple label lookup. Each FEC may be given a different type of service. At each hop, the routers and switches use the packet labels to index the forwarding table to determine the next-hop router and a new value for the label. This new label replaces the old one and the packet is forwarded to the next hop. As each packet exits the MPLS domain, the

label is stripped off at the egress router, and then the packet is routed using conventional IP routing mechanisms.

The router that uses MPLS is called a label switching router (LSR). A LSR is a high-speed router that participates in establishment of LSPs using an appropriate label signaling protocol and high-speed switching of the data traffic based on the established paths.
MPLS-TE has the following components and functionalities, as shown in Figure 11:

- The routing protocol (e.g. OSPF-TE, IS-IS TE) [KRBD02] [KORE03], collects information about the network connectivity (this information is used by each network node to know the whole topology of the network) and carries resource and policy information of the network. The collected information is used to maintain: The so-called Link-state database which provides a topological view of the whole network and the TE database which stores resource and link utilization information. The databases are used by the path control component. A constrained Shortest Path First (CSPF) is used to compute the best path.
- A signaling protocol (e.g. RSVP-TE or CR-LDP) [SBBB01] is used to set up LSP along the selected path through the network. During LSP setup, each node has to check whether the requested bandwidth is available. This is the responsibility of the link admission control that acts as an interface between the routing and signaling protocol. If bandwidth is available, it is allocated. If not, an active LSP might be preempted or the LSP setup fails.

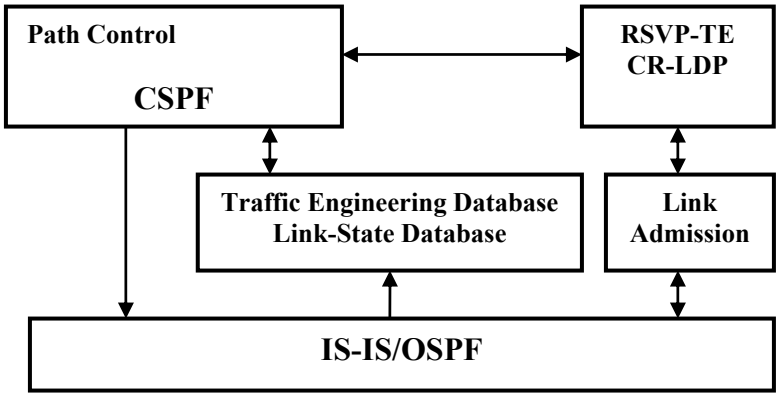


Figure 11: MPLS-TE functional components.

5.5.2 MPλS/GMPLS Control Plane

IETF proposed an extension to the MPLS-TE control plane to support optical layers in optical networks; this extension is called the multiprotocol lambda switching (MPλS) control plane [AWDU01]. In an MPLS network, the label-switching router (LSR) uses the label-swapping paradigm to transfer a labeled packet from an input port to an output port. In the optical network, the OXC uses switch matrix to switch the data stream (associated with the lightpath) from an input port to an output port. In both LSR and OXC, a control plane is needed to discover, distribute, and maintain state information and to instantiate and maintain the connections under various TE roles and policies.

The functional building blocks of the MPλS control plane are similar to the standard MPLS-TE control plane. The routing protocol (e.g. OSPF or IS-IS) with optical extensions, is responsible for distributing information about optical network topology, resource availability, and network status. This information is then stored in the TE database. A constrained-based routing function acting as a path selector is used to compute routes for LSPs through the optical mesh network. Signaling protocols (e.g. RSVP-TE or CR-LDP) are then used to set up and maintain the LSPs by consulting the path selector.

Another extension to the MPLS control plane is proposed to support various types of optical and other switching technologies. This extension is called Generalized Multi-Protocol Label Switching (GMPLS) [BANE01] [RVCA01]. In the GMPLS architecture, labels in the forwarding plane of Label Switched Routers (LSRs) can route the packet headers, cell boundaries, time slots, wavelengths or physical ports. The following switching technologies are being considered, as shown in Figure 12.

Packet switching: The forwarding mechanism is based on packet. The networking gear is an IP router.

Layer 2 switching: The forwarding mechanism is based on cell or frame (Ethernet, ATM, and Frame Relay).

Time-division multiplexing (time slot switching): The forwarding mechanism is based on the time frames with several slots and data is encapsulated into the time slots (e.g. SONET/SDH).

Lambda switching: λ switching is performed by OXCs.

Fiber switching: Here the switching granularity is a fiber. The networking gears are fiber-switch-capable OXCs.

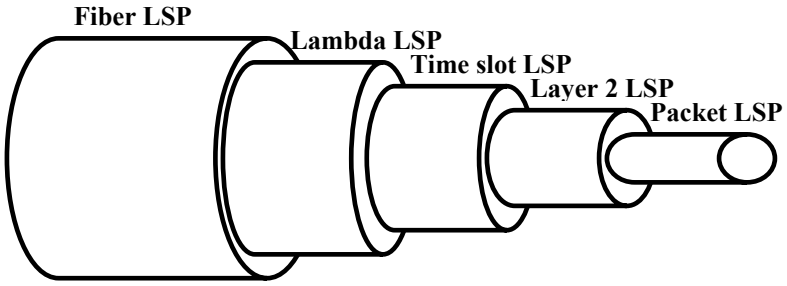


Figure 12: GMPLS label-stacking hierarchy.

The difference between MPλS and GMPLS is that the MPλS control plane focuses on Lambda switching, while GMPLS includes almost the full range of networking technologies.

5.6 Related Works

Most researches in the area of inter-domain TE networks apply intra-domain technologies (e.g. MPLS, GMPLS) over the inter-domain architecture [RVCA01] [AMAO99] [OHAC01] [SYMG02]. A modification of BGP was proposed for route distribution over the backbone, while MPLS is responsible for packet forwarding [RORE99]. BGP/GMPLS was another proposed solution for setting up circuit Label Switched Paths (LSPs) that span multiple ASs [XBXU02], where a modification of BGP was introduced to disseminate appropriate topology information, and GMPLS signaling was extended for inter-domain connections. Moreover, a third-party agent, called Bandwidth Management Point (BMP) [OHAC01], was proposed to exchange resource availability between ASs and to set up inter-domain LSPs using MPLS over Differentiated Service (DS) [BBCD98], while the Simple Inter-domain Bandwidth Broker Signaling Protocol (SIBBS) was used as inter-domain signaling protocol.

5.6.1 BGP and MPLS, GMPLS Technology

IETF, RFC 3107 [RERO01], specifies the way in which BGP can be used to distribute MPLS labels that are mapped to the route. If two adjacent routers in different domains are Label Switched Routers (LSRs) and are BGP peers, then, there is no need for label distribution protocol to distribute labels. Instead, BGP will distribute MPLS labels. A label mapping information for a particular route is piggybacked in the same BGP messages by using the BGP multiprotocol extensions attribute [BCKR98]. The label is encoded into the NLRI field and the Subsequent Address Family Identifier (SAFI) field, is used to indicate that NLRI contains a label. Two BGP routers must not advertise this capability of label distribution unless there is a LSP between them.

GMPLS technology was proposed for the inter-domain architecture [XBXU02] to support different types of clients (e.g. SONET/SDH, ATM, or IP). A Client Access Group (CAG) family of attributes was proposed and encoded in BGP-NLRI. The CAG address is a unique address for each domain and it is used to support different types of clients. This address is external and may be different from the client address that is used for internal routing. In this case, BGP will distribute LSPs and CAP attributes to its neighbors.

5.6.2 Third Party Agent

Another proposal to set up inter-domain LSPs was presented in [OHAC01]. But instead of using BGP to distribute MPLS-LSPs between domains, a bandwidth border agent called Bandwidth Management Point (BMP) takes this responsibility. BMP helps setting up inter-domain LSP's, while the Simple Inter-domain Bandwidth Broker Signaling Protocol (SIBBS) [QBon00] is used as inter-domain signaling protocol.

The BMP allocates and controls the bandwidth sharing between different domains and serving DSCP service classes. BMP calculates current service demands, available network resources, and their values. Using such measures, the BMP makes a decision or sets a policy for admission control, network resource provisioning, service level specification configuration, and bandwidth exchange. Based on its decision, the BMP acts as bandwidth

broker agent in each AS. The BMP interacts with MPLS to setup LSP's in a domain and communicates with peer BMP's to manage the inter-domain resources. A BMP consists of the following blocks:

- *The Bandwidth Measurement Base (BMB)* is a Management Information Base (MIB) that measures the traffic load and collects statistics about the BMP domain.
- *A Management Creation Point (MCP)* relates BMB with the Service Level Agreements (SLAs) of interconnections.
- A so-called *Interior-BMP (I-BMP)* component manages the resources within the domain.
- *Exterior-BMP (E-BMP)* manages the interconnection between BMP's. It is responsible for resource allocation and provisioning among multiple domains.
- *Service-Provisioning-Points (S-P-P)* are network nodes and interfaces. I-BMP and E-BMP can configure these nodes using their own management information and decisions.
- SLS-Enforcement-Points (S-E-P) and Service Level Specification (SLS**)** are network nodes and interfaces used for inter-domain resource allocation based on the SLS. The I-BMP and E-BMP configure these points and perform the admission tasks.
- *User-BMP-Interconnection-Points (UBIP)* are points at each user interconnection-access networks. Individual service flows are generated and terminated by the end hosts connected to the user network.

The signaling protocol proposed for inter-domain connection establishment is an extension to SIBBS [QBon00].SIBBS is an inter-domain signaling protocol that is still underdevelopment. It is responsible for label distribution between ASs. The BMP in a domain can receive a Resource Allocation Request message (RAR) from a host in the domain, from a peer BMP or from a third-party agent acting as a host or application. Then BMP responds by sending a Resource Allocation Answer message (RAA).

We can conclude from the previously described proposals that applying intra-domain TE approaches such as MPLS and GMPLS is a quite possible solution for the inter-domain context where small modifications of BGP to carry LSPs are possible. But we do not agree with the proposal that was presented in [OHAC01], where the BMP agent and the SIBBS

signaling protocol are presented as a good solution to replace the current BGP Internet protocol. It will be difficult to make a complete change to the current Internet architecture without affecting the current users.

Chapter 6

Traffic Engineering Algorithm

As we saw in the section on proposed architectures for inter-domain TE, most proposals apply MPLS technology in inter-domain networks. MPLS is well known for its TE capability for intra-domain architecture, where the network is controlled by a single administration. In inter-domain networks, there is no single management over the network resources between domains. A competition over network resources is quite possible; each domain administrator makes his best efforts to serve his customers while ignoring the traffic demands for other domains. Moreover, MPLS, with its provision of paths through the network and the use of additional IP header fields in the classification of packets into forwarding equivalency classes (FECs), provides the ability to route different types of traffic along very different paths through the networks, all that is achieved at the expense of significant additional processing by the network node. In this thesis, no intra-domain technology is used.

Our proposed algorithm is responsible of making decisions of setting up or tearing down lightpaths based on dynamic traffic measurements. From the architectural point of view, the algorithm consists of two parameters that are responsible for detecting the congestion and link usage over an established lightpath, the proposed parameters are: **H**, and **L**. The H represents a high traffic load boundary. If the traffic load reaches this threshold, the system will attempt to establish a new lightpath. The **L** represents a low traffic mark. If the traffic load reaches this threshold, a lightpath will be torn down.

Static re-engineering of the networks is insufficient to keep up with the very dynamic traffic patterns that are observed. Moreover, measurement is an important component of the TE function. Dynamic traffic re-engineering using dynamic measurements is considered to be the optimal solution for congestion avoidance. The main function of the proposed algorithm is to dynamically react to traffic instability (e.g. congestion) by establishing new lightpaths when congestion occurs at the IP level or by tearing a lightpath down when it

becomes under-utilized. We consider a dynamic traffic measurement system which measures the traffic load over every established lightpath. At the end of each observation period, if the traffic load exceeds the assigned value of **H,** a new lightpath will be established. When the load becomes lower than **L**, the lightpath will be torn down, as shown in Figure 13.

The DPM in each network node is responsible of making a decision of establishing or tearing down a lightpath. The IP router is responsible for packets forwarding and packets dropping in case of buffer overflow, and OXC is responsible for establishing and tearing down a lightpath. Section 6.2 describes the interactions between the DPM, the router, and the OXC.

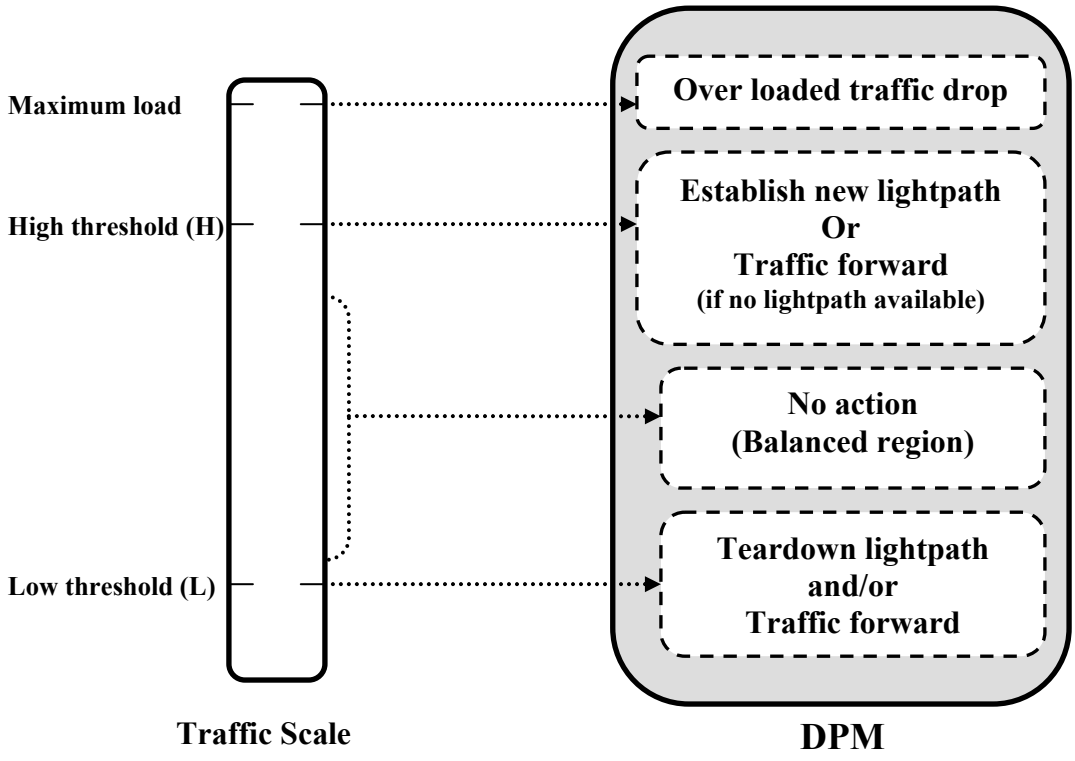


Figure 13: Decision process model.

6.1 Network Architecture

We assume a network consisting of multiple ASs or domains, as shown in Figure 14. Each domain is connected to its neighbor ASs using fiber link, where each optical link has a number of wavelengths. Each AS consists of multiple nodes, where each node could be an IP router, OXC or both. Each AS must have at least one boundary node connected to its neighbor AS where each boundary node consists of the following, as shown in Figure 15. *IP Router:* A device that is capable of routing IP packets. *Optical Cross Connect* (OXC): An OXC is a true system-level optical switch that can switch an optical data stream from an input port to an output port. Such a switch may utilize optical-electrical conversion at the input port and electrical-optical conversion at the output port, or it may be all-optical. We assume that OXCs do not provide a wavelength conversion capability.

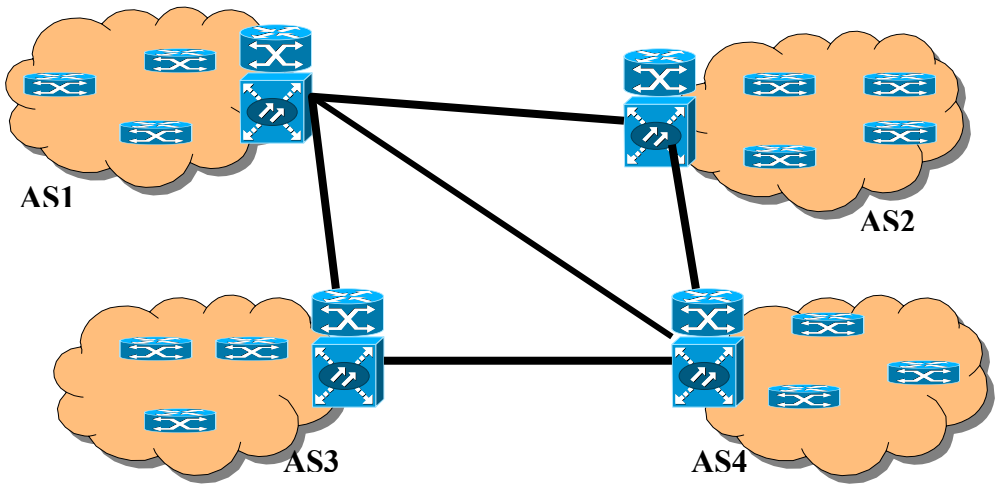


Figure 14: Network architecture.

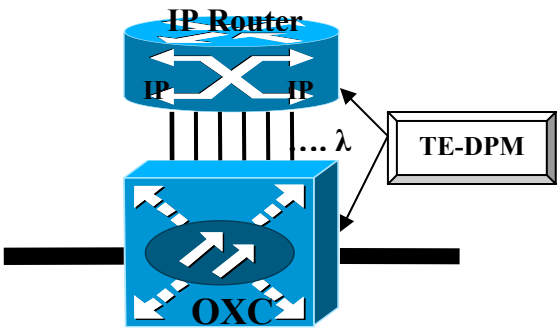


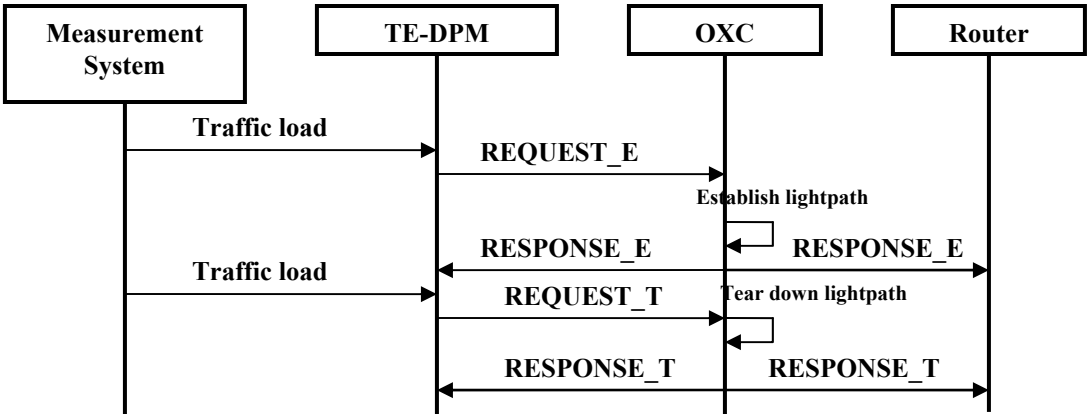
Figure 15: Network node architecture.

For each network node, the number of wavelengths between the router and OXC is fixed and may vary from one AS to the other. We assume that each AS is represented by a single network node. This node is the boundary node of the AS which is connected to other ASs. Each established lightpath has its own port on the router.

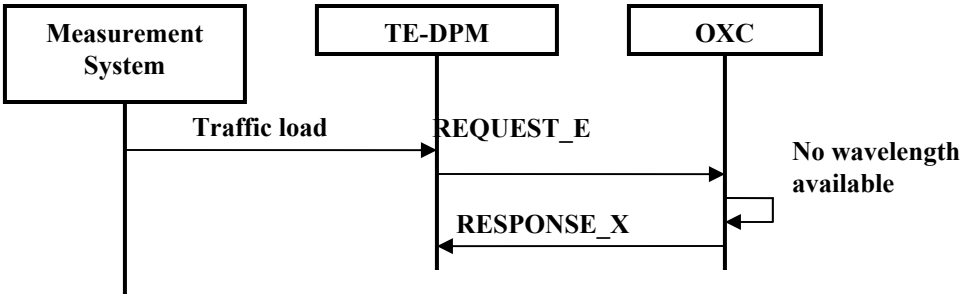
6.2 TE-DPM, Router and OXC Interaction

As we mentioned earlier, each network node consists of a TE-DPM, an IP router, and an OXC. The TE-DPM can be implemented inside the router or as a separate device. When the TE-DPM receives the information about the traffic load from the measurement system, the following can occur:

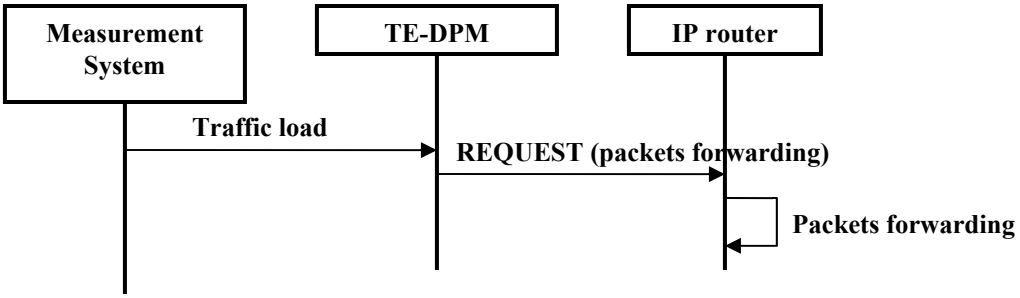
- The TE-DPM sends a REQUEST to the OXC for establishing or tearing down a light path (e.g. REQUEST_E message for lightpath establishing and REQUEST_T message for tearing down a lightpath). Then OXC sends a RESPONSE informing the DMP and the router that the request is complete (e.g. RESPONSE_E message for lightpath establishing and RESPONSE_T message for tearing down a lightpath), as shown in Figure 16-a.
- When there is a request for establishing a lightpath and there are no available wavelengths, the OXC sends a RESPONSE informing the TE-DPM about resource unavailability (e.g. RESPONSE_X message), as shown in Figure 16-b.
- If there is a need for a new wavelength but there is no wavelength available, the TE-DPM sends a REQUEST to the router for traffic forwarding, as shown in Figure 16-c



a) TE-DPM and OXC Interaction.



b) TE-DPM and OXC interaction.



c) TE-DPM and IP router interaction.

Figure 16: TE-DPM, IP Router and OXC interaction.

6.3 TE Decision Process Module

TE-Decision Process Module (TE-DPM) is the proposed algorithm for inter-domain TE in the optical network. After measuring the actual traffic load on a lightpath continuously (based on observation period), the measurement system [BMRU99] passes the measurement result to the DPM. The latter compares the traffic load value the assigned parameters value (**H, L**). Based on this process, DMP can make one or more of the following decisions: establish a new lightpath, tear down a lightpath, or traffic forwarding, as shown in Figure13.

6.3.1 System Parameters

The operation of the DPM is based on H and L. Based on the traffic load over the established lightpath and the assigned value of these parameters, the DPM will be able to determine if the lightpath is in a congestion state or in an underutilized state. The lightpath status at each router port could be in one of the following states, as shown in Figure 18**:**

- **Congestion state:** When the traffic load reaches the assigned value of H, DPM detects that the lightpath is in a congestion state. A decision of establishing a new lightpath or traffic forwarding will be taken.

- **Underutilized state:** when the traffic load reaches the assigned value of **L**, DPM detects that the lightpath is in an under-utilized state. A decision of tearing down this lightpath can be taken.

- **Balanced state:** When the traffic load is higher than the value of **L** and lower than the value of **H** (**L < Traffic Load < H**), DMP detects that the lightpath is in a balanced state (no congestion and no lightpath under-utilization). In this state, no change of the optical lightpath will be made.

6.3.2 Oscillation Interval

We call Oscillation factor the number of lightpaths established and torn down during a small observation period (Cycle) [LYSH00]. Frequently establishing and tearing down a lightpath in a small period of time, is undesirable behavior in the optical network. It is an indication of unstable network performance. In this perspective, to avoid this kind of network instability, we define an oscillation interval for the system parameters (L, H), as shown in Figure17. When the traffic load (α) reaches the assigned value of H or L, the DPM will not take a decision of establishing or tearing down a lightpath. Instead, the decision of establishing a lightpath will be taken when **α = A** or **α = C** and the decision of tearing down a lightpath will be taken when **α = D or α=B** (we tear down lightpath at B when we have more than one lightpath established to the same destination and these lightpaths are in balanced state).

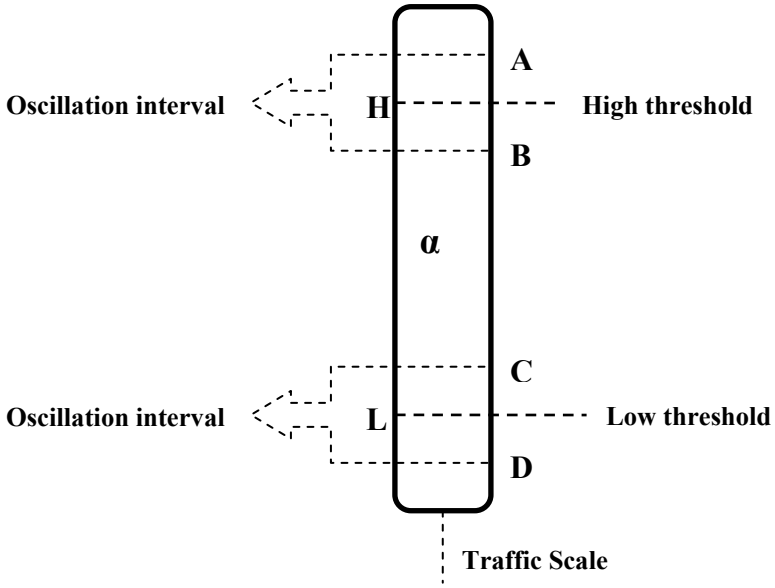


Figure 17: Oscillation interval.

6.3.3 Decisions Taken by the DPM

The proposed algorithm is based on dynamic traffic measurement. After each traffic observation time over an established lightpaths, the measurement system passes the measurement results to the DPM. The DPM compares the measurement result to the assigned value of H and L. Based on this comparison, the DMP detects the state of the established

lightpath. The lightpath could be in congested, underutilized or balanced state. Based on the lightpath state and resources availability, the DMP can make one or two of the following decisions: establishing a new lightpath, tearing down the given lightpath, forwarding the traffic to an intermediate node, or drop the traffic, as shown in Figure 13. The following will describe these decisions in more details.

Establishing a new lightpath

When the DPM compares the traffic load **(α)** (**α** - for a given lightpath) to **H** and if **α > H**, a decision will be taken based on the wavelengths availability, as shown in Figure 18.

A. If **α > H** and there are available resources (wavelengths) to the required destination, a new lightpath will be established. But, if there is more than one lightpath established to the same destination and there is one congested lightpath and another one is in the balanced state, the overloaded traffic over the congested lightpath will be forwarded over the non congested lightpath.
B. If **α > H** and there are no available resources to the required destination, one of two decisions could be taken:
    - A decision of forwarding the overloaded traffic will be taken. The overloaded traffic is forwarded to an intermediate AS (neighbor) then the intermediate AS forwards the traffic at the IP level to its final destination.
    - If there is no neighbor AS that accepts transit traffic to the requested destination for some reasons (e.g. the lightpath on a neighbor AS is in a congested state), some of the over-load traffic will be dropped.

Tearing down an existing lightpath

When the DPM compares the traffic load to **L** and if **α <L**, a decision will be taken based on the fallowing, as shown in Figure 18:

A. If there is more than one lightpath established to the same destination and if the traffic over the given lightpath is **α < L** (under-utilization state), then a decision of tearing down the under-utilized lightpath is taken.

B. If **α < L** and there is no residual traffic in the buffer (**α = 0**), the lightpath will be torndown.

C. If **α <L** and there is residual traffic in the established lightpath, one of two decisions could be taken:
   - Forward the residual traffic over an established lightpath if there is another lightpath to the same destination and this lightpath is in a balanced state.
   - The residual traffic is forwarded to an intermediate AS which forwards the traffic to its destination.

No action

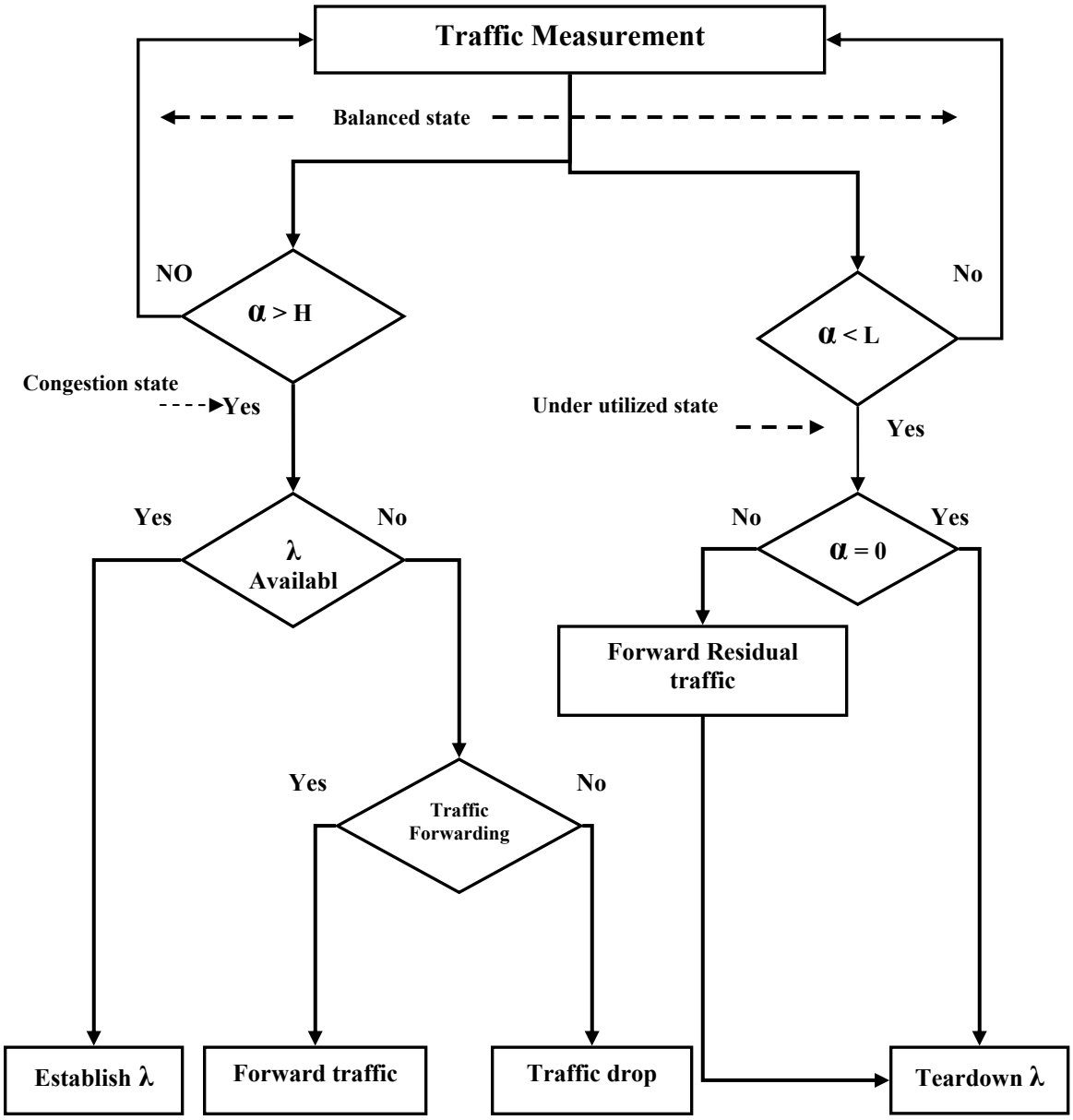
If the traffic load over the established lightpath is in the balanced region where **L < α < H**, then no decision will be taken. The DPM will wait until the next observation period (next measurement result).

Analysis

When the traffic load reaches **α > H** we establish a new lightpath. This can cause some delay. To reduce this delay, the algorithm can be modified as the following:

We can add a new parameter; we call it forwarding threshold (**F**). We start traffic load forwarding at the value **H > α > F.**

- If the domain can not forward the traffic load to an intermediate for some reasons (e.g. the intermediate domain is in a congested state) , then the algorithm must wait till **α>H** to establish a new lightpath
- If the domain own the network resources, then the value of H can be adjusted to avoid delay and the decision of establishing a new lightpath can be made when **α >F** or **α>H**



H- High threshold **λ** – Lambda (wavelength)
L- Low threshold **α** – Traffic load

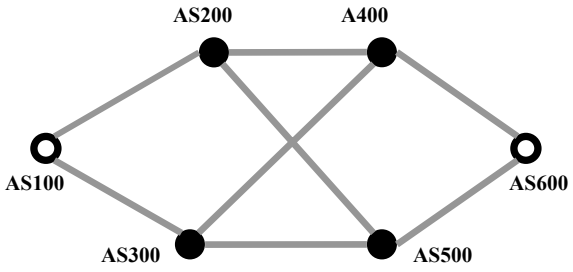
Figure 18: DPM algorithm.

6.4 An Example Network

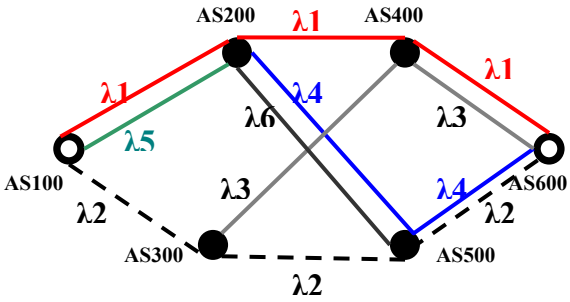
To clarify the TE-DPM algorithm, we consider in the following a network example and consider different scenarios. Each scenario describes some decisions taken by the DPM. The physical network consists of six nodes. Each node has an IP router and OXC. The network topology is shown in Figure 19.

- The physical topology of the network is shown in Figure 19-a. In the physical topology, each node represents a router and an OXC, and the optical fibers forms the physical links.
- The wavelengths connectivity is shown in Figure 19-b.
- The network virtual connectivity is shown in Figure 19-c. The lightpath connectivity between node pairs forms the virtual topology, where the node is represented by a router and the link is represented by the established lightpath between node pairs.

These figures will describe some scenarios for an ongoing traffic between AS100 and AS600. In our example we focus on one network node and we apply our algorithm over this node. AS100 will be the source node.

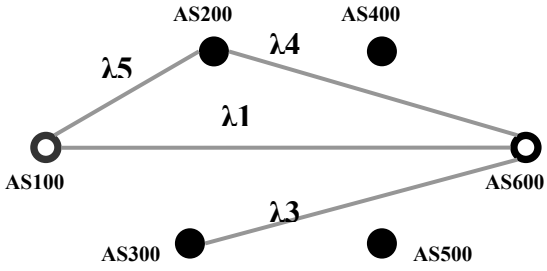


a) Network physical topology.



AS100 to AS600 λ1
AS100 to AS600 λ2 (available but not connected)
AS100 to AS200 λ5
AS200 to AS600 λ4
AS300 to AS600 λ3
AS200 to AS500 λ6

b) Wavelength connectivity.



c) Virtual topology for AS100 and AS600.

Figure 19: Network example.

Three scenarios will be explained, each one describes one or two decisions that will be taken by the DPM in the AS100 node based on the traffic load observation in AS100. The *First scenario* explains the ongoing traffic characteristic from AS100 to AS600. The *Second scenario* explains a case where traffic appears from AS100 to AS500 and there is no lightpath between these two ASs. *The Third scenario* explains a case where traffic load reaches **α<L** over an established lightpath from AS100 to AS600.

6.4.1 First scenario (traffic load larger than H)

In this scenario we have the following assumptions:

- There is an ongoing traffic from AS100 to AS600 over one wavelength (λ1).
- This wavelength starts at AS100 then traverses AS200 and AS400 till reaching AS600, as shown in Figure 19-b.When the connection is established between AS100 and AS600, it is called a lightpath. This lightpath represents the virtual end-to-end connection between these two ASs, as shown in Figure 19-c.
- Traffic load (α) over λ1 reaches the assigned **H** where **(α > H)**, as shown in Figure 20-a.
- Based on resources availability (wavelengths) in the network, the DPM can take one of the following decisions:

A) Traffic forwarding

If the traffic load is **α > H and** there is no available resources to establish a new lightpath. The traffic load could be sent to an intermediate node (AS) that sends this traffic to its final destination, as follows:

1) There is an established lightpath to neighbor AS200 through **λ5** , and the traffic load for **λ5** is in a balanced region (**L < α < H**), as shown in Figure 20-a.
2) The overload traffic is forwarded to AS200.
3) AS200 then forwards the traffic to AS600 (there is a lightpath established between AS200 and AS600 over **λ4**).

Analysis

As we deal with different ASs that have different administrations, and BGP protocol does not advertise bandwidth information. The question is, what *happens when the traffic load over* ***λ4*** *is* ***α>H*** *and how could AS100 know is?*

In our example, we assume that the ASs exchange the information about the available bandwidth using a routing protocol (e.g. BGP). In [XLWN02], an approach proposes an extension of BGP to advertise bandwidth availability information. It introduces a new QoS metric called Available Bandwidth Index (ABI). The ABI is used to perform bandwidth advertising and routing. If it is impossible to forward the overload traffic to an intermediate AS for some reasons (e.g. intermediate AS is in a congested state), then a decision of establishing a new lightpath will be considered by the DPM. Another possibility is just to forward the traffic to any intermediate AS without knowing the traffic load at that intermediate AS which will then be responsible for forwarding the traffic.

B) Establishing lightpath

Establishing a new lightpath to the same destination becomes a critical issue for congestion avoidance over an established lightpath to that destination. If the traffic load reaches the value **α > H** and if there are available resources in the network, a new lightpath must be established. Then the overloaded traffic is sent over this lightpath [SPAR01], as follows:

1. The wavelength **λ2** available to the same destination.
2. DPM makes a decision of establishing a lightpath to AS600.
3. Part of the traffic load from **λ1** is removed to **λ2,** as shown in Figure 20-b.

Analysis

If there is an available wavelength for an AS to the required destination, there will be no problem of establishing a new lightpath. But if the AS has no wavelength available to the

required destination, and the intermediate AS can not accept transit traffic for some reasons (e.g. not enough bandwidth to the same destination) [XLWN02], the only decision that is left to DPM is a traffic dropping.

C) Traffic dropping

Traffic dropping is the last and final decision that DMP could take if the two previously described decisions can not be taken (traffic forwarding or establishing a new lightpath). In this case, traffic will be dropped when the maximum buffer size, is reached for the given destination.

From the QoS perspective, the packets may be given different treatment before being dropped. If we apply DiffServ, as described in Section 3.3, we can define a dropping policy based on the SLA between the ASs (e.g. packets with low service level will be dropped and packets with high service level will be transmitted).

6.4.2 A second scenario (traffic to a new destination)

In the first scenario we assumed that there is an already an established lightpath and there is ongoing traffic from source to destination. In this scenario, we make a different assumption as follows: New traffic appears to a destination to which there is no lightpath established to that destination. In our network example, as shown in Figure 20-c, we can have the following question. W*hat decision must be taken when there is traffic from **AS100** to the destination **AS500** and there is no lightpath established between these two ASs?* If we look at Figure 20-c, we can see that there is no lightpath established between these two ASs. This traffic must be sent to its destination. Based on resource availability and the traffic load, DPM can take one of the following decisions.

A) Traffic forwarding

From the resource reservation point of view, if the incoming traffic load **$\alpha < L$,** there is no need to establish a new lightpath. This traffic could be forwarded to an intermediate AS that will forward this traffic to its final destination, as follows:

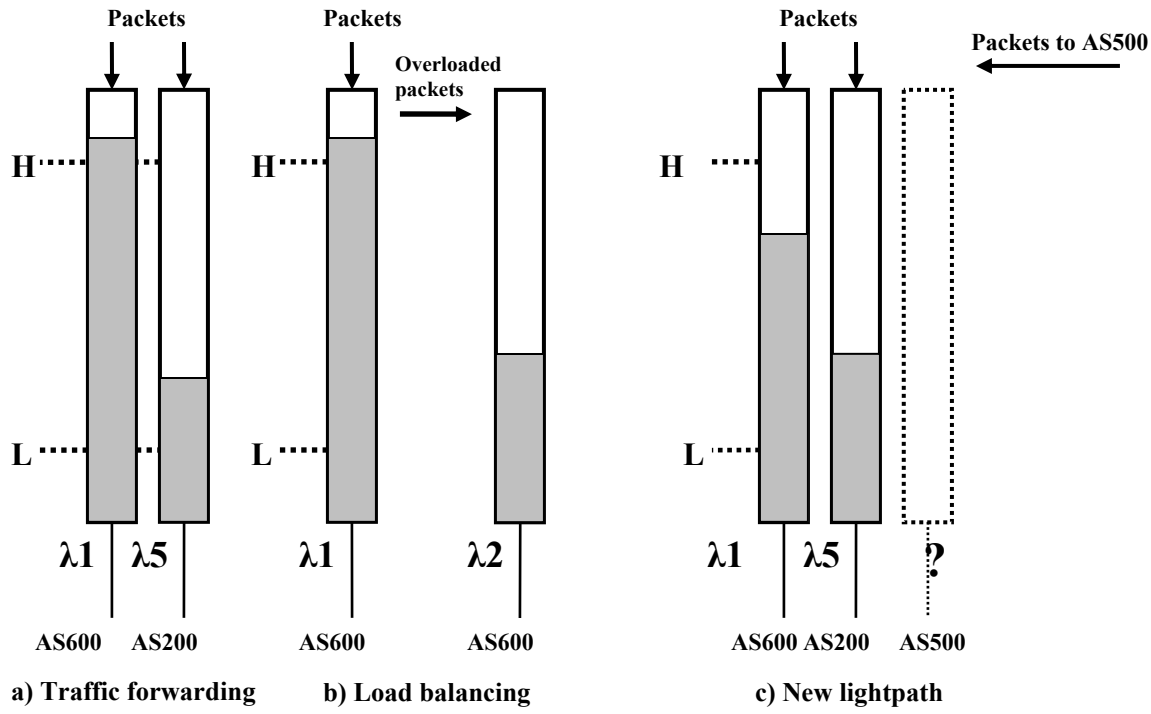


Figure 20: Buffers for AS100.

1) The DMP verifies that the traffic load is **α<L.**
2) There is a lightpath to the neighbor AS200 through **λ5** and the traffic is in a balanced region (**L<α<H**), as shown in Figure 20-c.
3) The decision to forward this traffic is taken.
4) The traffic is forwarded to AS200.
5) AS200 forwards the traffic to AS600.

Analysis

What happens if AS200 can not accept transit traffic for some reasons (e.g. no enough bandwidth to accept transit traffic to the requested destination) [XLWN02]. Establishing a new lightpath becomes an alternative for the DPM.

B) Establishing a new lightpath

If there is traffic to AS500 with **α < L** and an intermediate AS can not accept transit traffic, the following happens:

1) DPM verifies that **α <L**.
2) There is **λ2** available (not reserved).
3) DPM takes the decision of establishing a lightpath to AS500.
4) Traffic is sent to its destination.

C) Traffic dropping

Traffic dropping is the last decision that the DPM could take. The following takes place:

1) There are no available wavelength to establish a lightpath to AS500,
2) If it is impossible to forward the overloaded traffic to an intermediate AS, the traffic is dropped.

Chapter 7

Simulations Based performance

In order to study the effect of applying TE-DPM on network performance, some simulation studies were performed for different network topologies (NSFNET, simple networks, and a star network). The objective of these simulations was to evaluate the network performance in terms of IP packet loss rate and delay, resource utilization, and blocking probability for the establishment of new wavelength connections. The networks have the following properties:

- Each node represents an AS.
- There is at most one fiber between two nodes.
- There is no wavelength conversion in the network.

The simulation software we used was developed by our research group using Java programming language. The simulation works as the following:
For the simulation we run the network under different load. The x-coordinate for all figures (traffic load) represents the number of packets generated per milliseconds for each node.
The traffic (IP traffic with packets of fixed size of 64 k) are uniformly distributed over all the network nodes. The packet arrival follows a Poisson distribution and the arrival rate is increased and the results are collected. Each simulation runs for a certain period of time (400 ms).

7.1 Performance Metrics

The performance metrics we used are as follows:

- *Number of packets dropped:* This metric represents the number of packets that can not be served.
- *Average number of lightpaths established*: This metric shows the number of lightpaths established in the network.
- *Average packet delay* (described in Section 3.1.2).

- *Blocking probability*: This is the probability that blocking occurs when a source node tries to establish a new lightpath to a destination; it could not achieve that because there is no wavelength is available.
- *Oscillation factor*: see Section 6.2.2.

7.2 Simulation Results

7.2.1 NSFNET Network

The National Science Foundation Network (NSFNET) consists of 14 nodes, 21 unidirectional links and 8 wavelengths per link, as shown in Figure 21. We compare the simulation results for TE-DPM with the following other TE techniques;

- *Standard-TE* (STE). This is a simple algorithm. It is not sensitive to the network traffic load. Whenever there is traffic to a destination, a lightpath will be established. If there is no packet in the queue to a destination, then the lightpath will be torn down.

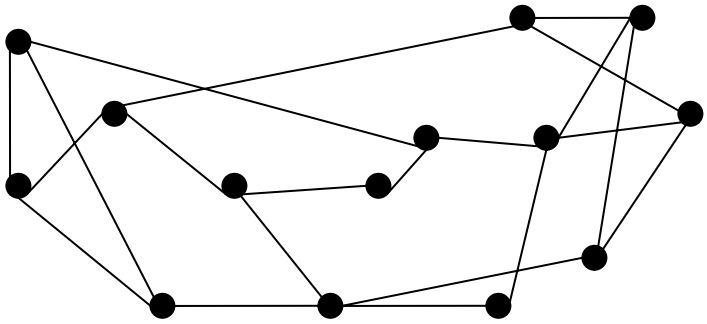


Figure 21: NSFNET network

- *The Optimal-TE* (opTE) [CHSH95] algorithm assumes that the traffic demands are known for all nodes, and then the source-destination pairs are sorted by traffic load, in a way that the priority of establishing a lightpath will be given to source-destination pairs that have more traffic.

7.2.1.1 No Traffic Forwarding

No traffic forwarding means that we do not forward the traffic load to an intermediate domain. If the traffic load is $\alpha > H$, a new lightpath will be established.

Discussion

Figure 22 shows the number of packets lost in the network. As the traffic load increases, the packets loss increases. In the **TE-DPM** network, the packets loss is higher than with **opTE** until a certain level where the number of lost packets becomes equal for all scheme which is normal, because all network resources become exhausted (the traffic load is very high) and there are no more lightpaths available to serve the network. For **STE**, the packets loss is the highest, because the network resource are used with no control.

Figure 24 shows the blocking rate of the network. For **opTE** the blocking rate is lowest. For **STE** the blocking rate is the highest because there is no control over the resource utilization. As traffic appears from source to destination, a lightpath will be established that leads to increase the blocking rate.
For **TE-DPM**, the blocking rate stays equal to 0 up to a certain traffic level. The TE-DPM algorithm controls the resource utilization depending on the traffic load. If the load increases, new light paths must be established. For this reason we see the blocking rate increases as the traffic load increases. As the traffic load becomes very high, all schemes become equal in term of blocking. (The network resources become exhausted).

Figure 25 shows the average delay in the network. **TE-DPM** starts with a high delay then the delay becomes stable. The reason is the following: The new lightpath is not immediately established and we can not forward the traffic to an intermediate system. The packets are kept in the buffer until they reach the L value, then the lightpath will be established. That causes a delay. To avoid this delay, we establish a new lightpath as soon as traffic arrivers. **opTE** is the optimal one because the buffer with the highest traffic load will be served first; this reduces the delay. For **STE** the delay is the worst. It starts with no delay

when the traffic is low, because STE serves the packets as they appear in the buffer, then the delay increases as the traffic load increases.

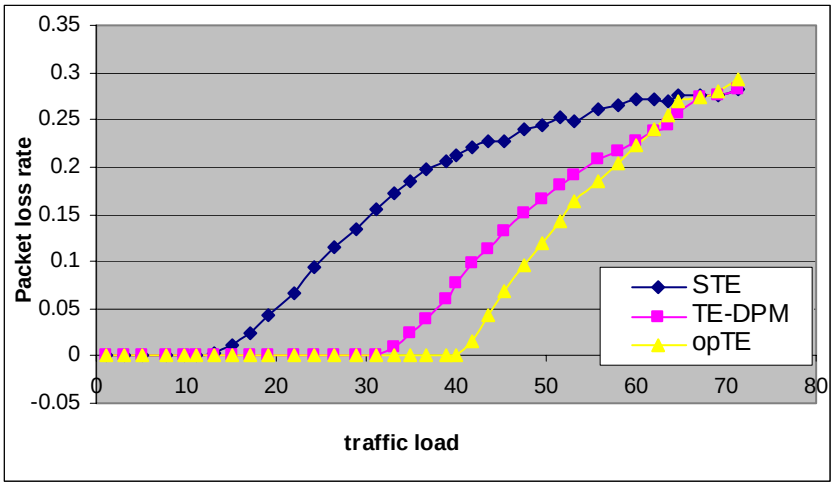


Figure 22: Fraction of packets lost with different TE techniques and no traffic forwarding (NSFNET network).

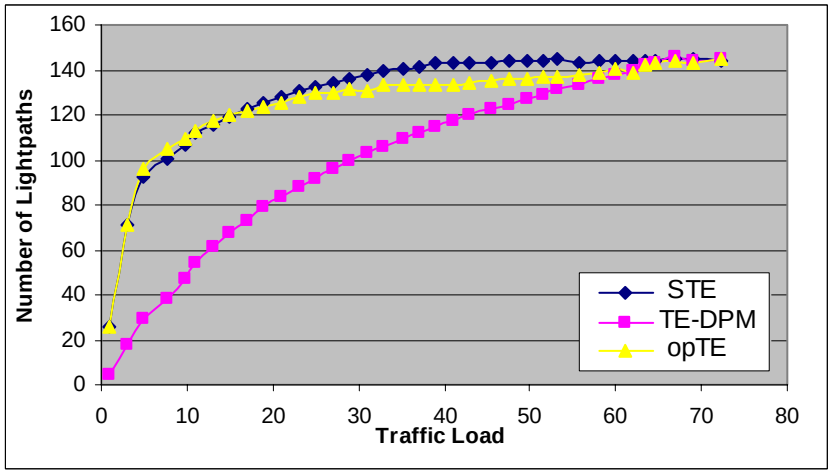


Figure 23: Average lightpaths established with different TE techniques and no traffic forwarding (NSFNET network).

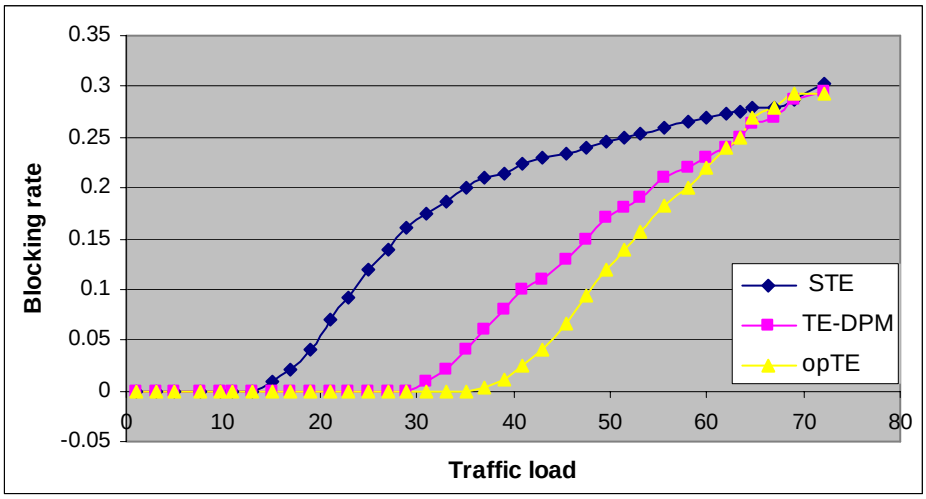


Figure 24: Blocking rate with different TE techniques and not traffic forwarding (NSFNET network).

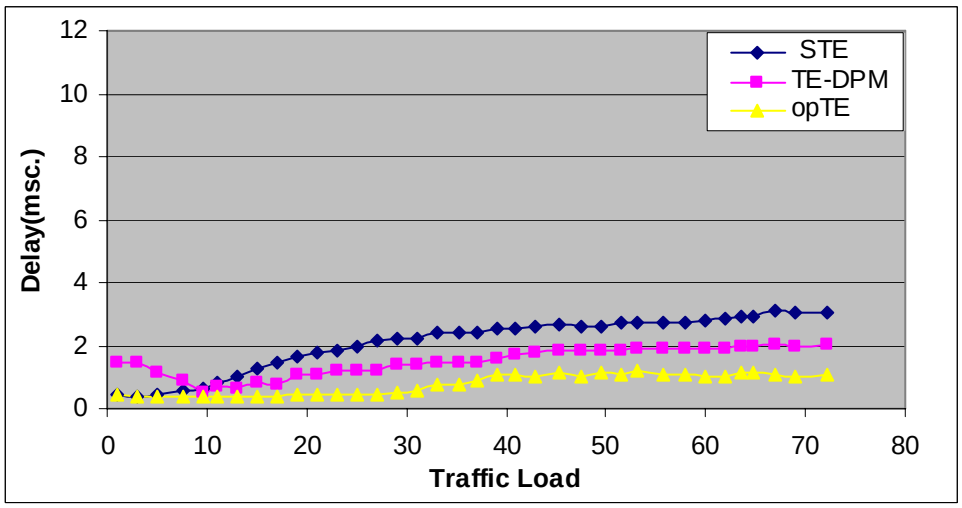


Figure 25: Average delay with different TE techniques and no traffic forwarding (NSFNET network).

7.2.1.2 With Traffic Forwarding

Some ASs do not accept transit traffic while others do. The decision of accepting transit traffic by an AS is based on business agreements with other ASs. In this section we assume that traffic can be forwarded to an intermediate domain, and we assume that each node knows the traffic load status (congested or balanced) of the intermediate domain. If the

intermediate domain is in a congested state, traffic is not forwarded to that domain. The objective of the simulation in this section is to compare three cases:

1. There is no traffic forwarding to an intermediate domain (AS).
2. There is traffic forwarding to an intermediate AS with the assumption of knowing the traffic load of the intermediate AS (in the congested or in the balanced state). The TE-DPM chooses the intermediate AS which has a lightpath to a final destination and the lightpath is in a balanced state.
3. There is traffic forwarding; the traffic will be forwarded to a neighbor AS without any knowledge about its traffic load status.

Discussion

Figure 26 shows the number of lost packets in the network. For TE-DPM the loss rate is improved compared to the case of no traffic forwarding. Forwarding blindly increases the packets loss rate. Here the source node forwards the overloaded traffic to an intermediate AS, without knowing the traffic load status of the intermediate AS. The intermediate AS might be in a congested state. Receiving extra traffic from its neighbor will cause an increase in traffic loss. The traffic loss becomes equal in all cases when the traffic load is very high (2120180 packets).

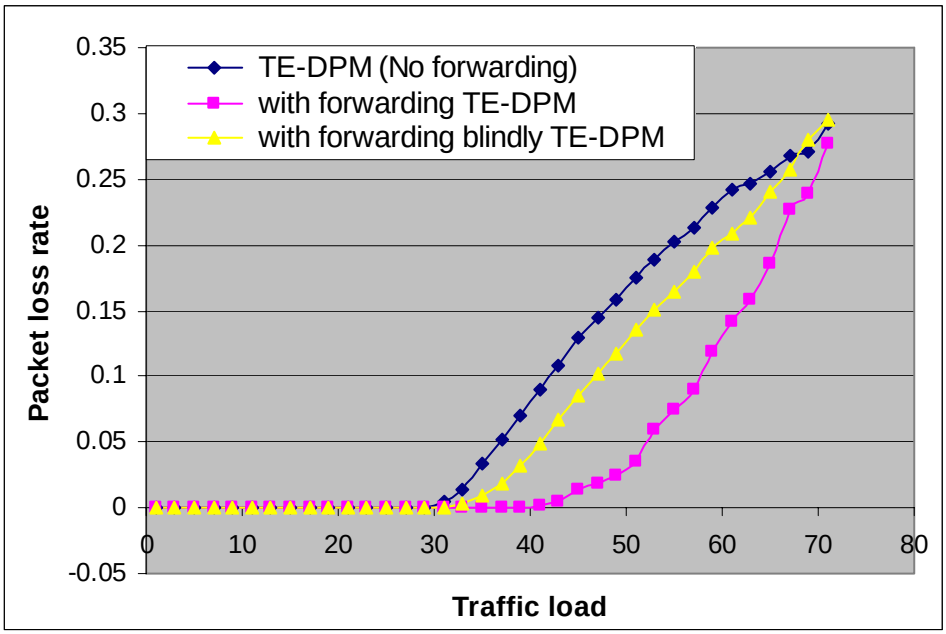


Figure 26: Packet loss rate (NSFNET network).

Figure 27 shows the average number of lightpaths established in the network. With TE-DPM, where traffic forwarding is considered, more lightpaths are established. The intermediate AS will have more traffic to serve (its own traffic and the transit traffic) which results in more lightpaths established. As the traffic load increases more lightpaths are needed to serve the traffic.

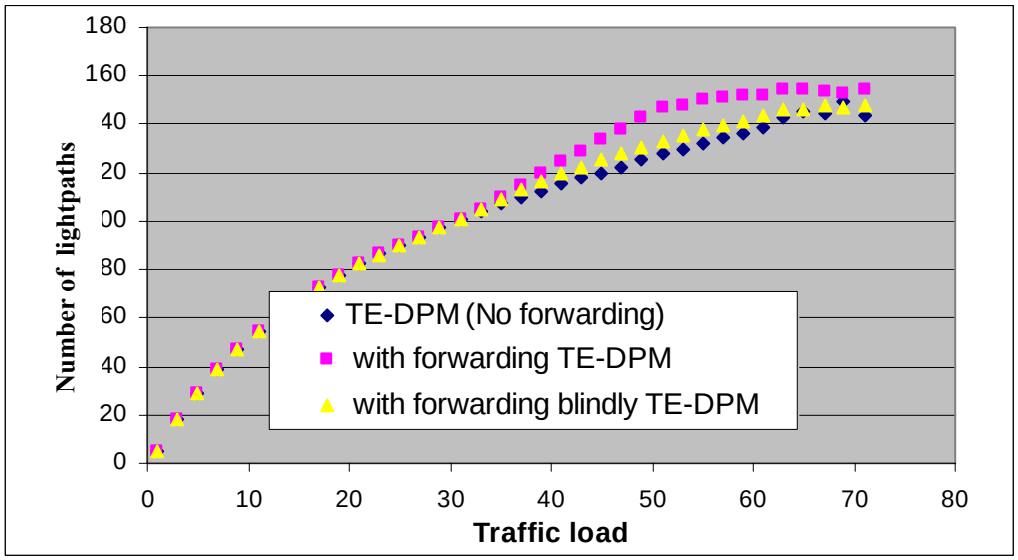


Figure 27: Average number of lightpaths established (NSFNET network).

Figure 28 shows the blocking rate when an attempt is made to establish a new lightpath. The best result is achieved when traffic is forwarded with knowledge about the traffic load status for the intermediate domain. With forwarding blindly, blocking will increase, because the intermediate AS might be in a congested state (while others are not) and it will get more traffic to serve. This will result in the requirement of establishing more lightpaths to serve the traffic load (this leads to the increase of the blocking probability). While other neighbors could have less traffic load and can accept more traffic without establishing an additional lightpath.

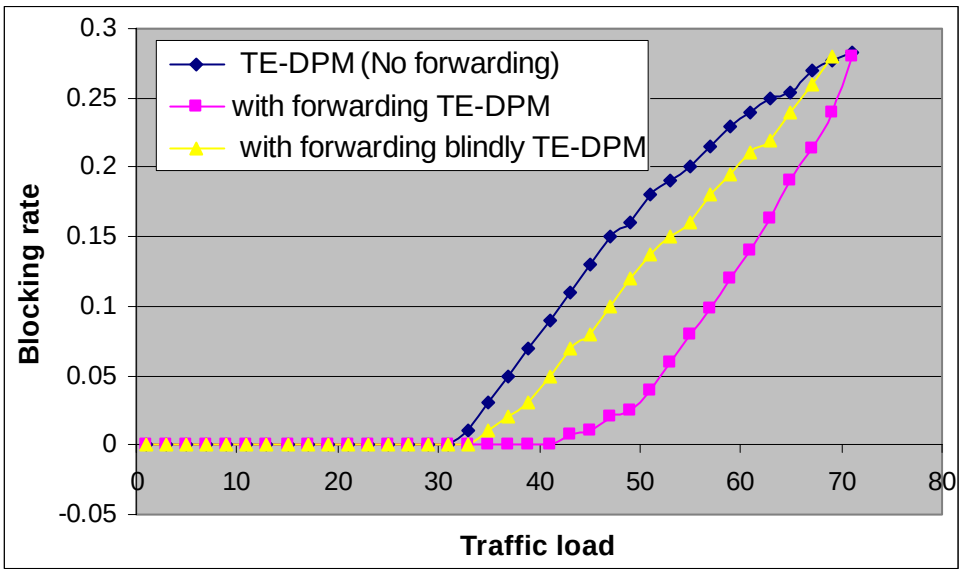


Figure 28: Blocking rate (NSFNET network).

Figure 29 shows the average delay, all cases are similar because the traffic is always kept in the buffer before making a decision for traffic forwarding or not.

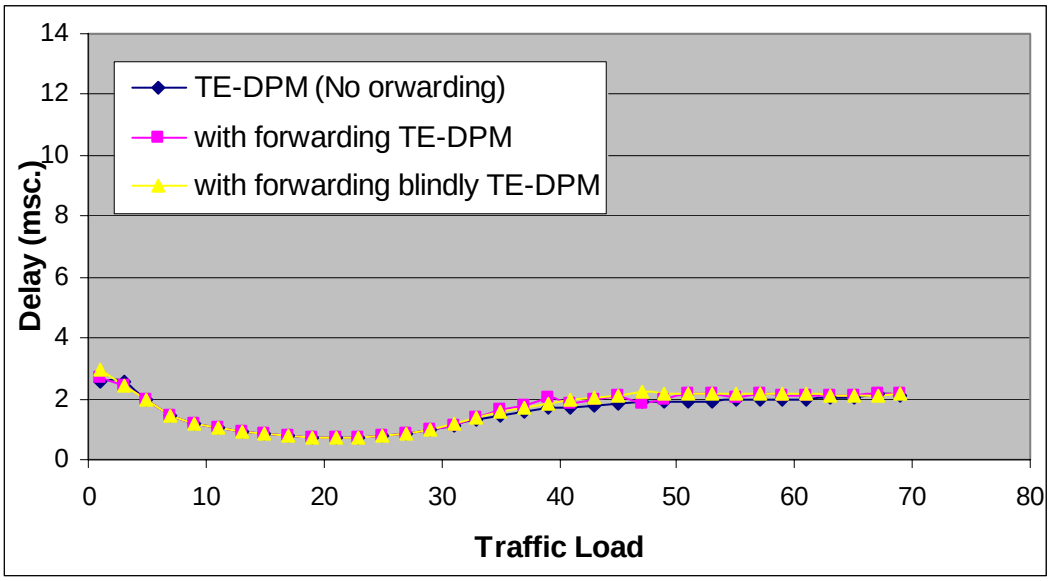


Figure 29: Average delay (NSFNET network).

7.2.2 Simple Network

In order to study the effect of applying TE-DPM on different network topologies, we apply the simulation to different network topologies:

- The first network consists of 7 nodes, 11 unidirectional links and 8 wavelengths per link, as shown in Figure 32-a. The bandwidth is 2.5Gb/s. Figure 31, 32, 33, 34 show the simulation results for packet loss, average number of lightpaths established, blocking probability and delay.
- The second network consists of 5 nodes, 6 unidirectional links and 6 wavelengths per link, as shown in Figure 32-b. The bandwidth is 2.5 Gb/s. Figure 35, 36, 37, 38 show the simulation results for packets loss, average number of lightpaths established, blocking probability -and delay.
- The third network (Star network) [BLLB02], consists of 6 nodes, (the central node represents a OXC and the edge nodes essentially consist of a router), 6 unidirectional links and 7 wavelengths per link. The bandwidth is 2.5 Gb/s.

Figure 36, 37, 38, 39 show the simulation results for packets loss, average number of lightpaths established, blocking probability and delay.

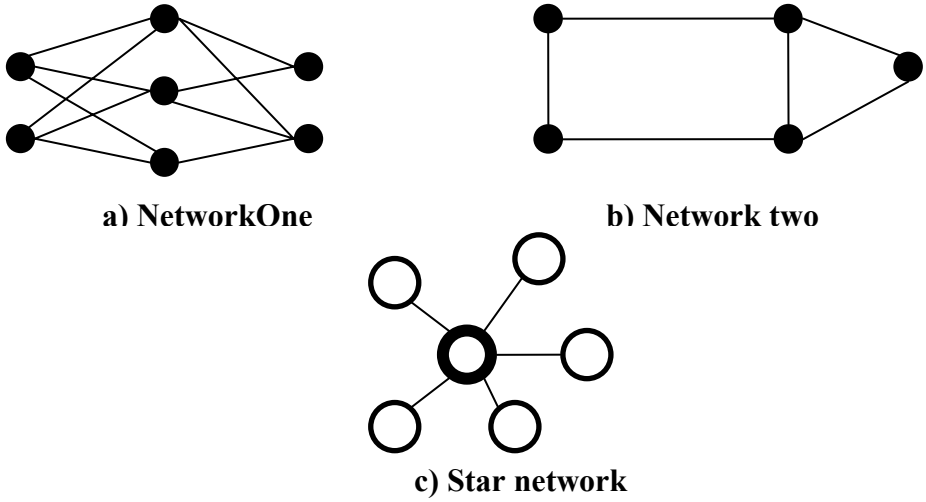


Figure 30: Simple Network.

Discussion

If we compare the results between the three networks presented in this section, we can notice the following.

Figures 31, 32 and 33 show the fraction of lost packets in the networks. All networks show similar results in pattern. At the beginning there are zero lost packets till a certain traffic load. As the traffic load increases the lost packets increase. When the traffic load increases, more lightpaths will be established. If the source node can not establish lightpath to a required destination and it can not forward the traffic to an intermediate node, the packets will be dropped.

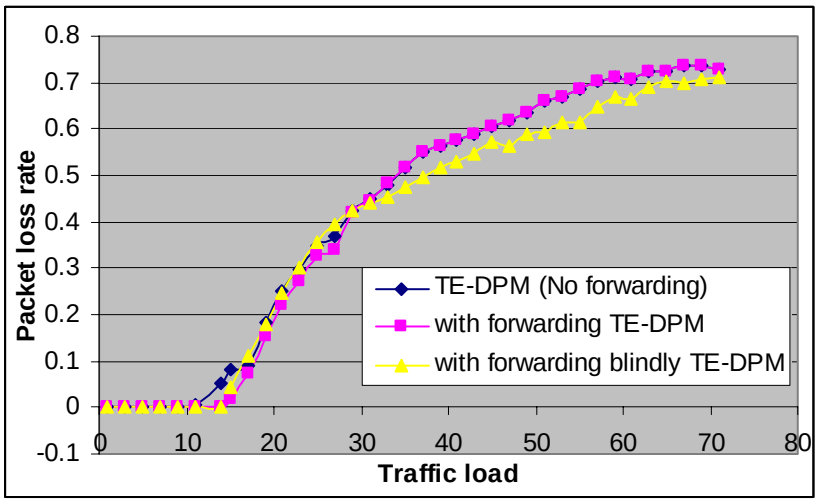


Figure 31: Packet loss rate (simple network one).

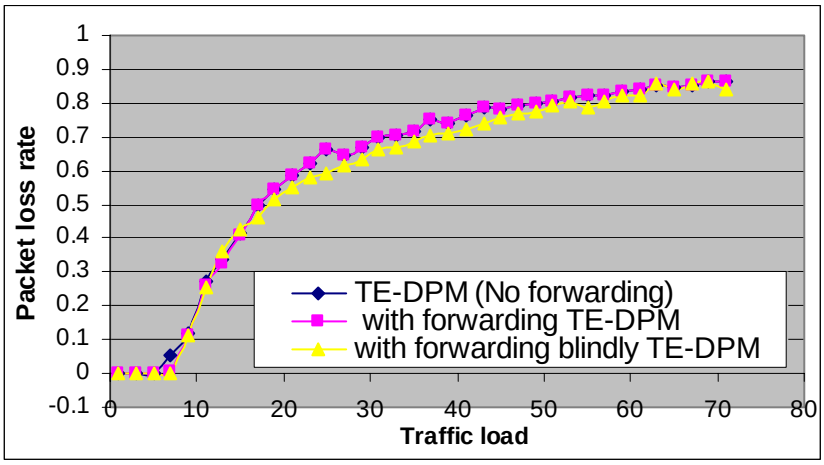


Figure 32: Packet loss rate (simple network two).

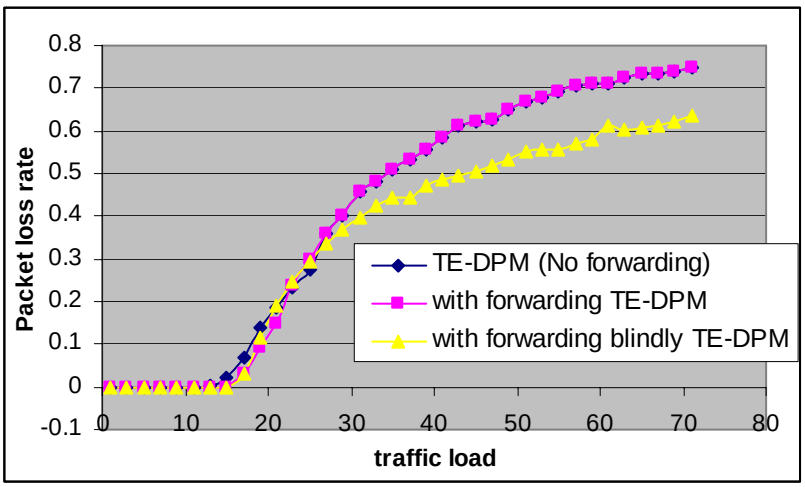


Figure 33: Packet loss rate (s

tar network).

Figures 34, 35, and 36 show the average number of lightpaths established by the TE-DPM. In the first network where traffic forwarding is considered, more lightpaths are established compared to the second network. But in general, the average number of lightpaths established, for the three networks and with different forwarding considerations, is similar in pattern. As the traffic increases, more lightpaths will be established to serve the traffic

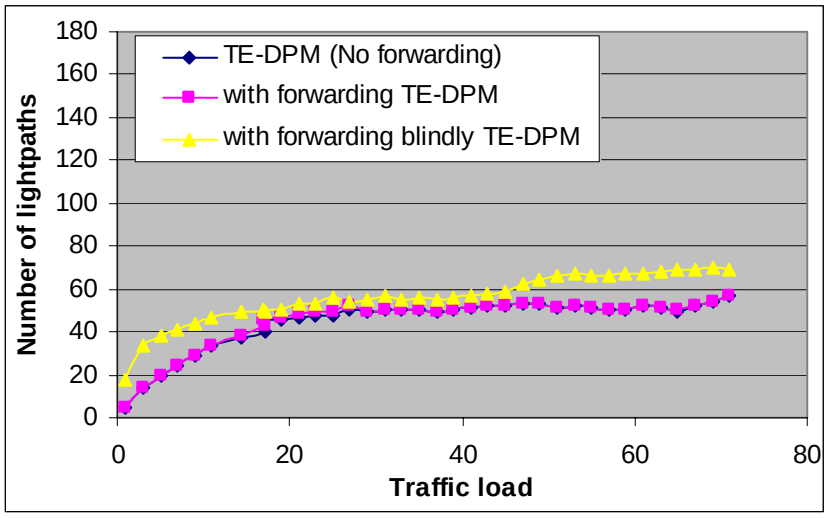


Figure 34: Average number of lightpaths established (simple network one).

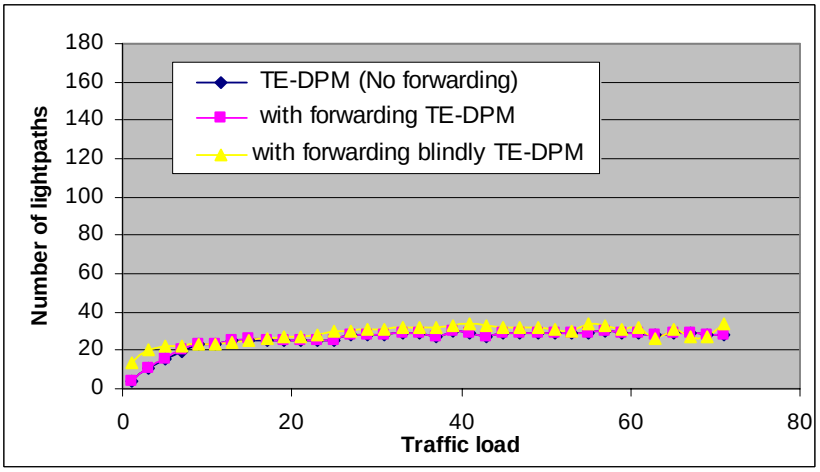


Figure 35: Average number of lightpaths established (simple network two).

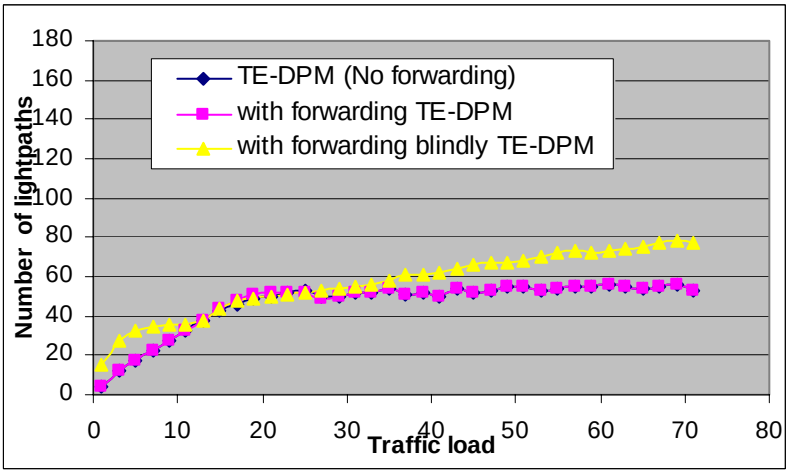


Figure 36: Average number of lightpaths established (star network).

Figures 37, 38, and 39 show the blocking rate. This result shows the similarity in pattern between the three networks in term of blocking probability. All networks start with zero delay till a certain traffic load where the blocking increases as the traffic load increases. This occurs because as the traffic load increases, more lightpaths will be established to service the traffic which results in increasing the blocking probability.

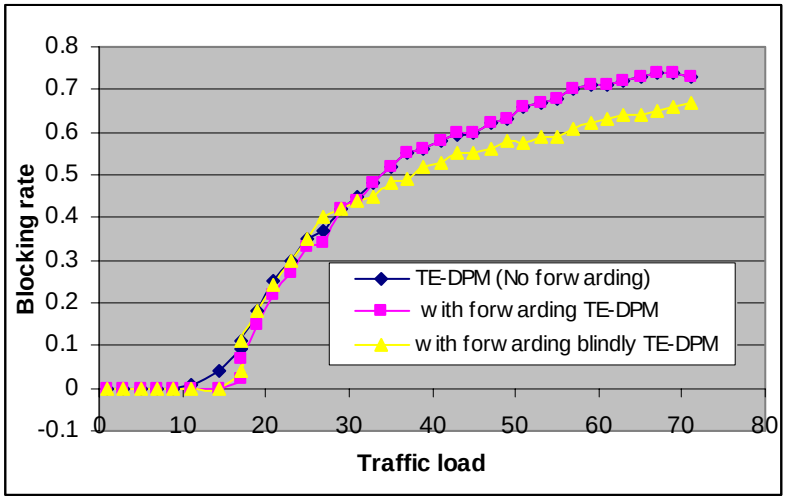


Figure 37: Blocking rate (simple network one).

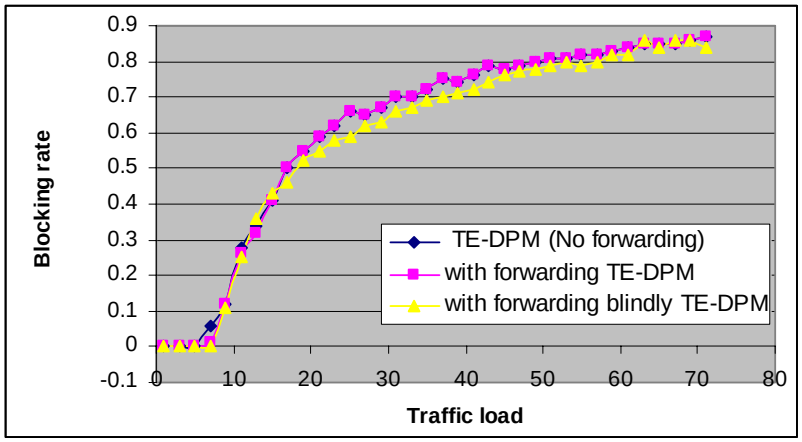


Figure 38: Blocking rate (simple network two).

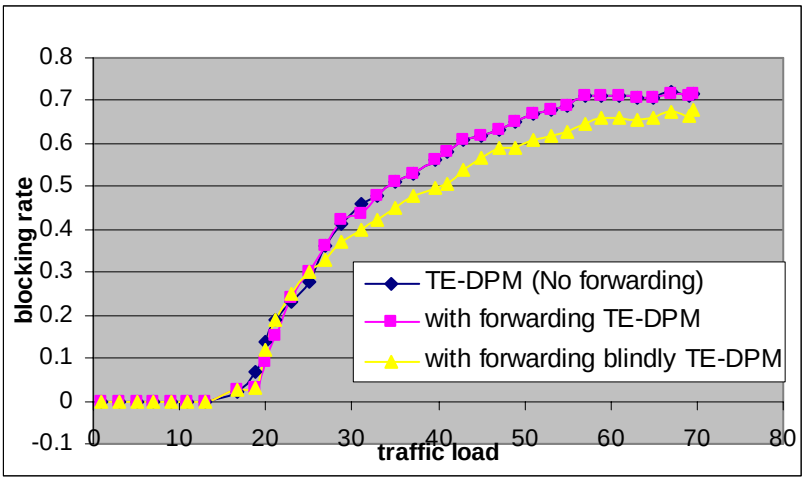


Figure 39: Blocking rate (star network).

Figure 40, 41, and 42 show the average delay; all networks have approximately similar delay patterns. In the beginning, there is a start delay, to avoid this delay we establish a new lightpath as soon as traffic arrives. Then the delay reaches a certain level that does not change, this occurs because the traffic load is very high and the network resources are all used.

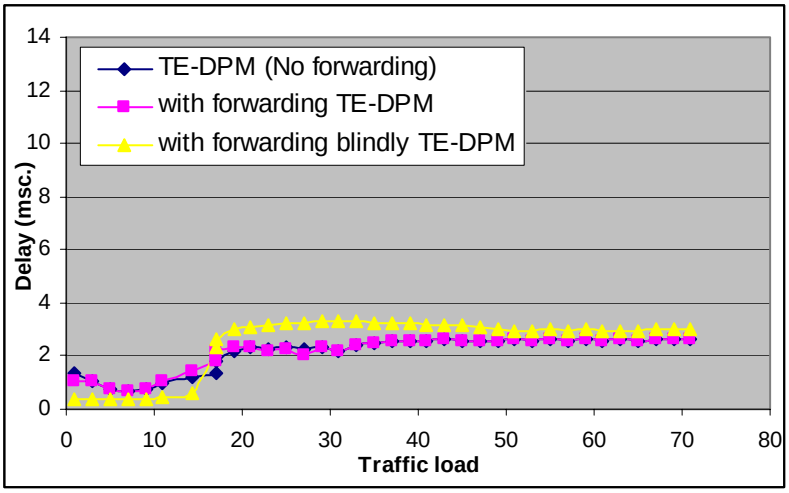


Figure 40: Average delay (simple network one).

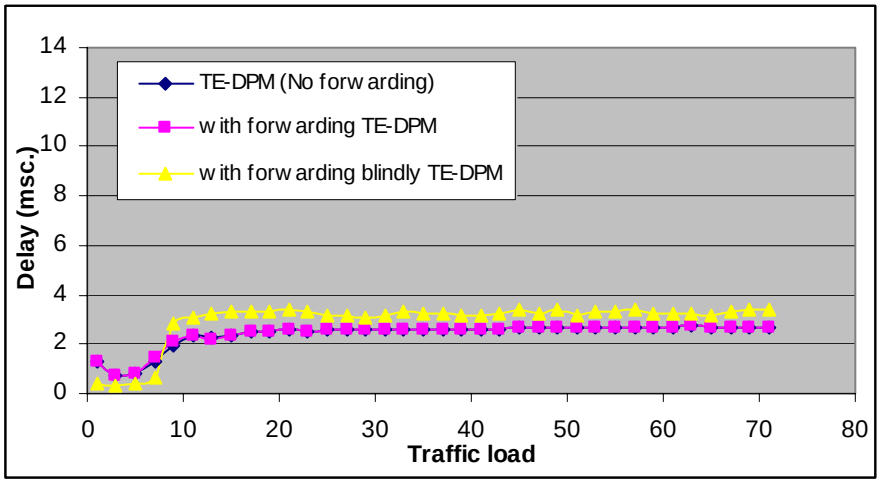


Figure 41: Average delay (Simple network two).

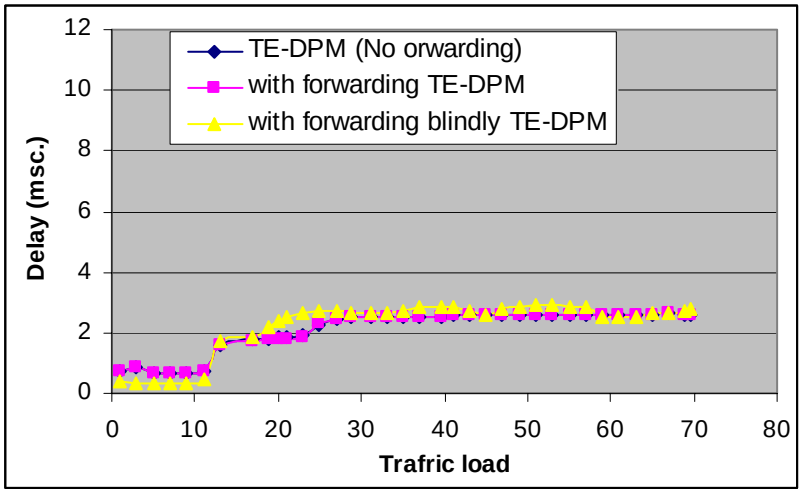


Figure 42: Average delay (Star network).

7.2.3 Simulations to study the effect of oscillations

As described in Section 6.3.2, the oscillation interval is the threshold added to the system parameters to make the process of establishing and tearing down lightpaths more stable, as shown in Figure 43. Let us give an example to explain the oscillation interval. Assume we have H = 16, L= 6 and an oscillation interval = 4 .For H we have: A = 16 + 2 =18 and B = 16 - 2 =14. For the L we have: C = 6 + 2 = 8 and D = 6 – 2 = 4

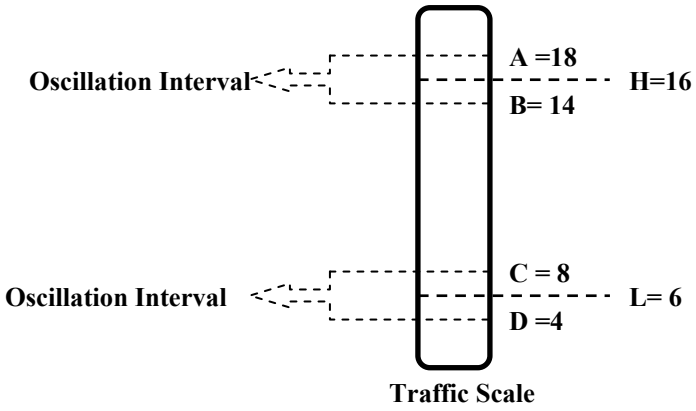


Figure 43: Oscillation interval example.

Figure 44 shows the effect of oscillation for two oscillation intervals with values 2 and 4, and without an oscillation interval. We call Oscillation (Y axis) the number of lightpaths established and torn down during a small observation period. We see that with a bigger oscillation interval we achieve better results (interval = 4). Without an oscillation interval, the oscillation in the network is very high, especially when the traffic load is low. Finally, we have to mention here that the oscillation interval does not affect any other performance parameters neither the TE-DPM algorithm.

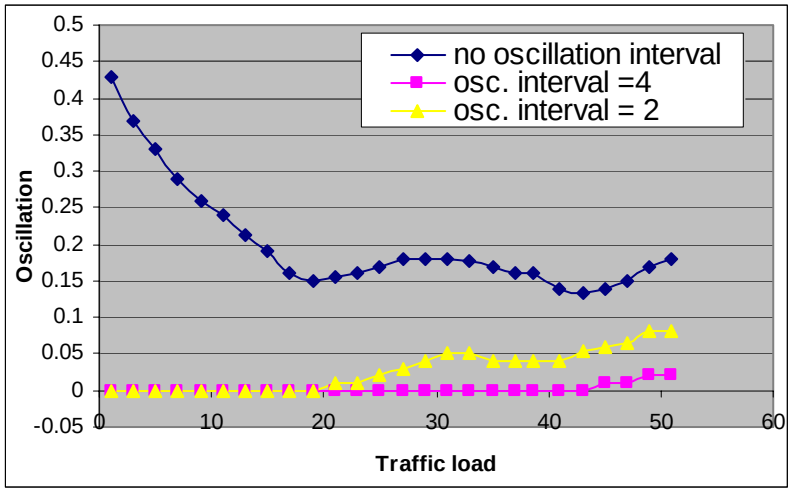


Figure 44: Oscillation factor.

Chapter 8

Conclusions of the Thesis

In this thesis we proposed and studied a TE algorithm for dynamic lightpath establishment between routers belonging to different ASs. We called this algorithm Traffic Engineering Decision Process Model (TE-DPM). As this algorithm is independent of any routing and signaling protocol, it could be applicable in any network architecture. From the simulation results that we obtained for different networks topologies, we conclude that packet loss and delay are improved by our algorithm. Instead of keeping the traffic lined up in the buffer or dropping the overload traffic when congestion occurs, we establish a new lightpath to serve the overload traffic. Even when there are no available resources (wavelengths) to establish new lightpaths, TE-DPM informs the IP router of traffic forwarding decision. Then the router forwards the overload traffic to an intermediate domain.

Resource utilization in optical network is an important issue. The significant achievement that our algorithm brought is the improvement in term of wavelength utilization, as shown in Figure 24. The DPM makes a decision of establishing a lightpath only if the traffic load reaches a certain threshold, otherwise the traffic is forwarded to an intermediate node which forwards the traffic to its final destination. In inter-domain networks, competition frequently occurs between different domains for network resources (wavelengths), where each network administrator tries to serve his customers without caring about resources utilization between domains. This competition decreases the efficiency of resource utilization and increases the blocking probability. Our algorithm solves the competition problem by giving an equal control of the network resources to all domains based on the traffic load that enters the inter-domain network.

Our TE algorithm is simple and easy to implement. Its simplicity comes from its simple design. There is no complex mathematical calculation needed to run the algorithm. And there are no control messages required to be exchanged between the network nodes to

run the algorithm. Each node running the algorithm works independently from other nodes. A traffic measurement device or software is required to dynamically measure the local traffic

8.1 Contributions of the Thesis

The contributions to this thesis are:

- We developed a TE algorithm for inter-domain optical networks called TE-DPM. This algorithm has the following advantages:
  - It is applicable over any network topology.
  - It performs dynamic lightpath establishment based on dynamic traffic measurements which lets TE-DPM adapt to dynamic changes.
  - The algorithm can be easily integrated into the existing Internet. It can run independently of any routing and signaling protocol.

- We performed simulation studies of the TE-DPM algorithm for different network topologies. The simulation evaluated the network performance in terms of IP packet loss rate and delay, resource utilization, and blocking probability for the establishment of new wavelength connections. The simulation results show similarity in pattern for different network topologies in terms of packet loss, average number of lightpaths established, blocking probability and delay.

8.2 Topics for Future Study

In our proposed TE algorithm, we did not take DiffServ into account. All traffic has the same service level. Moreover, as our work focuses on inter-domain traffic, the SLA may be different among domains based on their respective policies and business interests. Applying DiffServ in our TE scheme needs more investigation. The DiffServ for TE-DPM can be considered as a topic for future study.

The TE-DPM uses two parameters H and L to control the establishment of new lightpaths. H represents a high traffic load boundary. If the traffic load reaches this threshold, the system will attempt to establish a new lightpath. The L represents a low traffic mark. If the traffic load reaches this threshold, a lightpath will be torn down. If a domain owns the

inter-domain link, the network administrator will have more flexibility to assign the wavelengths based on the internal and external traffic load. Meanwhile, in the global inter-domain context, the network resources are shared between domains. Each domain may adjust the parameters H and L based on internal traffic, transit traffic and resource availability between domains. How to select the optimal values of H and L can be considered as another topic for future study.

References

[ACEW02] D. Awduche, A. Chiu, A. Elwalid, I. Widja, X. Xiao, "Overview and Principles of Internet Traffic Engineering," IETF, RFC 3272, May 2002.

[AHHC00] B. Arnaud, R. Hatem, W. Hong, J. Coulter, M. Blanchet, et all, "Optical BGP," Ca*Net3 News Archive, July 2000.

[AMAO99] D. Awduche, J. Malcolm, J. Agogbua, M. O'Dell, J. McManus, "Requirements for Traffic Engineering Over MPLS," IETF, RFC 2727, May 1999.

[ATJZ00] N. Agrawal, E.S. Tentarelli, N.A. Jacckman, L. Zhang, " Demonstration of Distributed mesh restoration and auto provisioning in a WDM network using a large optical cross connect," Proc. OFC200, PD37-1, pp.37.1-37.3, Baltimore, Maryland, March 2000.

[Awdu99] D. Awduche, "MPLS and Traffic Engineering in IP Networks," IEEE Communications Magazine, December 1999.

[AWDU01] D. Awduche et all, "Multiprotocol lambda switching: Combining MPLS traffic engineering control with optical cross connect," IEEE communications, pp. 111-116, March 2001.

[BAHU96] R. A. Barry, P. A. Humblet, "Modules of Blocking probability in All Optical Networks with and without Wavelength Changes," IEEE, Journal on Selected Area in Communication, vol., 14, no. 5, pp. 858-867, June 1996.

[BANE01] A. Banerjee et all, "Generalized multi-protocol label switching: An overview of routing and management enhancement," IEEE communications, pp. 144-150, January 2001.

[BBCD98] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services," IETF, RFC 2475, December 1998.

[BCDW98] S. Blake, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services," IETF, RFC. 2475, December 1998.

[BCKR98] T. Bates, R. Chandra, D. Katz, Y. Rekhter, "Multiprotocol Extensions for BGP-4," IETF, RFC 2283, February1998.

[BCSH94] R. Braden, D. Clark, S. Shenker, "Integrated Services in the Internet Architecture: an Overview," IETF, RFC 1633, June 1994.

[BEIJ02] I. V. Beijnum, "Building Reliable Networks with the Border Gateway Protocol," Book, 0-596-00254-8, September 2002.

[BECK00] T. Bates, Y. Ekhter, R. Chandra, D. Katz" Multiprotocol Extensions for BGP-4," IETF, RFC 2858, June 2000.

[BLAC02] U. Black, "Optical Networks Third Generation Transport Systems," Prentice Hall PTR, 2002, (ISBN 0130607266).

[BLLB02] F. J. Blouin, A. W. Lee, M. Beshai, "Comparison of two optical-core networks," Journal of optical networking vol.1, No.1, January 2002.

[BMRU99] N. Brownlee, C. Mills, G. Ruth," Traffic Flow Measurement: Architecture," IETF, RFC 2722, October 1999.

[BWW99] F. Baker, W. Weiss, J. Wroclawski," Assured Forwarding PHB Group," IETF, RFC 2597, June 1999. Section WD5, pp.56-58, Baltimore, Maryland, Mar. 2000.

[BZBH97] R. Braden, L. Zang, S. Berson, S. Herzong, S. Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification," IETF, RFC 2205, September 1997.

[CFZH96] L. Chlamtac, A. Farrago, T. Zhang, "Lightpath (wavelength) routing in large WDM networks," IEEE Journal on Selected Area in Communications, 5:909-913, 1996.

[CHGU02] H. J. Chao, X. Guo, "Quality of Service control in High-Speed Networks," A Wiley- Interscience Publication, 2002, (ISBN 0471003972).

[CHSH95] C.S. Chang, P. Heidelberger, P. Shahabuddin, "Fast simulation of packet loss rates in a shared buffer communications switch," ACM Press, p. 306-325, 1995.

[CHSC00] R. Chandra, J. Scudder, "Capabilities Advertisement with BGP-4," IETF, RFC 2842, May, 2000.

[COYA90] D. Comer, R. Yavatkar, "A Rate-Based Congestion Avoidance and Control Scheme Packet Switched Networks," Proc. Of 10th ICDCS, IEEE, pp. 390-397, 1990.

[CRJA02] G. Cristallo, C. Jacquenet, "Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI Attribute," IETF, internet draft, March 2002.

[DARB81] DARPA Internet Program, “Internet Protocol,” IETF, RFC 791, September 1981.

[DAVI97] J. Davidson, “An introduction to TCP/IP,” Springer – Verlage, 1997, (ISBN 038796651X).

[EBBB97] A. Ed, F. Backer, B. Braden, S. Bradner et all, “Resource Reservation Protocol (RSVP) Version 1 Applicability Statement Some Guidelines on Deployment,” IETF, RFC 2208, September 1997.

[EIMO00] J. H. Elmirghani, H. T. Mouftah, “Technologies and Architectures for scalable Dynamic Dens WDM Networks,” IEEE Communications Magazine, Vol.38, No.2, pp.58-66, February 2000.

[ERLI00] V. Eramo, M. Listani, “Packet loss in a bufferless optical WDM switch employing shared tunable wavelength converters,” IEEE/OSA Journal of Lightwave Technology, 18(12):1818-1833, December 2000.

[FEHU98] P. Ferguson, G. Huston, “Quality of Service: Delivering QoS on the Internet and in Corporate Networks,” Willey Computer Publishing, 1998, (ISBN 0471243582).

[FEPE01] N. Feamster, J. Pexford, “Network-Wide BGP Route Prediction for Traffic Engineering,” AT&T Labs-Research, NJ, USA, 2001.

[FLJA93] S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion. Avoidance,” IEEE/ACM Transactions on Networking, Vol. 1 Nov. 4, p. 387-413, August 1993.

[FPHL02] M. J. Francisco, L. Pezoulas, C. Huang, L. Lambadaris, “End-to-End Signaling and Routing for Optical IP Networks,” IEEE, 2002.

[G.872] ITU-T Rec. G872, “Optical Transport Networks,” February 1999.

[GJJM00] D. Goderis, T. Jones, Y. Jacquenet, C. Memenios, et all, “Specification of a Service Level Specification (SLS) Template,” IETF, internet draft, November 2000.

[GRHJ00] A. Greenberg, G. Hjalmtysson, J. Yates, ”Smart routers-simple optics a network architecture for IP over WDM,” Optical Fiber Communication Conference, 2000.

[GRLR00] GRAR00, “Quality of Service in IP Network,” April 2000, ISBN:1578701899

[HBWW9] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, “Assured Forwarding PHP Group,” IETF, RFC 2597, June 1999.

[HKWE01] Haward, H. Kim, M. Weiss, “Interprovider DiffServ Interconnection Management Modules in the Internet Bandwidth Commodity Market,” Telematics and Informatics, Special Issues on EC, 2001.

[HLHU02] C. F. Hsu, T. Lung, N. Huang, “Performance Analysis of Deflection Routing In Optical Burst-Switching Networks,” IEEE, 2002.

[HNCA99] D. K. Hunter, H. M. Nizam, M. C. Cia, I. Andonovic, et all, “A wavelength switched packet network,” IEEE communications, 37(3), March 1999.

[JACO89] V. Jacobson, “Presentations to the IETF Performance and Congestion Control Working Group,” August 1989.

[JACW02] B. Jamoussi, L. Andersson, R. Callon, L, Wu, et all, “Constraint-based LSP using LDP,” IETF, RFC 3212, January 2002.

[JNPO99] V. Jacobson, K. Nichols, K. Poduri, “Expedited Forwarding PHB,” IETF, RFC 2598, June 1999.

[KHIR04] M. Khair, “An Implementation Approach for an Inter-Domain Routing Protocol for DWDM,” Master thesis, SITE, University of Ottawa, March 2004.

[KORE03] K. Kompella, Y. Rekhter, “IS-IS Extensions in Support of Generalized MPLS,” Internet draft, <raft-ietf-isis-gmpls-extensions-16.txt>, work in progress, June 2003.

[KRBD02] K. Kompella, Y. Rekhter, A. Banerjee, and all, “OSPF Extensions in Support of Generalized MPLS,” Internet Draft, <draft-ietf-ccamp-ospf-gmpls-extensions-04.txt>, work in progress, February 2002.

[LYSH00] S. E. Lyshevski, “Control System Theory with Engineering Applications,” Birkhauser, November 2000, ISBN 081764203X).

[MANK90] A. Makin, “Random Drop Congestion Control Proc.,” ACM SIGCOMM, pp.1-7, September, 1990.

[MARA91] A. Makin, K. Ramakrishnan, “Gateway Congestion Control,” IETF, RFC-1252, August 1991.

[MCKE91] P. McKenny, “Stochastic Fairness Queuing,” Internetworking: Research and Experience, vol. 2 pp. 133-131, January 1991.

[MCSP98] D. E. McDysan, D. L.Spohn, “ATM Theory and Application” McGraw-Hill

Osborne Media, 1998, (ISBN 0070453462).

[NAGL 85] J. Nagle, "On Packet Switches With Infinite Storage," IETF, RFC 970, December 1985.

[NBBB98] K. Nichols, F. Baker, S. Blake, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers" IETF, RFC 2474 December 1998.

[OHCA01] I. T. Okumus, J. Hwang, H. A. Mantar, S. J. Chapin, "Inter-Domain LSP Setup Using Bandwidth Management Points," Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE, Volume: 1, 25-29, November 2001.

[POST81] J. Postel, "Transmission Control Protocol-DARPA Internet Program Protocol Specification," STD7, RFC793, DARPA, September 1981.

J. Post, "Internet Control Message Protocol," IETF, RFC-297, September 1981

[PRPO87] W. Prue, J. Posrel, "Something a Host Could Do with Source Quench," IETF, RFC 1016, July 1987.

[QBon00] QBone, Signaling Design Team, http://qbone.internet2.edu/meetings/houston2000/.

[QIYO99] C. Qiao, M. Yoo, "Optical Burst Switching (OBS) – a new paradigm for the optical Internet," IEEE Journal of High Speed Networks, 8(1), 1999.

[RAJA90] K. Ramakrishnan, R. Jain, "A Binary Feedback Scheme for Congestion Avoidance in Computer Networks," ACM Trans. On Computer Systems, vol. 8, no.2, , pp. 158-181,May 1990.

[RAJA87] K. Ramakrishnan and R. Jain, "Congestion Avoidance in Computer Networks with A Connectionless Network Layer: Part IV: A Selective Binary Feedback Scheme for General Topologies," DEC Technical Report DEC-TR-510, 43 pp. August 1987.

[RASI96] R. Ramswami, K. N. Sivarajan, "Design of Logical Topologies for Wavelength Routed Optical Networks," IEEE Journal on Selected Areas in Communications, vol. 14, pp. 840-851, January 1996.

[RELE95] Y. Rekhter, T. Li, "Border Gateway Protocol 4 (BGP-4)," IETF, RFC 1771, March 1995.

[RLAW03] B. Rajagopalan, J. Luciani, D. Awduche, "IP over Optical Networks: A

Framework," Internet Draft, Work in progress, 2003.

[RORE99] E. Rosen, Y. Rekhter, "BGP/MPLS VPNs," IETF, RFC 2547, March 1999.

[RVCA01] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture," IETF, RFC 3031, January 2001.

[SBBB01] P. A. Smith, A. Banerjee, L. Berger, G. Bernstien et all, "Generalized MPLS Signaling RSVP-TE Extensions," Internet Draft, <draft-ietf-mpls-generalized-rsvp-te-06.txt>, work in progress, November 2001.

[SEST01] C. Semeria, J. W. Stewart, "Supporting Differentiated Services Classes in Large IP Networks," Juniper networks, part number: 200019-001, 2001.

[SPAR01] Sparrow, Global load Balancing for Linux, http://www.supersparrow.org/, software, Initial Release January 2001.

[SWAL99] G. Swallow, "MPLS advantages for traffic engineering," IEEE Communications Magazine, Volume: 37, Issue: 12, page(s): 54 -57, December 1999.

[TURN99] J. S. Turner, "Terabit burst switching," International Journal High Speed, Networks 1999.

[WEDZ02] G. Wedzinga, "Photonic Slot Routing in Optical Transport Networks," Kluwer Academic Publishers, December 2002, 222 pp (ISBN 1402073488).

[WPSW98] J. Wei, M. Post, C. C. Shen, B. Wilson, et all, "Network control and Management of reconfigurable WDM All-Optical Network," IEEE/IFIP 1998 Network Operations and Management Symposium; 15.2.-20.02.1998.

[WROC97] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," IFTF, RFC 2210, September 1997.

[XBXU02] Y. Xu, A. Basu, Y. Xue, "A BGP/GMPLS Solution for Inter-Domain Optical Networking," IETF, Internet draft, June 2002.

[XLWN02] L. Xiam, K. S. Lui, J. Wang, K. Nahrstedt, "QoS extension to BGP," Network Protocols, 2002. Proceedings. 10th IEE International Conference, pages: 100-1009, 2002.

[XYDQ01] C. Xin, Y. Ye, S. Dixit, C. Qiao, "An emulation-based comparative study of centralized and distributed control schemes for optical networks," SPIE ITCom2001, Aug. 19-24, Denver; Proc. SPIE 4527: 65-72. 2001.

[YARE95] C. Yang, A. Reddy, "A Taxonomy for congestion Control Algorithms in Packet Switching Networks," IEEE Network Magazine, p. 34-45, 1995.

[YHMA99] Y. Yamada, S. Mino, K. Habara, "Ultra fast clock recovery for burst-mode optical packet communication," proc. OFC 99, pp.114-116, 1999.

[YOQI98] M. Yoo, C. Qiao, "A new optical burst switching protocol for supporting quality of service," Proc. SPIE98, All Optical Networks: Architecture, Control, and Management Issues, pp.396-405, Boston, Massachusetts, November 1998.

[YXMP02] S. Yao, F. Xue, B. Mukherjee, S. J. Ben Yoo, S. Dixit, "Electrical ingress buffering and traffic aggregation for optical packet switching and its effect on TCP-level performance in optical mesh networks," IEEE communications Magazine, 40(9):66-72, September 2002.

[ZJMU00] H. Zang, J. P. Jue, B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," Optical Networks, 1(1):47-60, January 2000.

Appendices

A- List of Acronyms and Abbreviations

ABI	Available Bandwidth Index
ADM	Add Drop Multiplexing
AF	Assured Forwarding
ARIN	American Registry for Internet Number
AS	Autonomous System
ATM	Asynchronous Transfer mode
BGP	Border Gateway Protocol
BMP	Bandwidth management Point
CIDR	Classless Inter-domain Routing
CR-LDP	Constraint-based Routing using Label Distribution Protocol
DSCP	Differentiated Service Code Point
DWFQ	Dynamic Weighted Fair Queuing
EF	Expedited Forwarding
ENNI	Exterior Network to Network Interface
FEC	Forward Equivalence Class
FIFO	first-in, first-out
GMPLS	Generalized MoltiProtocol Label Switching
IETF	Internet Engineers Task force
INNI	Interior Network to Network Interface
IP	Internet Protocol
ISA	Integrated Service Architecture
IS-IS	Intermediate System to Intermediate System
ISP	Internet Service Provider
LMP	Link Management Protocol
LSP	Label-switched Path
LSR	Label-switched Router
MED	Multi Exit Discriminator
MPLS	MoltiProtocol Label Switching
MPλS	MoltiProtocol Lambda Switching
NLRI	Network Layer Ratability Information
NC	Network Control
NC&M	Network Control & Management
NLRI	Network Layer Reachability Information
NNI	Network to Network Interface
ONT	Optical Network
OLS	Optical Label Switching
OBGP	Optical Border Gateway Protocol
OADM	Optical Add-Drop Multiplexing
OBS	Optical Burst Switching
OCh	Optical Channel

OMS	Optical Multiplex Section
opTE	Optimal Traffic Engineering
OSPF	Open Shortest Path First
OXC	Optical Cross Connect
OTS	Optical Transport Section
PHB	Per Hop Behavior
PTE	Proposed Traffic Engineering
QoS	Quality of Service
RAA	Resource Allocation Answer
RAM	Random-access memory
RAR	Resource Allocation Request
RED	Random Early Detection
RSVP-TE	Resource reservation protocol- Traffic Engineering
SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SONET	Synchronous Optical network
SRSO	Smart Router-Simple Optics
STE	Standard Traffic Engineering
STE	Standard Traffic engineering
TCA	Traffic Conditioning Agreement
TCP	Transmission Control Protocol
TE	Traffic Engineering
TE-DPM	Traffic Engineering -Decision Process Module
ToS	Type of Service
UDP	User Datagram Protocol
UNI	User Network Interface
WDM	Wave Division Multiplexing
WS	Wavelength signaling
WR	Wavelength routing

B- Terminology and Concepts

In this section we introduce terminology pertinent to this thesis and some related concepts.

- Asynchronous transfer mode (ATM): is a dedicated-connection switching technology that organizes digital data into 53-byte cell units and transmits them over a physical medium using digital signal technology. Individually, a cell is processed asynchronously relative to other related cells and is queued before being multiplexed over the transmission path.

- *Congestion:* A state of a network resource in which the traffic incident on the resource exceeds its output rate.

- *Dense Wavelength Division Multiplexing (DWDM):* A process by which different colors, hence different wavelengths, of an optical signal are multiplexed onto one strand of optical fiber. Due to the unique wavelength of each signal, simultaneous transmissions of different types of signals are possible.

- *Domain (Autonomous System):* is defined as a set of routers or networks that are technically administrated by a single organization such as corporate network, a campus network, or an Internet Service provider (ISP) network.

- *Fair queuing:* The controlling of congestion in gateways by restricting every host to an equal share of gateway bandwidth. Fair queuing does not distinguish between small and large hosts or between hosts with few active connections and those with many.

- *Explicit routing:* This scheme relies on a network made of switches, IP routers or ATM switches. A predefined path is specified in advance for a packet. This is a virtual circuit in the ATM world. Since the path is predefined, the packet is switched at each node, thus eliminating the need to make routing decisions at every node along the path. Explicit routing is useful for traffic engineering, QoS (Quality of Service), and the prevention of routing loops. It requires path setup in advance, something that can be done in IP networks with MPLS (Multiprotocol Label Switching).

- *Exterior Gateway protocol (EGP):* A protocol which distributes routing information to the routers which connects ASs. Border Gateway Protocol (BGP) is an EGP defined in RFC 1267 and RFC 1268. Its design is based on experience gained with EGP. BGP is the routing protocol used to exchange routing information across the Internet.

- *Interior Gateway Protocol (IGP):* An Internet protocol which distributes routing information for inter-domain routing to the routers within an autonomous system. The term "gateway" is historical; "router" is currently the preferred term.
- *IP router:* A router is a physical device that joins multiple networks together. Technically, a router is a Layer 3 gateway, meaning that it connects networks (as gateways do). Routers use the Internet Protocols (IP) to reach a desired destination.
- *Lightpath or Optical Channel:* A lightpath is a point-to-point optical layer connection between two access points in an optical network. It is also called an optical channel or channel.
- *Link Management Protocol (LMP):* Is defined as a component of GMPLS. It is a protocol supporting control channel management, link connectivity verification, link property correlation, and fault management between adjacent nodes. In the context of optical network, LMP can run between OXCs and, between an OXC and an optical line system (e.g. WDM multiplexer).
- *Optical Cross Connect (OXC):* An OXC is a true system-level optical switch that can switch an optical data stream from an input port to a output port. Such a switch may utilize optical-electrical conversion at the input port and electrical-optical conversion at the output port, or it may be all-optical. An OXC is assumed to have a control-plane processor that implements the signaling and routing protocols necessary for computing and instantiating optical channel connectivity in the optical domain.
- *SONET*: The synchronous optical network, known in Europe as synchronous digital hierarchy, or SDH. This is a 20-year-old standard .Traffic is carried in an electronic packet transported over fiber, using dual counter-rotating rings. The systems are relatively inexpensive to implement, but they were designed in the 1970s for primarily circuit-switched voice traffic, so they make very inefficient use of bandwidth in carrying traffic that is mostly IP-encapsulated data.
- *Transmission Control Protocol (TCP):* A primary part of the TCP/IP set of protocols which forms the basis of communications on the Internet. TCP is responsible for breaking large data into smaller chunks of data called packets. TCP assigns each packet a sequence number and then passes them on to be transmitted to their destination. Because of how the Internet is set up every packet may not take the same path to get to its destination. TCP has the

responsibility at the destination end of reassembling the packets in the correct sequence and performing error-checking to ensure that the complete data message arrived intact.

- *User Datagram Protocol (UDP):* is a communications protocol that offers a limited amount of service when messages are exchanged between computers in a network that uses the Internet Protocol (IP)). UDP is an alternative to the Transmission Control Protocol (TCP) and, together with IP, is sometimes referred to as UDP/IP.

- *Wavelength Division Multiplexing (WDM):* A WDM is a technology that allows multiple optical signals operating at different wavelengths to be multiplexed in onto a single optical fiber and transported in parallel through the fiber.