A Framework for Traffic Engineering of Diverse Traffic Granularity Entirely on the Optical Layer Terms

by

Antonis Hadjiantonis

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

2006

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

_____	_____
Date	**Prof. Mohamed A. Ali** Chair of Examining Committee
_____	_____
Date	**Dean Mumtaz K. Kassir** Executive Officer

Prof. Samir Ahmed	(Department of Electrical Engineering, The City College of New York, CUNY)
Prof. Neophytos Antoniades	(Department of Engineering, Science and Physics, The College of Staten Island, CUNY)
Prof. Leonid Roytman	(Department of Electrical Engineering, The City College of New York, CUNY)
Dr. Andrew Wallace	(Principal Technical Staff ember, AT&T Research Labs Optical Network Performance and Reliability. 200 Laurel Ave, Middletown, NJ07748)

Supervisory Committee

Abstract

A Framework for Traffic Engineering of Diverse Traffic Granularity Entirely on the Optical Layer Terms

by

Antonis Hadjiantonis

Adviser: **Prof. Mohamed A. Ali**

This thesis addresses the important problem of how to migrate the networking functionalities and intelligence down to the optical layer, including traffic engineering, switching, and selective dynamic provisioning/restoration of diverse traffic granularity, all supported entirely on the optical layer's terms. Specifically, this work examines the technological requirements and assesses the performance analysis and feasibility for implementing a novel, simple, and scalable optical networking paradigm that can

efficiently support fully automated pure layer 1 optical networking service at any bandwidth granularity.

To realize such an ambitious initiative, we devise four optical networking innovations:

- A fully intelligent and agile optical transport layer,
- A novel hybrid optical node architecture,
- An integrated control plane that manages both layers (later-1 and layer2/3), which must be owned by the optical layer rather than by the IP/MPLS routers as it is the case in the peer model,
- A fully distributed integrated routing and signaling framework for dynamically provisioning diverse traffic granularity (on a per-call basis including both full-lambda and sub-lambda traffic flows) entirely on the optical layer's terms.

To the best of our knowledge, developing integrated routing and signaling algorithms and protocols for provisioning sub-lambda connection requests at the optical layer is a new and challenging issue that has never been addressed before. The proposed integrated routing and signaling protocols go beyond those being developed within GMPLS by the IETF and OIF. Provisioning of diverse traffic granularity entirely on the optical layer's terms, as this work will show, introduces numerous new challenges including additional control plane complexities that need to be addressed.

While both technology trends and ongoing debate and activities within both the research communities and standard bodies point compellingly to network intelligence moving up to higher layers, e.g. IP layer (favoring the intelligence of routers over optical switches), we argue, as this work will show, that moving the networking functionality and intelligence down to the optical layer (favoring the intelligence of optical switches over

routers), is more compelling in terms of simplicity, scalability, overall cost savings, and the feasibility for near-term deployment.

Finally, we show the feasibility of implementing several significant and far-reaching practical applications enabled by the proposed optical networking paradigm.

ACKNOWLEDGEMENTS

I would like to express my sincere admiration, and respect for my mentor**, Prof. Mohamed A. Ali**, for he has been a truly inspiring figure in the last five years, in both a humane and an academic context.

Many thanks to my fellow students
A. Shami, **A. Khalil** and **C. Assi**
for the stirring exchange of ideas we had.

I am deeply grateful to
Prof. Samir Ahmed and **Prof. Georgios Ellinas**
of the Department of Electrical Engineering
at The City College of New York, CUNY.

This thesis is dedicated to **my family** for all their support.

TABLE OF CONTENTS

TABLE OF FIGURES

Chapter 1

INTRODUCTION

1.1 Introduction

Recent phenomenal advances in Wavelength Division Multiplexing (WDM)-based optical networking technologies are currently beginning to shift the focus from static point-to-point WDM networking architecture (first generation optical networking) toward more dynamic, reconfigurable and switched architectures (third generation optical networking). However, the "ultimate vision" of realizing a fully intelligent optical networking infrastructure can only be achieved when most of the networking functionalities and intelligence, including switching, protection, traffic engineering, selective restoration and real-time provisioning of diverse traffic granularity, are migrated down to an agile optical networking layer.

One of the most serious limitations facing the implementation of such a vision is the large disparity between the coarse/fixed granularity bandwidth offered by the optical layer to clients (full wavelength channel, e.g. OC-48, OC-192 and OC-768), and the bandwidth requirement of a typical connection request, which is only a fraction of the wavelength channel (e.g., STS-1, OC-3, OC-12, etc.). Clearly, traffic demands with finer bandwidth granularity are the rule, and those requiring full wavelength capacity are the exception. In fact, the lack of true economic drivers for a network that can only support provisioning of coarse/fixed granularity bandwidth has been the main impediment to the adoption and deployment of a truly agile optical networking layer.

One of the most important considerations of a carrier in designing and deploying its transport network is the reliability offered by the network to the services and customers it supports. WDM-based optical transport networks have met the challenge of accommodating the phenomenal growth of Internet Protocol (IP) data traffic and, at the same time, have provided novel services such as rapid provisioning/restoration of very high bandwidth circuits, and bandwidth on demand. It is only the logical conclusion that the impact of network failures on this traffic grows at the same pace and, therefore, service reliability considerations are profoundly critical when high capacity WDM transport technologies are involved, since certain single WDM transport system failures may affect millions of connections.

There is an emerging consensus that IP-over-WDM networking architecture is the most suited for the transport and protection of the rapidly growing IP traffic [1-2]. These architectures combine the higher reachability of IP (currently, data transfers are

overwhelmingly dominated by IP traffic), and the vast bandwidth that WDM technologies enable for fiber transmission. In this scenario, the role of synchronous digital hierarchy/optical network (SDH/SONET) will diminish, and future IP networks will evolve towards a model that consists of high-performance IP/MPLS routers attached to an optical transport network that will directly provide a global transport infrastructure for legacy and new IP services. Standardizing bodies, like the Optical Internetworking Forum (OIF) and the Internet Engineering Task Force (IETF), have proposed several architectural options on how the IP routers must interact with the optical layer to achieve end-to-end connectivity, including overlay, augmented, and peer-to-peer models [1-2], which will be examined further in Chapter 2.

The issue of how to efficiently provision sub-lambda connection requests has recently received considerable attention in the context of addressing the problem of traffic grooming in mesh-based WDM networks [1-10]. To support traffic grooming, two independent networking domains must be considered: the WDM-based *optical network*[1] and the attached *client networks*[2] [7]. The role of the optical network is to provide the clients with lightpaths at the full wavelength granularity, when, in fact, client traffic has heterogeneous data rates of sub-wavelength granularity. A lightpath is an end-to-end optical pipe connection that may possibly span a number of physical fiber links. In this pipe, the traffic signal is maintained in the optical domain. Switching is achieved by the use of All-Optical Cross-Connects (OXCs), which could employ various technologies

[1] The terms "*Optical Core*", "*Optical Transport Network (OTN)*", "*Optical Layer*", "*Optical Network*" and "*Physical Layer*" will be used interchangeably to refer to the carrier WDM-based fiber optical network.

[2] Also referred to as the "*Logical Layer/Network*", "*Client Layer/Network*" and "*IP layer/Network*".

(like Micro-Electro Mechanical Switches (MEMS), Semiconductor Optical Amplifiers (SOAs), etc.). The set of established lightpaths forms the logical topology (logical connectivity between distant client networks) over which client traffic is transported.

1.2 Thesis Motivation

In general, provisioning of connections requires two mechanisms. These are *routing algorithms* for path selection, and *signaling protocols* for requesting and establishing connectivity within the network along the chosen path. While provisioning sub-lambda requests has been the main objective of all previous work on traffic grooming, all of the reported work in the literature has only considered the routing mechanism. In addition, except for a few works where requests are routed over a combination of existing lightpaths and a newly created lightpath (hybrid approach) [9-10], sub-lambda requests are always assumed to be routed over the logical topology (IP/MPLS layer) using a Single-Hop (SH) approach (traffic is routed over a single existing lightpath) and/or a Multi-Hop (MH) approach (traffic is routed over multiple existing lightpaths) [2-7]. If a route is selected, then the request is deemed successfully provisioned. Typically, routing at the IP layer is independent of routing at the optical layer and takes into account only the resource usage information at the IP layer, rendering the selected route sub-optimal.

In fact, despite the common assumption of a fully distributed networking infrastructure, none of the previous work on traffic grooming has considered the most important component of a truly provisioning process. That is the signaling mechanism needed to

setup and release resources for sub-lambda requests within the network along a chosen path.

To the best of their knowledge, the authors are not aware of any previous work on traffic grooming that has addressed or implemented a signaling protocol for provisioning sub-lambda requests over both the logical topology (IP/MPLS layer) and the optical layer. The authors believe that understanding the challenges, potential and limitations of sub-lambda signaling is of considerable importance. Furthermore, as will be shown here, performance results when taking into account the signaling component for sub-lambda provisioning are different and are not as ideal as those obtained when only the routing component is considered.

It is important to emphasize that most of the GMPLS-based routing and signaling algorithms and protocols which have been reported by standards bodies as well as research communities [11-18], were developed to provision full wavelength channels (lightpaths) at the optical layer only (the Routing and Wavelength Assignment (RWA) problem). While a number of works in the literature have dealt with the initiative of a unified control plane (peer interconnection model proposed by IETF) to integrate the optical and IP/MPLS layers into a single administrative domain that runs a single instance of routing and signaling protocols [14-18], the authors are not aware of any work that has addressed the specifics of implementing integrated signaling protocols that can simultaneously provision both full lambda (lightpaths) and sub-lambda connection requests at a single domain in a unified manner. Furthermore, any specific MPLS-based

signaling standards implementations for setting up Label Switched Paths (LSPs) are industrial proprietary and have not been published in the literature.

Within the above context, it is natural for one to assume that if the optical layer were capable of provisioning these low-speed connection requests, not only would the optical transport carriers have full control of the whole networking functionalities, but they would also be able to dynamically and efficiently support diverse traffic granularity requirement including selective provisioning and restoration.

1.3 Thesis Statement

It is the purpose of this work to present an innovative fully distributed global information-based integrated framework for real-time provisioning of diverse traffic granularity (on a per-call basis including both full-lambda and sub-lambda traffic flows) entirely on the optical layer's terms. Provisioning of diverse traffic granularity entirely on the optical layer's terms, as this work will show, introduces numerous new challenges including additional control plane signaling complexities that need to be addressed. To implement the proposed vision of an agile optical networking layer capable of supporting integrated routing and signaling algorithms to dynamically provision diverse traffic granularity, the following two salient features must be implemented:

1) Most of the networking functionalities and intelligence (including switching, traffic engineering, and selective restoration and protection on a per-call basis) must be migrated down to the optical layer, and

2) The optical layer must also own and manage both the physical connectivity (optical resources) and logical connectivity (IP resources) [19-20]. It is the author's belief that moving the networking functionality and intelligence down to the optical layer (favoring the intelligence of optical switches over IP routers, in contrast to the peer model), is more compelling in terms of simplicity, scalability, overall cost savings, and feasibility for near-term deployment.

For this reason, a network architecture model is envisioned and presented which through minimal modifications can be generalized to also address any other kind of client traffic. The implications of this proposed model, however, do not stop there. The re has recently been a lot of research on the promising and attractive idea of transferring Native Ethernet frames between end users. This is propelled by the recent advancements in fast Gigabit Ethernet. Having the "optical layer controls both wavelength and sub-wavelength connectivity" approach, one can easily interchange IP traffic with Gigabit Ethernet (GigE) traffic and explore further the above idea.

The thesis is organized as follows: in Chapter 2, current-trend IP-over-DWM interconnection models, and lessons learned from them, are examined; in Chapter 3 the proposed networks architecture is presented. Chapter 4 examines the routing component, while Chapter 5 presents a novel signaling framework fro integrated provisioning using the proposed model. Chapter 6 presents practical implications of the network model including the direct interfacing of Ethernet over the optical network. Finally, Chapter 7 presents the conclusions and discusses future work.

Chapter 2

LESSONS LEARNED FROM IP-OVER-WDM INTERCONNECTION MODELS

2.1 Introduction

The key aspect on delivering service from the client layer across a carrier optical layer (i.e. an optical network core) is how these two layers interface and interact with one another. There is an emerging consensus that the best way to achieve this is to adopt the existing IP topology self-discovery, addressing schemes and routing capabilities to the optical network environment [1-9]. Ongoing research focuses on the use of distributed management schemes such as Multi-Protocol Label Switching (MPLS) to provide the control plane necessary to ensure automated provisioning, maintenance of connections and manage the network resources. To this end, there has been quite some research on the

issue of how to extend the IP/MPLS protocol suite. This provides quality of service guarantees across an IP cloud and performs fast forwarding based on packet labels embedded by MPLS. Major optical industry organizations, like the OIF and IETF, are currently working on extending this label meaning (the conventional electronic label attached on every IP packet) of the MPLS-framework in order to support not only devices that perform packet switching (IP), but also those that perform switching in time (synchronous networks), wavelength (OXCs), and space. This extension is referred to as Generalized-MPLS (GMPLS) [4, 6-8, 10].

The OIF and IETF have proposed several architectural options on how the IP routers must interact with the optical layer to achieve end-to-end connectivity. Several interconnection models have been proposed, sometimes causing a subject of great disagreement. This is because an interconnection model dictates where the inter-domain boundaries are set and which, if any, layer "gains" in terms of significance. Finally, the overlay, augmented, and peer-to-peer models [1-14] have emerged as major candidates for standardization.

The *overlay* model is currently the simplest way of interconnection and it treats the optical layer as a completely separate entity from the IP layer. In this model, the optical layer provides point-to-point connections (lightpaths) to the IP domain. The client routers request high-bandwidth connections (lightpaths) from the optical network, via some User-to-Network Interface (UNI), and are provided with no knowledge of the optical network topology or resources. A more sophisticated model that offers a tighter integration between IP and optical layers (*peer* model) collapses the two layers into a

single integrated layer managed and traffic engineered in a unified manner. Finally, the *augmented* model is one that allows a partial flow of information across domain boundaries.

The following sections provide a more thorough examination of the Overlay and Peer models and present some of their advantages and limitations.

2.2 Overlay Model

Due to the isolation it provides between the two layers, the overlay model is the most practical for near-term deployment. This is because it is appropriate for the current telecom infrastructure that consists of multiple administrative domains, where there is clearly a need to maintain topology and control isolation between the optical transport and the service layers since they are most likely to be under different administrative controls (or ownership) and policies. Furthermore, this model provides the optical layer, which may consist of multiple sub-networks, with well-defined interfaces to its client layers. This permits each sub-network to scale independently. These, along with the simplicity of this model have led to its early endorsement by the OIF and the International Telecommunication Union (ITU). Despite the numerous benefits cited above, the simplicity of the overlay model comes at the expense of the inefficient use of network resources due to the complete separation of the state and control information between the boundaries of the optical and client layers. The problem is further compounded due to the fact that, since inter-layer isolation is in place, a single failure within one domain might cause multiple unrelated failures in other domains.

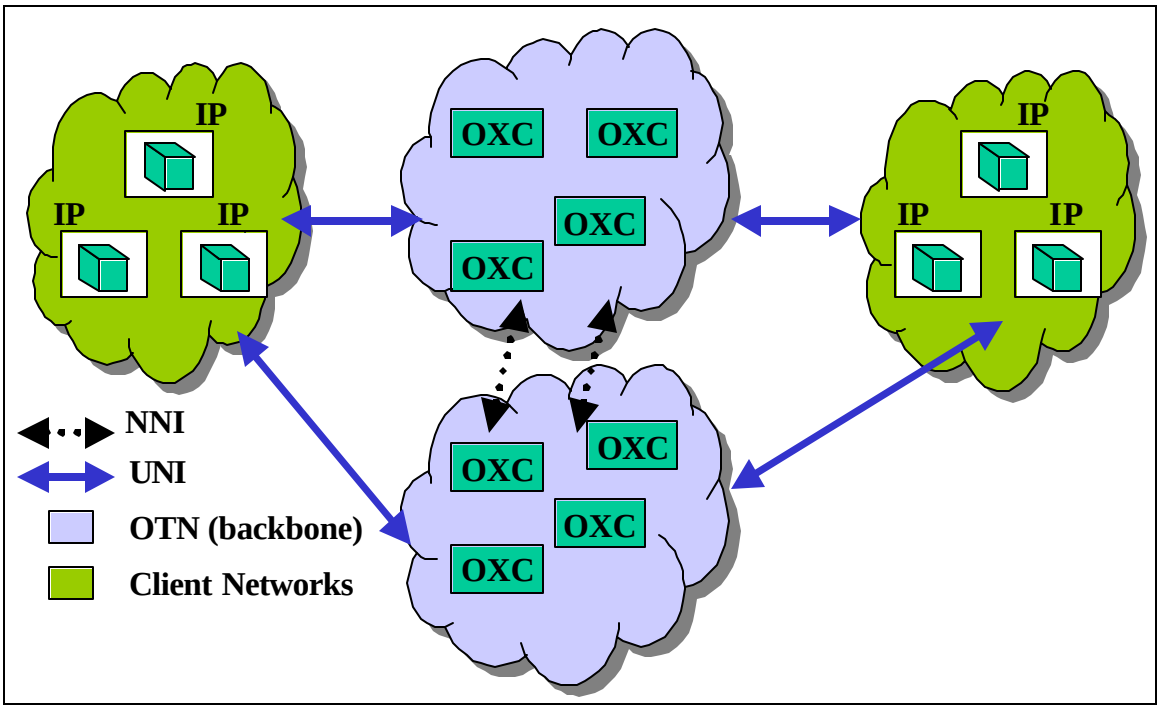

Figure 2.1: The Overlay Model

As shown in Figure 2.1, the overlay model requires two interfaces for appropriate operation. The first is the User-to-Network Interface (UNI). The UNI governs the communication between the client networks and how these require connectivity from the optical core. The second is the Network-to-Network Interface (NNI), which governs the appropriate communication between optical networks of different ownership and/or administration. Client networks are registered with the optical core through UNI and the latter provides access to the destination networks through optical pipes.

2.3 Peer Model

Unlike the overlay model, the peer model supports an integrated interconnection model where both the client and optical layers are considered one (see Figure 2.2). There is a single instance of routing and signaling mechanisms running, and the combined knowledge of resource and topology information at both the IP and optical layers are taken into account for decision-making [5]. In this regard, OXCs and routers act as peers, and using a unified control plane they establish paths that could traverse any number of routers and OXCs with complete knowledge of network resources. Thus, from a routing and signaling point of view, there is no distinction between the UNI, NNI (network-network-interface), and router-to-router interface; all network elements are direct peers and fully aware of the network topology and resources. Using an enhanced Interior gateway protocol (IGP) such as Open shortest path First (OSPF) supported by GMPLS, the edge routers can collect the resource usage information at both layers.

The unified IP/MPLS-based control plane in the peer model simplifies control coordination and fault handling among network elements with different technologies. Furthermore, this model can support end-to-end protection and failure restoration, and efficient use of resources in a network composed of multiple technologies. The peer model does, however, present two major problems:

- a scalability problem due to the amount of state and control information to be handled by any network element within an administrative domain, and

- the fact that it is highly unlikely that the carrier service provider, who owns the OTN, would ever want to give a client insight into the structure of its optical network (i.e., full access to the topology and resources of the optical network)

These problems, perhaps the second more than the first, overshadow the many advantages of this model and render its development and realization a rather distant future.

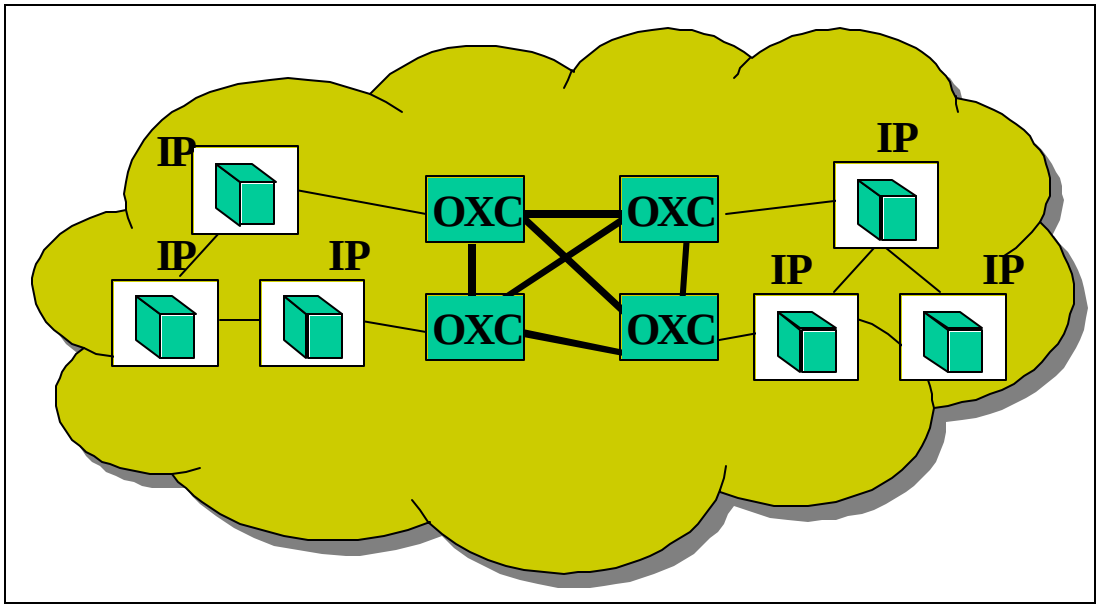

Figure 2.2: The Peer-to-Peer Model

2.4 Lessons Learned

It is clear from the above discussion that the viability of a unified control plane of an IP-over-WDM model rests entirely on avoiding the two major obstacles that hindered the practical implementation of the integrated control plane of the IP-over-WDM

interconnection peer model. Specifically, the following are the two lessons that will guide the process of devising a unified control plane:

1. The edge IP/LSR router of the integrated control plane of the peer model is the device that manages all network resources including both the physical and logical resources. Thus, the edge routers are choked by constant barrage of network state updates and optical network topology and resources, leading to a major scalability problem. So a first lesson is that ***when devising an integrated control plane that manages both IP routers and optical switches, never delegate this task to the IP routers for they are traffic bearing and costly to upgrade with high-speed electronics.***
2. Since the boundaries between the OTN and client data network are impenetrable and since it is highly unlikely that IP routers (especially those owned by a customer) would have the ability to "see" the topology and resources of the optical network and make changes, a second lesson is that ***topology isolation between the OTN and the client layer must be maintained***.

With the above in mind, one can easily reach the following conclusion: ***Devise an optical layer-based unified control plane that manages both IP routers and optical switches (analogous to the peer model), while still retaining the client/server relationship with the network (the customer has no network visibility and depends on network intelligence) and the simplicity of the optical UNI of the overlay architecture.***

The above discussion and the lessons learned from it form critical drivers that led the authors to devising a model that tries to combine the pros and avoid the cons of the overlay and the peer models. In the next chapter, the devised model will be presented.

Chapter 3

THE PROPOSED NETWORK MODEL ARCHITECTURE

3.1. Introduction

To promote and expedite the practical near-term implementation of IP-over-optical networks, this work proposes a simple and scalable IP-over-optical network interconnection model that takes the best features from both the overlay and peer models while avoiding their limitations. Specifically, the proposed model utilizes optical layer-based unified control plane that manages both routers and optical switches (analogous to the peer model), while still retaining the complete separation between the optical and IP layers of the overlay model. This is to say that the proposed model retains the advantages of these two models while avoiding their limitations.

Third generation OXC switches have the ability to perform routing and wavelength assignment as well as maintain certain operational and monitoring functionalities. This is made possible be the use of intelligent electronic modules attached to the OXCs, which are responsible for the optical fabric switch operation and monitoring. These controllers are making decisions about routing, pass switching instructing to the OXCs, test the switch configurations etc. In other words, these electronic controllers (which are non-traffic bearing) are the brains of the sophisticated optical switch.

3.2 Optical Node Architecture

It is exactly this crucial component of the OXC switch that the proposed model uses for implementation, with the assumption that it is IP/MPLS-aware and non-traffic bearing. The proposed model is achieved by shifting the control plane functionalities, previously associated with the IP layer, to these OXC controller modules located within the optical domain. These functionalities include building the logical topology and maintaining and updating forwarding tables that keep track of established sub-lambda connections between edge routers. Thus, both the logical and physical topologies now belong to a single administrative domain, leading to the creation of a unified control plane that can efficiently support both full-lambda and sub-lambda routing, signaling and survivability functionalities in the optical domain. An integrated dynamic routing algorithm that takes into account the combined topology and resource usage information, including router capacities, at both the IP and optical layers, is, then, required in order to route the data in this proposed architecture.

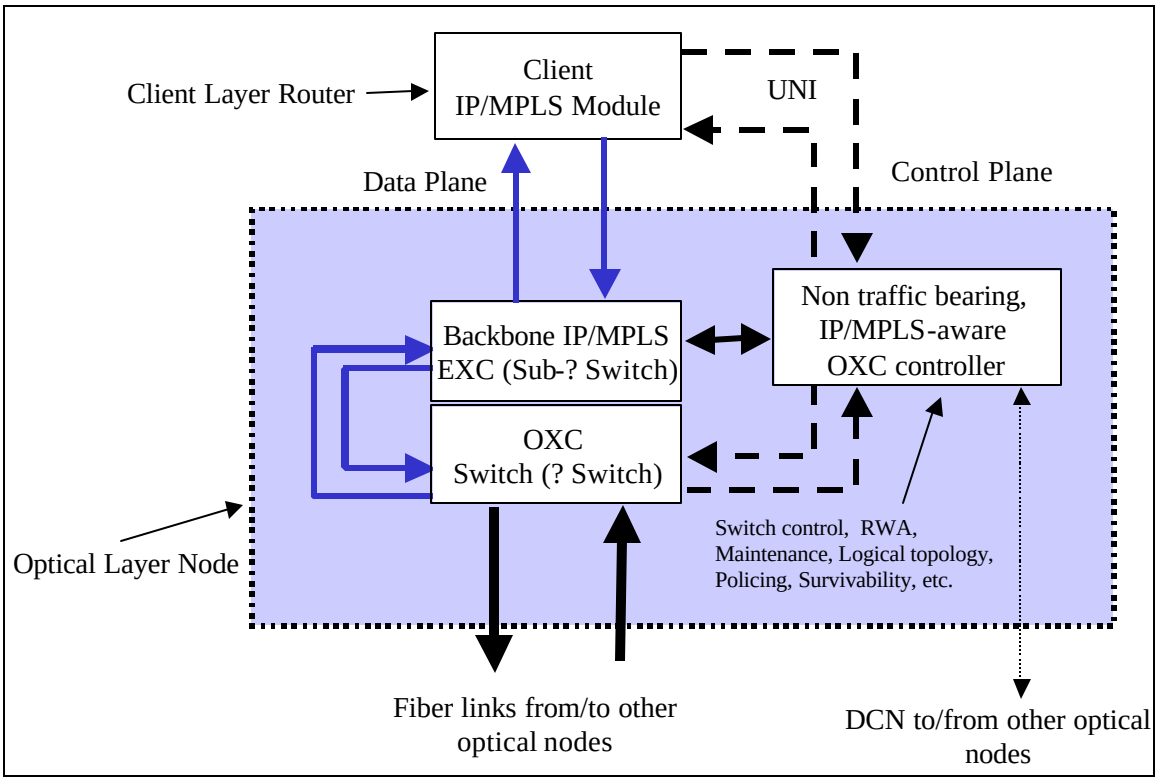

Figure 3.1: The Proposed Model Optical Node Architecture

Figure 3.1 depicts the proposed optical node architecture. As it can be seen, this architecture is composed of three components. A backbone IP/MPLS router used for Electronic Cross-Connections (EXC), shown as the sub-wavelength, or sub-?, electronic switch, an OXC switch (the ? optical switch), and an IP/MPLS-aware, non-traffic bearing OXC controller module. The backbone IP/MPLS router belongs to the optical layer and is a high-speed router capable of statistically multiplexing data streams up to the capacity that can be supported by the OXC (several full wavelength channel speeds). This router is attached to the optical switch and can generate and terminate the traffic to/from a lightpath. The OXC performs wavelength switching and wavelength multiplexing/demultiplexing. An incoming wavelength-multiplexed signal on a fiber is

first demultiplexed and then, based on its wavelength channel, will be either optically switched (i.e. continue riding the same lightpath if this is not terminated at this optical node), or dropped to the backbone IP/MPLS switch for electronic processing. From there, the signal can be either switched and added to a new lightpath (in the case of Multi-Hop logical routing), or dropped to the client network (i.e. to the local IP/MPLS router, operating in the client layer) if his is the egress optical node.

3.3. A Fully Intelligent Agile Optical Networking Layer

To realize the "ultimate vision" of an agile, fully intelligent optical networking layer capable of supporting integrated routing and signaling algorithms for real-time provisioning/restoration of connection requests at any bandwidth granularity (on a per-call basis including both full-lambda and sub-lambda traffic flows), the following two salient features must be implemented [10-13]:

1. Most of the networking functionalities and intelligence must be migrated down to the optical layer including switching, protection, traffic engineering, Operation, Administration, Maintenance and Protection (OAM&P) capabilities, provisioning of both full lambda and sub-lambda connection requests, and selective restoration (differentiated resilience for different classes of service), all supported entirely on the optical layer's terms, and
2. The optical layer must own and manage both the physical connectivity and resources (layer-1 optical resources) and logical connectivity and resources (layer-2 and 3, IP/MPLS resources). Thus, both the logical and physical topologies now belong to a single administrative domain managed and controlled by the optical layer, leading to the creation of a unified control plane with the optical layer running a single integrated routing/signaling protocol instance.

3.4. Optical Layer-Based Unified Control Plane Architecture

The main intelligence component of the proposed model is the OXC controller. This is responsible for creating, maintaining and updating *both* the physical *and* logical connectivity and with it, the optical node is capable of provisioning on-demand both lightpaths (full wavelength channels), as well as low-speed (sub-lambda) connection requests. The client IP/MPLS router responsibility is, then, simply to request a service from the Optical Transport Network (OTN) via a predefined UNI that will encompass a Service Level Agreement (SLA) to govern the specifics of the service requests. The OTN is responsible for providing this service, but in whichever way it deems optimum. For instance, it might:

- open up a new lightpath to service the new traffic (Single-Hop RWA),
- open a sequence of lightpaths to service the new traffic (Multi-Hop RWA),
- use one or more existing lightpaths to multiplex the new traffic on (Single- or Multi- Hop logical servicing), or
- use a combination of existing lightpath(s), together with setting up new ones to serve the new traffic (hybrid provisioning).

The OXC controllers can communicate with each other over a control network (Data Communication Channel, or DCN), either in- or out-of-band with data using a specific lower-speed supervisory wavelength channel.

Under this architecture, the optical core can be thought of as an autonomous system (AS) whose members (OXC controllers) are hidden completely from the outside domains. In other words, both the logical and physical topologies now belong to a single

administrative domain and all of the networking intelligence (both physical and logical connectivity) has now been shifted to the optical layer, leading to the creation of a unified control plane that can efficiently manage both optical switches and routers at the optical layer. GMPLS can support the unified control plane to provide lambda/sub-lambda routing, signaling and survivability functionalities in the optical domain. Thus, better decisions can be made for provisioning and managing network resources, leading to their more efficient use. Below, the main advantages of the proposed model are numbered:

1. Unlike the overlay model that separates the routing at each layer (routing at the IP/MPLS layer is independent of routing of wavelengths at the optical layer), the proposed model supports an integrated routing approach in the OTN where the combined knowledge of resource and topology information at both the IP and optical layers are taken into account. Note, however, that this approach is different than the peer model as well as the integrated approaches proposed in [5-6]. In those schemes the integrated routing approach is achieved through the exchange of a significant amount of state and control information between the IP and optical layers, that renders their implementations more time consuming and complex and raises the well-known scalability problem. The proposed integrated routing approach requires no exchange of information between the boundaries of the two layers except for that of the simple UNI. Thus, shifting most of the intelligence and burden from the IP layer (traffic-bearing edge routers) to the optical layer (IP/MPLS-aware, non-traffic bearing OXC controllers) renders the proposed model simpler, less expensive and most importantly alleviates the scalability problem associated with the peer model.

2. The proposed integrated routing approach is now able to service calls by selecting a mixture of existing (logical routing) and new (physical routing) lightpaths in a way that optimizes resource usage. This introduces the important concept of hybrid provisioning which will be explained below.

3. The impact of this approach on the network protection/restoration strategy is far reaching. The conventional notion that the optical layer is operating independently and has no awareness of a router failure and no role in the restoration of such a failure is no longer applicable. The physical layer now keeps updated database about both the logical and physical connectivity, including all connection requests. Thus, in the case of an edge traffic-bearing router failure, the physical layer can independently restore all the disrupted traffic streams traversing the router (sub-lambdas and/or full lambdas). It is important to note that this is not possible if the optical core is not maintaining the logical connectivity, in which case, when a forwarding router fails, the forwarded traffic can only be restored at the IP layer, which is much slower restoration process.

With the introduction of the above model, many of the integrated routing algorithms already discussed in the literature are now possible. In the following chapters we will discuss the routing and signaling mechanisms necessary for automated provisioning under a fully agile optical layer.

Chapter 4

THE ROUTING MECHANISM

4.1 Introduction

Service provided by the optical layer is governed by two key components; namely, the routing and signaling components. This chapter deals with the routing component and will attempt to provide insights on both layered and integrated routing approaches. In layered sequential routing, a call request attempts to find a valid path from its source to its destination on a given layer's topology. In other words, if the call is tried on the logical topology, the routing algorithm will only have the logical links (lightpaths) and nodes (in case of IP, IP routers) available; if, on the other hand, it was to be tried on the physical topology the physical topology components would be visible. In integrated routing, the two layer topologies are collapsed to a single integrated topology containing

both physical and logical residual resources, and on which the call request is tried. It is clear that the knowledge of both layer resources would be beneficial as more insightful decisions can be made on the path (set of links) to be chosen, and its characteristics (logical vs. physical links).

Logical routing is based on the logical topology, which is made up from the client edge routers (logical nodes) and the existing setup lightpaths (logical links). Single-Hop approach is when a call is only using one existing lightpath with available bandwidth resources, whereas the Multi-Hop approach allows the call to use sequentially more than one logical links. Apart from the bandwidth resources on logical links, logical routing can also be constrained by the electronic speed capacity of the IP/MPLS routers, since these act as the electronic multiplexing and demultiplexing agents to and from wavelength channels. With respect on how the logical topology is maintained, the logical topology is called static (or incremental) if no logical links are torn down once they have no traffic on them, or dynamic otherwise.

In general, when dealing with WDM-based optical networks, physical routing refers to solving the Routing and Wavelength Assignment (RWA) problem. As the name implies, this is made up of two sub-problems, which deal with path discovery (routing), and assigning wavelength channel(s) along its fiber links. In networks where wavelength translation (or conversion) is not supported, one wavelength must be assigned along the whole path in observance of the wavelength continuity constraint. Many ways of solving the RWA problems have been proposed including treating the sub-problems separately,

or trying to solve them simultaneously [1, 4]. Clearly, in the presence of full wavelength conversion at the OXCs, the wavelength assignment problem is relaxed.

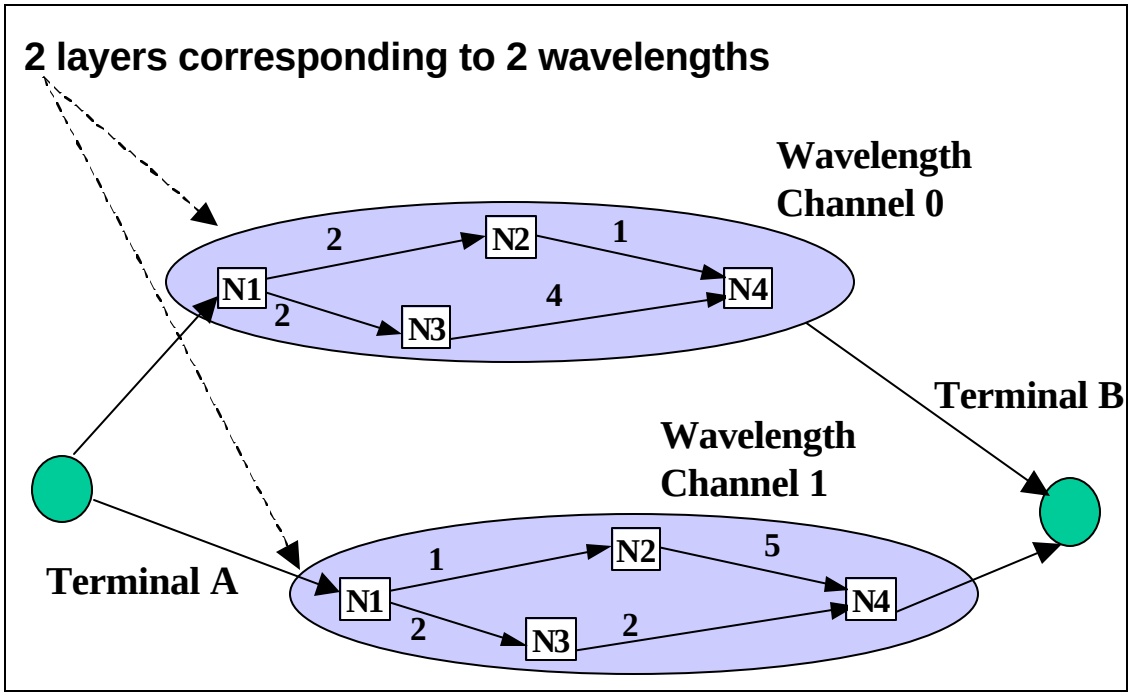

Figure 4.1: Modeling different wavelength channels as Layered Graphs when jointly solving the RWA problem

When the sub-problems of RWA are treated separately, a shortest path routing algorithm (i.e. Dijkstra, or Bellman-Ford) is used to get the path. Then, the wavelength assignment is added based on some fashion (like First-Fit, Random-Fit, Most-Used, Least-Used etc). In the absence of wavelength translation, and when RWA is attempted to be solved jointly, each wavelength channel represents a distinct (layered) graph of the physical resources. These graphs include the same set of nodes but only include physical fiber links that are available on the given wavelength channel. Once the shortest paths are obtained for each wavelength channel, the one path that is shortest among them will also yield the wavelength assignment. Observing Figure 4.1, there are 2 wavelength channels

on a simple 4-node physical topology. If a request is generated between Terminal A and B, then the layered approach for solving the RWA is depicted. With respect to the same Figure, wavelength channel 0 shortest path is N1-N2-N4 (cost is 3) while wavelength channel 1 shortest path is N1-N3-N4 with cost 4. Clearly the first option is of lower cost and will dictate both path N1-N2-N4 and wavelength channel 0 as the solution to the RWA problem. Whenever sparse wavelength translation is present, it would be depicted as links connecting the two layers at the nodes who accommodate it.

4.2 Sequential Routing Approach

Under the sequential routing approach, the logical and physical layers maintain their distinct topologies. This is to say that if a call is tried on the logical topology, it can only use existing setup lightpaths for its service, and if it was to be tried on the physical topology, there would be a need to solve the RWA problem in order to have a valid set of links and a valid wavelength channel on it. To this end, many algorithms can be proposed, and below we present a small contribution.

4.2.1 Interchanging the Search Order

Rather than trying to optimize the conventional performance parameters (number of transceiver arrays at each border router, forwarding speed of electronic routers/switches, number of wavelengths per fiber, etc.) that should lead to an optimum logical topology, we try to intrinsically optimize the logical topology by means of interchanging the search

space between logical and physical layers. We present three search schemes and compare their performances. In all of the three algorithms presented below (in Figures 4.2, 4,3 and 4.4), a connection request is constrained to a maximum allowable logical links of 2. This is because, as will be shown below, Multi-Hop logical routing up to 2 hops seems to give the best performance. The logical routing approach uses the shortest-widest criterion. That is the shortest path is in terms of number of links, and ties are broken by using the path whose narrowest logical link is the widest among the candidates (in terms of residual bandwidth).

Step1:	For each connection request, invoke [shortest-widest] logical routing. If a path is returned and that path has logical links less or equal to a maximum allowable logical links, reserve bandwidth and go to step5. Else go to step2.
Step2:	Invoke the RWA to setup a new lightpath between the s-d pair. If blocked, go to step4 else go to step3.
Step3:	Setup the lightpath update the logical connection list.
Step4:	Call blocked.
Step5:	Call succeeded.

Figure 4.2: The Logical-First Algorithm

4.2.1.1 The Logical-First Algorithm:

With this algorithm, the network first tries to service the call on the logical layer making use of the already existing connections. The algorithm is summarized in Figure 4.2.

The Physical-First Algorithm:

With this approach, the network looks to service a call from the physical layer first. Only upon failure of the RWA algorithm will the logical topology be created and searched. The algorithm is summarized in Figure 4.3.

Step1	For each connection request, invoke the RWA to setup a lightpath between the s-d pair. If blocked, go to step2 else go to step3.
Step2	Invoke [shortest-widest] logical routing. If a path is returned and that path has logical links less or equal to a maximum allowable logical links, reserve bandwidth and go to step5, else go to step4.
Step3	Setup the lightpath update the logical connection list.
Step4	Call blocked.
Step5	Call succeeded.

Figure 4.3: The Physical-First Algorithm

4.2.1.3 The Interchange Approach:

The proposed interchange algorithm first tries to route a call on the logical layer if and only if a single logical connection (pre-established lightpath) from the source to the destination exists. If that fails, then the RWA algorithm is invoked and if it, in turn, is unsuccessful in servicing the call, the algorithm tries to service the call on a two-logical-link path on the logical topology. The algorithm is summarized in Figure 4.4.

Step1: For each connection request, set a call_flag DOWN. Invoke [shortest-widest] logical routing. If a path with exactly one logical link is returned, reserve bandwidth and then go to step6. If a path with exactly two logical links is returned, set the call_flag UP, and save the path. Go to step2.

Step2: Invoke the RWA to setup a new lightpath between the s-d pair. If blocked, go to step3 else go to step4.

Step3: If the call_flag is UP, reserve bandwidth on the saved path and go to step6, else go to step5.

Step4: Setup the new lightpath and update the logical connection list.

Step5: Call blocked.

Step6: Call succeeded.

Figure 4.4: The Interchange Algorithm

The performance of the proposed algorithms is evaluated via simulation of the mesh-based 14-node NSF network, shown in Figure 4.5. Each adjacent node pair is interconnected by 2 unidirectional fibers from east to west and another 2 unidirectional fibers from west to east (total of four fibers). Each fiber carries 4 wavelengths and each channel speed is assumed to be 2.5Gbps (OC-48). Each node generates flows randomly and those are equally likely to be destined to any node in the network. Their demand in bps follows a normal distribution centered at 400Mbps with a standard deviation of 200 Mbps and exponential holding time distribution with mean $1/\mu$, where μ varies. The

overall blocking performance is used as the metric for comparing the performance of the propose algorithms. Each IP/MPLS border router has a processing speed of N Gb/s and M arrays of transceivers. Since the main objective of this work is to explore the effect of interchanging the search space between logical and physical layers the values of M and N are chosen such that the overall blocking probability is independent of both of them.

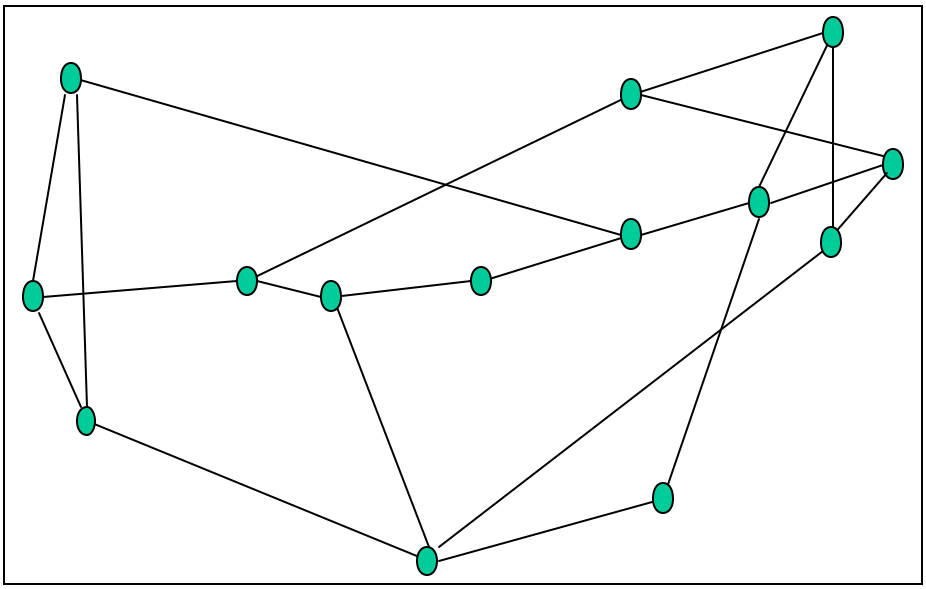

Figure 4.5: The topology fro the 14-node NSF Network

Figure 4.6 investigates the effect of the maximum number of logical hops that a connection request is allowed to be routed on, for logical servicing. As the results indicate, up to 2 logical hops seems to exhibit the best performance (which has been an initiative for constructing the interchange algorithm). This is because limiting logical routing to 2 links achieves a balance of efficient use of the wavelength bandwidth while limiting wasting physical resources by using extra logical links. Note that when the network is operating at very high loads (> 1600 erlangs, not included in Figure 4.6), the

blocking probability improves when the number of maximum allowable logical links decreases, with single logical hop giving the best performance. This is due to the fact that, at very high loads, not utilizing the optimum routes from a source to destination is increasing the probability of future calls to be blocked.

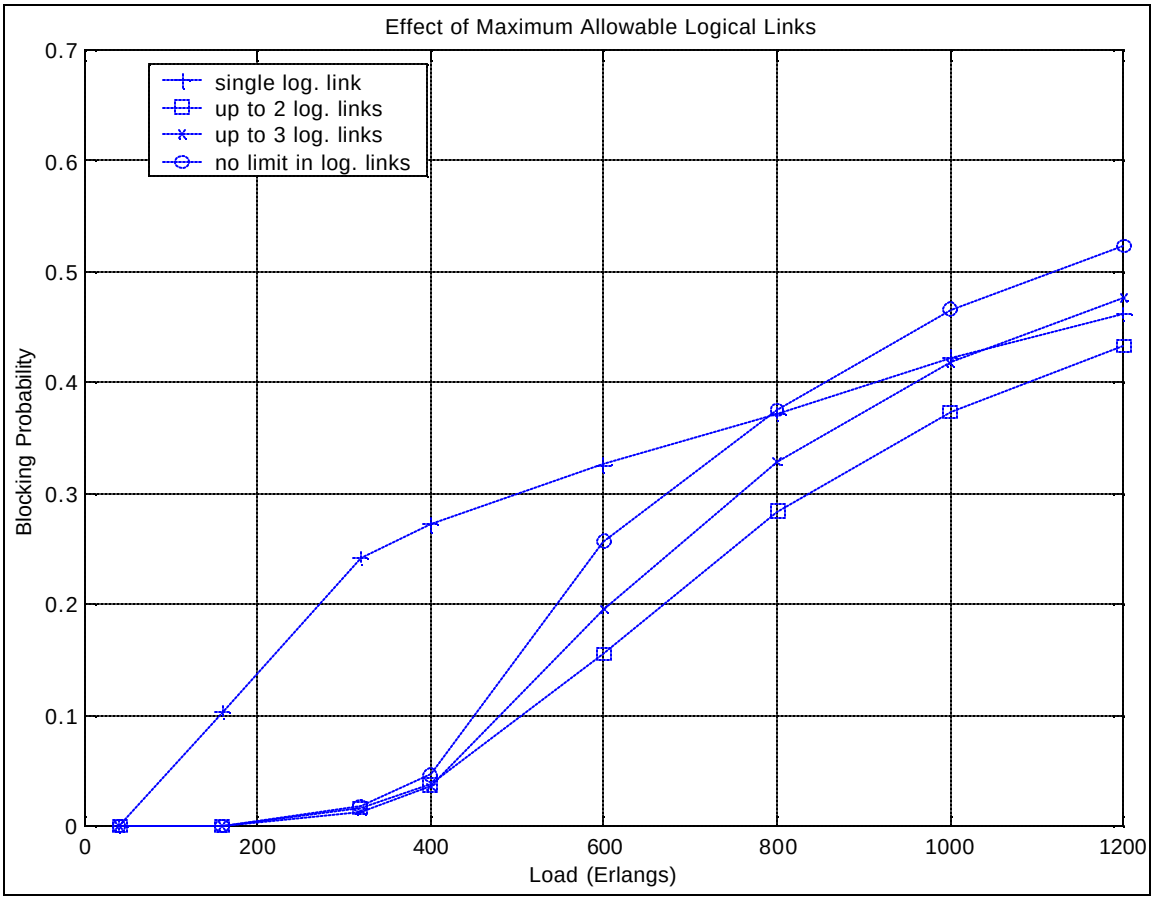

Figure 4.6: Logical-First Algorithm; Comparison of the effect of maximum number of logical links in Routing

Figure 4.7 compares incremental and dynamic logical topologies when the logical-first algorithm is used for various maximum logical link constraints. The graph indicates that dynamic logical topology creation improves performance. However, the improvement

vanishes as the number of logical links allowed increases. This is because when more logical links are allowed, the number of torn-down lightpaths (the only difference between incremental and dynamic topologies) vanishes, and so incremental and dynamic topologies converge in performance.

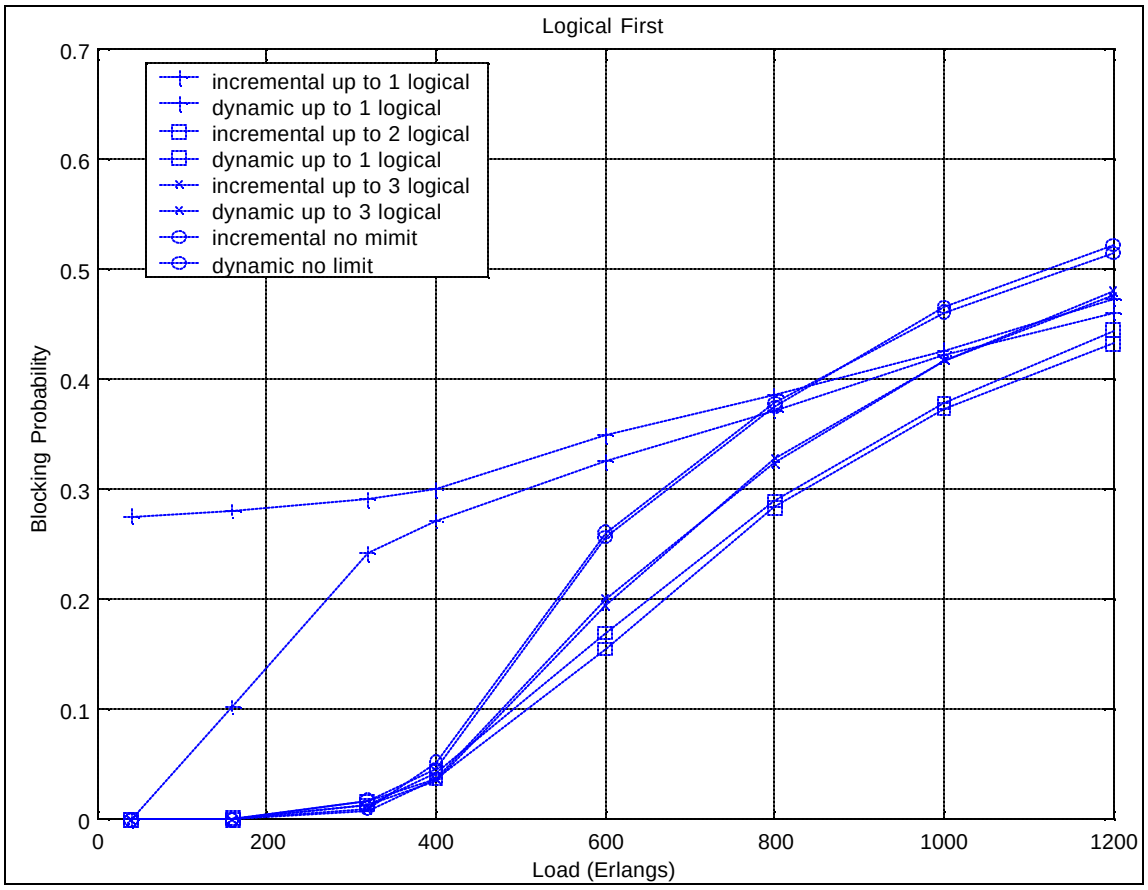

Figure 4.7: Comparing Incremental and Dynamic Logical Topologies

Figure 4.8 compares the logical-first, physical-first and interchange algorithms for dynamic topology construction. The results show that the interchange algorithm improves performance as it attempts to efficiently use bandwidth without wasting network resources by first trying single-hop logical routing. Then the algorithm tries to set up a new lightpath that enriches logical connectivity, and only if these two attempts fail will it

attempt to service the call on two logical links. It is important to emphasize that this conclusion is always true, regardless of the values chosen for both M and N.

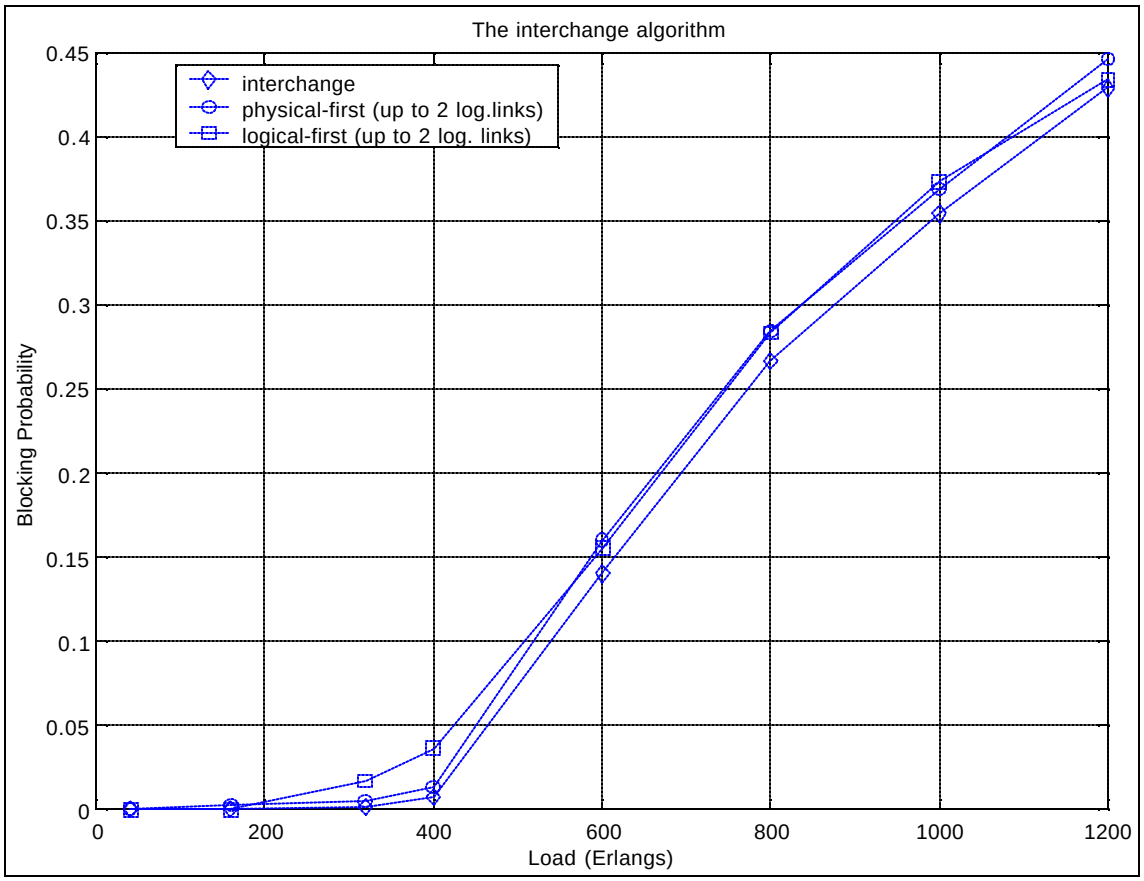

Figure 4.8: Comparison between the three sequential algorithms

4.3 Hybrid Approach (On the Integrated Graph)

This routing approach makes partial use of the model proposed in Chapter 2 that combines resource information across the layers. Since the optical layer keeps updated databases of both the logical and physical connectivity, including all connection requests, calls can now be serviced by selecting a mixture of existing lightpaths (logical routing) and setting up new lightpaths (by solving the RWA problem) in a way that optimizes

resource usage. The goal of this hybrid provisioning approach is to find an intermediate node *I* that splits the source *S* to destination *D* path into an existing segment (obtained from the existing logical topology) and a segment to be setup (created using physical routing - RWA). In this way the number of lightpaths that are used to service a call under any approach will not exceed two. Three new algorithms for the hybrid approach are proposed; namely *Shortest Path First-Fit*, *Shortest Path Exhaustive Search* and *Network--Wide Exhaustive Search*. Below we present them in more detail.

4.3.1 Shortest Path First-Fit (SPFF)

In order to search for an intermediate node *I*, this algorithm examines intermediate nodes on the static physical shortest path, to determine if there already exists a lightpath (an existing connection with adequate residual bandwidth to multiplex the new request) either connecting the source *S* and *I*, or *I* and the destination *D*. If such a path exists, the algorithm tries to complete the remainder of the *S*-*D* path by invoking the dynamic RWA used in this work with *I* as the source or destination, accordingly. If the RWA is successful, the search stops and the valid hybrid path is used.

4.3.2 Shortest Path Exhaustive Search (SPES)

In order to search for an intermediate node *I*, this algorithm examines all possible intermediate nodes on the shortest path that minimizes the hop-count, to determine if there already exists a lightpath either connecting *S* and *I*, or *I* and *D*. If such a path exists,

the algorithm tries to complete the remainder of the *S-D* path by invoking the dynamic RWA used in this work with I as the source or destination, accordingly. If the RWA is successful, the hybrid path is stored and the next *I* on the shortest path is examined. When all possible intermediate nodes are examined, a list of the feasible hybrid paths is constructed. The algorithm then chooses the one with the smallest hop-count in its "lightpath-to-set" portion. In case of a tie, the hybrid path is chosen randomly from the candidate paths.

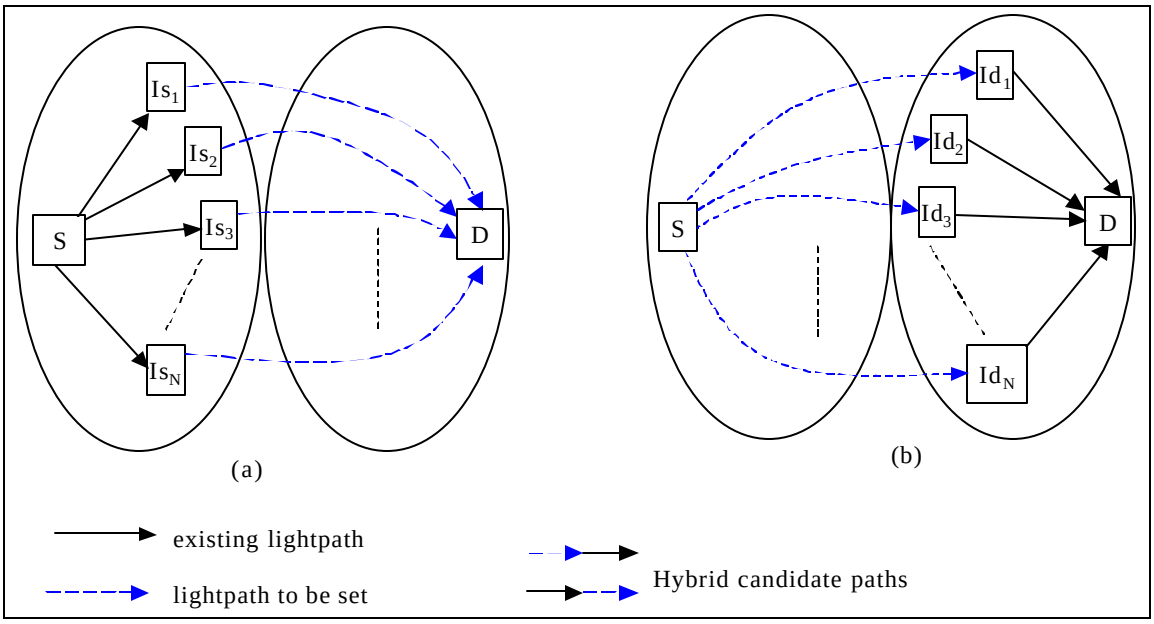

Figure 4.9: Depiction of Logical Topology Partitioning for NETWES

4.3.3 Network-Wide Exhaustive Search (NETWES)

This algorithm is a generalization of SPES. It is not limiting the search space for *I* on the static shortest path, but, instead, searches for reachability points across the whole network. It is easier to visualize this by partitioning the network into two regions: one

containing nodes that *S* has an existing lightpath to (excluding *D*, even if there exists a lightpath to *D*), and one with the remaining nodes (including *D*).

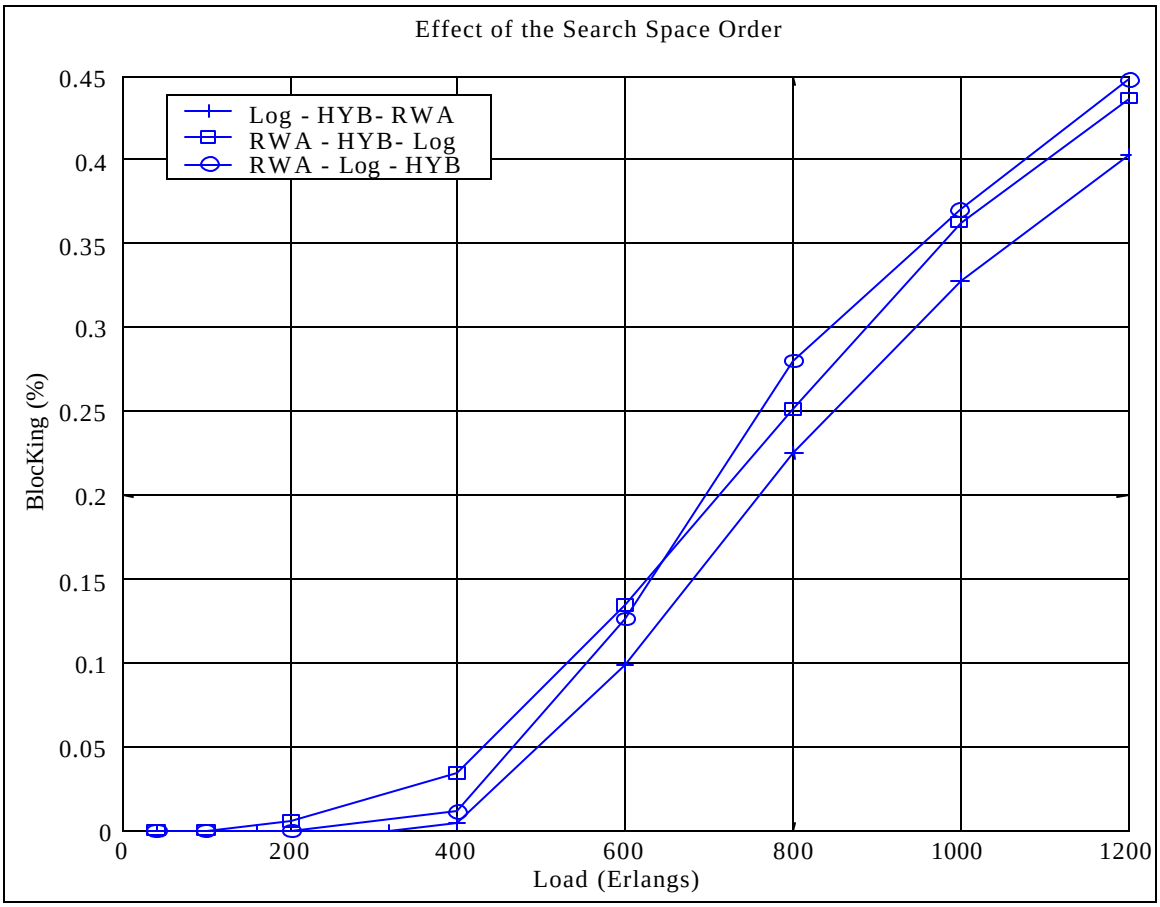

Figure 4.10: Interchanging the order including the hybrid approach

This is shown in Figure 4.9 (a). From the nodes in the first partition we store hybrid paths for the ones that have a successful RWA invocation (from Is_i to *D*). Then, a similar partitioning is performed but now the candidate reachability nodes are the ones that have existing paths to *D*, as illustrated in Figure 4.9 (b). Here, the possibility of having existing paths between Id_i and *D* (and completing the hybrid paths by setting new lightpaths from *S* to Id_i) is explored. As before, the hybrid path with minimum hop-count on its "lightpath-to-set" portion is chosen.

This work assumes sequential provisioning to service a given request. For instance, LOG-HYB-RWA provisioning specifies that the logical search is performed first and if it fails, the hybrid algorithm is invoked, and if that fails the RWA algorithm is invoked. A call is blocked if all three approaches fail sequentially. A dynamic RWA is assumed that attempts to solve the routing and wavelength assignment problems jointly by means of dynamic routing over multi-layered graphs (each layer represents a single wavelength) under a multi-fiber environment [1]. The logical topology construction is performed whenever the provisioning algorithm attempts to service a call using the logical topology, since the latter changes every time a lightpath is set-up or torn-down. The logical link cost is based on the normalized used bandwidth of the link after the call is accommodated.

The performance of the proposed approach is evaluated by simulating the mesh-based NSF network consisting of 14 nodes and 21 bi-directional links (Figure 4.5). Adjacent nodes are connected through a bi-directional physical link that consists of 4 fibers (two in each direction), where each fiber is assumed to have 4 wavelengths. We use a dynamic traffic model in which call requests arrive at each node according to a Poisson process and the session holding time is assumed to be exponentially distributed. The wavelength channel capacity is assumed to be OC-48 (~2.5 Gb/s). The sub-lambda requests have bps demands that are normally distributed around 400 Mbps with a standard deviation of 200 Mbps, in multiples of 50Mbps. The backbone routers are assumed to have enough interfaces and process all the traffic that can potentially pass through them (this assumption can be relaxed to account for the cases where routers have limited processing

capabilities). We further assume a dynamic logical topology; i.e. lightpaths are torn-down whenever the last call utilizing them departs the network.

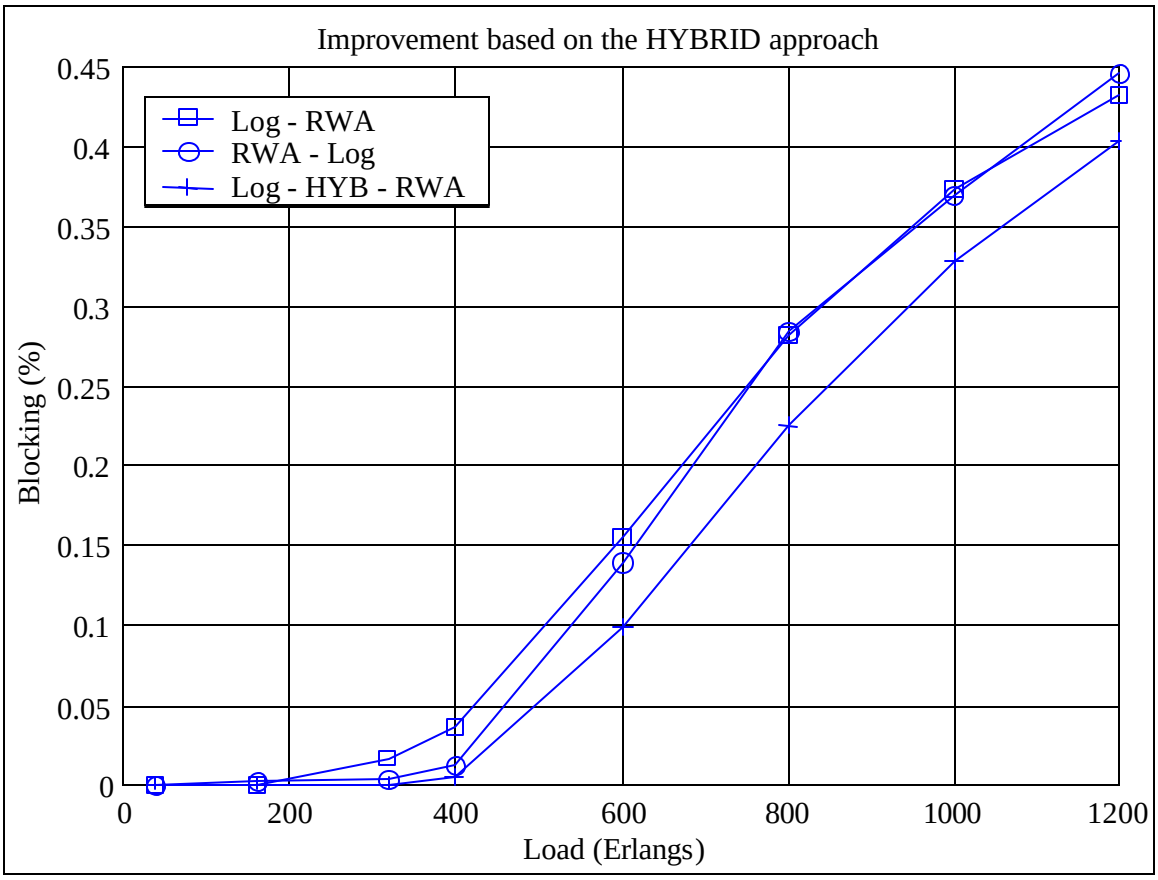

Figure 4.11: Improvement introduced by the hybrid approach

Investigation of the effect of changing the sequential search order for provisioning a call has yielded the order LOG–HYB-RWA as the one that exhibits the best performance in terms of blocking probability (the investigation is shown in Figure 4.10). The results reveal that it is important to have the hybrid approach before the RWA as this enriches the logical topology connectivity. Using this result, Figure 4.11 now compares the

performance of the conventional overlay sequential search (LOG-RWA or RWA-LOG) with that of the proposed one (LOG-HYB-RWA). It is clear that when the hybrid approach is presented, the performance of the network improves significantly compared to that of the overlay model.

Figure 4.12 compares SPFF, SPES and NETWES algorithms and as expected, NETWES is performing better with the SPES following. This is because exhausting the search for an optimum *I* optimizes the selection over the search space (nodes on the shortest path, or all the nodes) whereas the first-fit algorithm relies on the first success only.

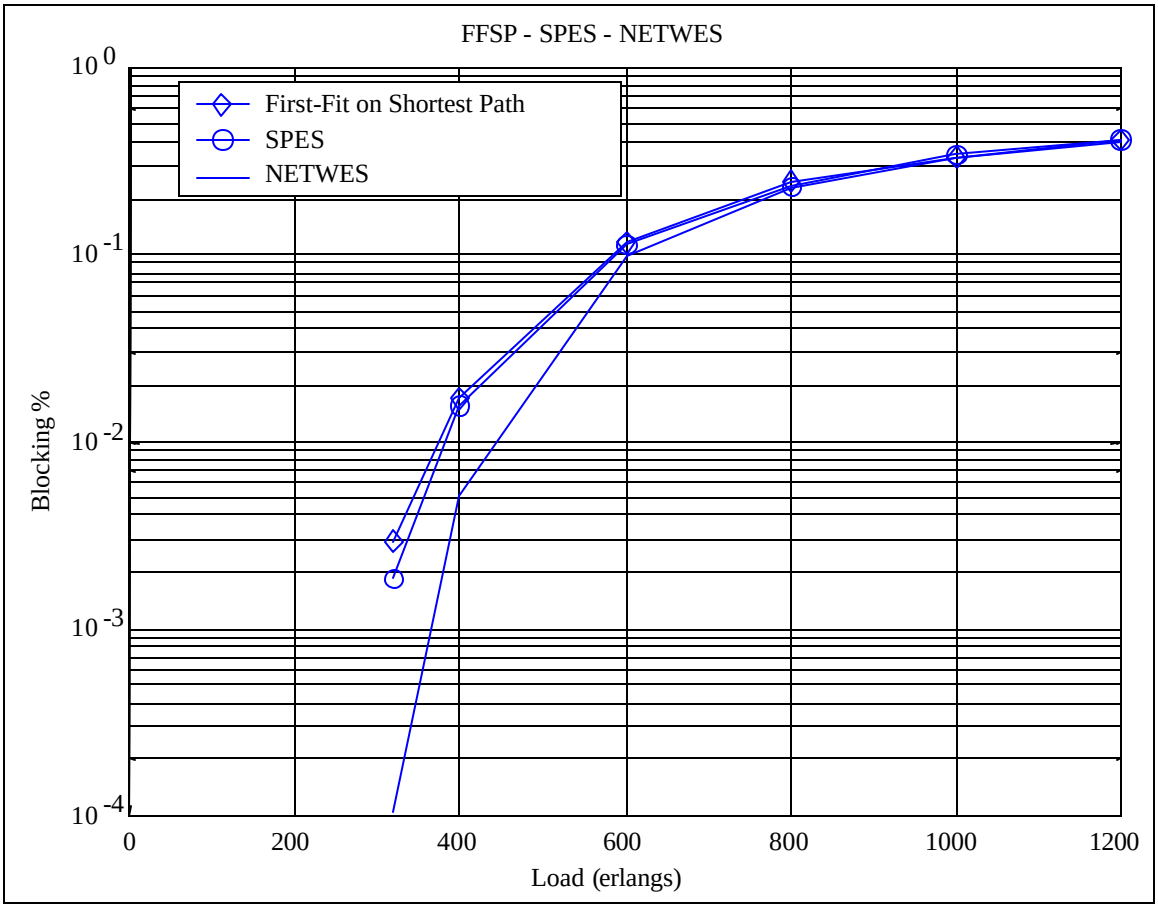

Figure 4.12: Comparing the three hybrid algorithms

4.4 Integrated Routing

There are few studies in the literature that have addressed the integrated routing approach [2-3]. The authors in [2] proposed two dynamic integrated routing algorithms, namely the Integrated Min-Hop (IMH) and the Maximum Open Capacity (MOCA). The IMH algorithm considers only a route with minimum hops and doesn't take into account the network resources. The MOCA selects a route that maximizes the residual capacity between the router pairs. Although MOCA outperforms IMH, however, the selection of a route involves lengthy and complex computation of the maximum flow values for all router pairs in the network. In [3], the authors introduce a varying optimization parameter k to balance the physical layer cost with that of the logical layer. The optimum value of k is not well defined and varies drastically with network load. In contrast to previous approaches that only consider the integrated routing in single-fiber IP/WDM networks, our proposed approach is simple and considers multi-fiber environment.

4.4.1 Integrated Graph Construction

The network is modeled as an integrated directed layered-graph where each layer represents a wavelength [1, 4]. On the integrated layered-graph, an established lightpath on wavelength channel $?_{layer}$ between a pair of ingress-egress nodes is modeled by a cut-through edge connecting them. Once the path is created, the corresponding physical links connecting the same ingress-egress pair on $?_{layer}$ are removed from the original graph. Thus, a cut-through path in the integrated layered-graph represents a logical link (LL) in

the IP layer while a non cut-through path represents a physical link (PL) in the optical layer. An integrated link (IL) represents a logical or physical link. Since multi-fiber WDM networks is assumed, a wavelength channel (WC) is identified here by a triple <*L*, *?*, *f*> identifying the wavelength of the channel *?*, and fiber *f* on link ID *L*.

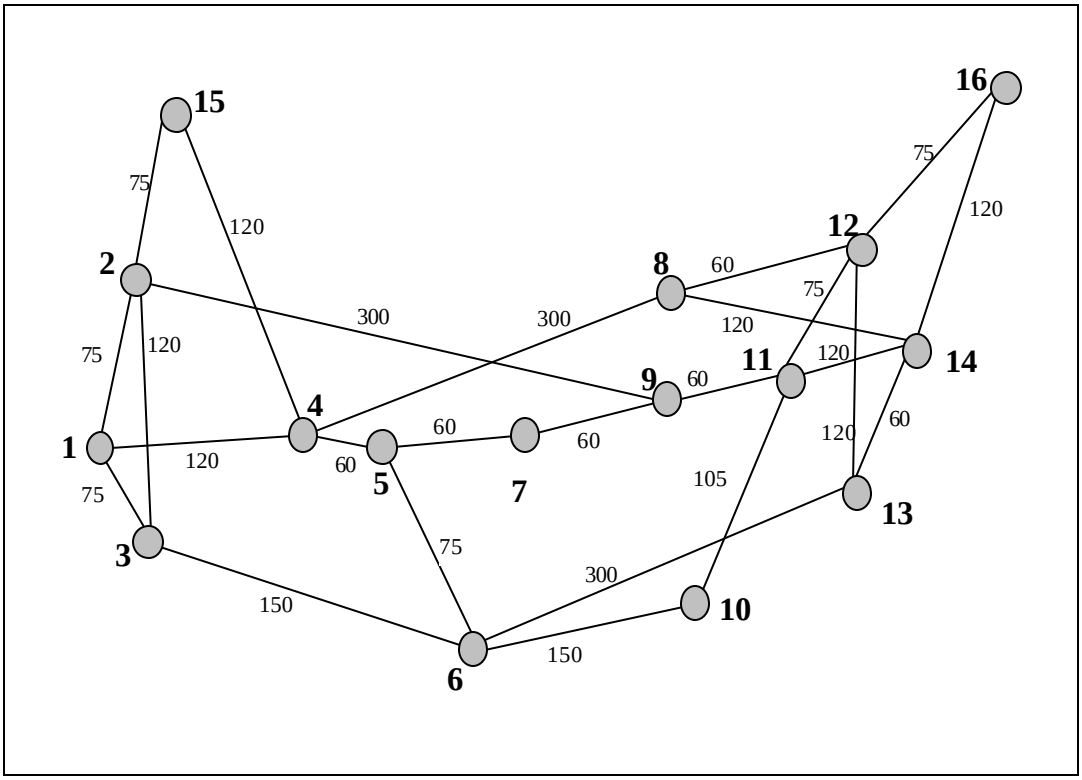

Figure 4.13: The NSF16 Network representing a US backbone

The idea behind the integrated graph is that a call arriving at the network can be served by setting up a new lightpath (purely RWA provisioning), by using one or more existing lightpaths (purely logical provisioning), or by using a combination of new and existing lightpaths (hybrid provisioning). This means that if no restrictions are placed on the creation of the integrated topology, routing decisions will have no limit on the physical and logical links used (referred to as "no constrains", or NC approach). However, to employ a practical integrated routing approach, the specifics of the routing path need to

be considered. Specifically, the number of segments (existing or to be setup using RWA) that will serve a call could degrade the data signal (physical transmission impairments) and thus the path should be limited in length.

For this reason, the maximum number of segments comprising a path is restricted to a maximum of three segments; at most a single newly created lightpath and a maximum of two existing segments (referred to as Single RWA Segment, or SRWAS approach). To realize this, given a source *s* and destination *d* pair, the integrated graph is constrained to only include existing logical links that *originate at s, or terminate at d.* Furthermore, the wavelength continuity constraint is assumed to be obeyed (i.e. optical switches have no wavelength conversion capability). In this way, the routing algorithm on a given wavelength's integrated graph (e.g. $?_{layer}$) is forced to return one of the following six options:

- an existing lightpath from s to d,
- a sequence of physical links from s to d on $?_{layer}$ (setting up a new lightpath)
- a sequence of two existing lightpaths (two logical hops) from s to d through an intermediate node i,
- a sequence of physical links $?_{layer}$ from s to i, and an existing lightpath form i to d,
- an existing lightpath from s to i and a sequence of physical links on $?_{layer}$ from i to d,
- an existing lightpath from s to i_1, a sequence of physical links from i_1 to i_2 on $?_{layer}$, and an existing lightpath from i_2 to d.

4.4.2 Integrated Routing Algorithms

Each link in the integrated graph (both logical and physical) is assigned a cost to reflect the weight of that link. The link cost can be a function of a number of metrics such as number of hops, residual link bandwidth, router interface capacities, the number of O-E-O conversions, etc. After a cost is assigned to each link, our proposed integrated routing algorithm uses the simple shortest path computation (Dijkstra's algorithm) to calculate the explicit route. The shortest path is computed taking into account both the logical and the physical links. Thus, a selected path may use an existing lightpath (logical routing), set up a new lightpath (RWA), or use a mixture of existing and new lightpaths (hybrid approach) in a way that optimizes resource usage. In the following section, two different cost metrics will be developed, namely, Least-used Cost (LUC) and Future-Based Cost (FBC), and examine the tradeoffs between them to decide on an optimum or near-optimum link cost. The challenge is how to assign a uniform cost across both the logical and physical links.

In general, the total cost of a path, C_{path}, is defined as:

$$C_{path} = \sum_{PL_i, LL_j \in path} (C_{PLi} + C_{LL\,j}),$$

where C_{PL} and C_{LL} are the costs of a PL and LL, respectively. Once the costs of all links (logical and physical) are assigned, minimum cost routing is performed using the Dijkstra's algorithm, and the cheapest route is selected.

Further assumptions are:

a) Routing is performed on a layered graph where each layer represents a wavelength channel, so the PLs on a routing decision are all on the same wavelength.

b) The network is multi-fiber.

c) The costs are assigned after a call request arrives.

d) The LLs considered are pruned (i.e. only the ones that are able to carry this call's demand in bandwidth are considered).

e) The network traffic bandwidth demand statistics are somehow estimated.

4.4.3 Integrated Cost

In the proposed cost functions, the additive cost entity is C_{PL}. On a given layer (wavelength channel $?_{layer}$) assume the following notation:

f_{total} : the number of unidirectional fibers connecting a pair of neighboring nodes (assumed constant),

BG : the base granularity (e.g. 50Mbps), assumed constant,

BW_{total} : the wavelength channel capacity, in number of BGs (assumed constant),

$f_{PL_i}^{used}$: the number of used $?_{layer}$ channels on PL_i at the time of arrival of this call (i.e. the number of fibers on which $?_{layer}$ is used),

$f_{PL_i}^{available}$: the number of available $?_{layer}$ channels on PL_i (i.e. the number of fibers on which $?_{layer}$ is available, equal to $f_{total} - f_{PL_i}^{used}$),

$BW_{LL_i}^{available}$: the available BW on LL_i (in number of BGs) at the time of arrival of this call,

$BW_{LL_i}^{used}$: the used BW on LL_i (in number of BGs) at the time of arrival of this call, equal to $BW_{total} - BW_{LL_i}^{available}$,

$N_{LL_i}^{PLs}$: the number of PLs in LL_i

CR_{BW} : this call's BW requirement (in number of BGs).

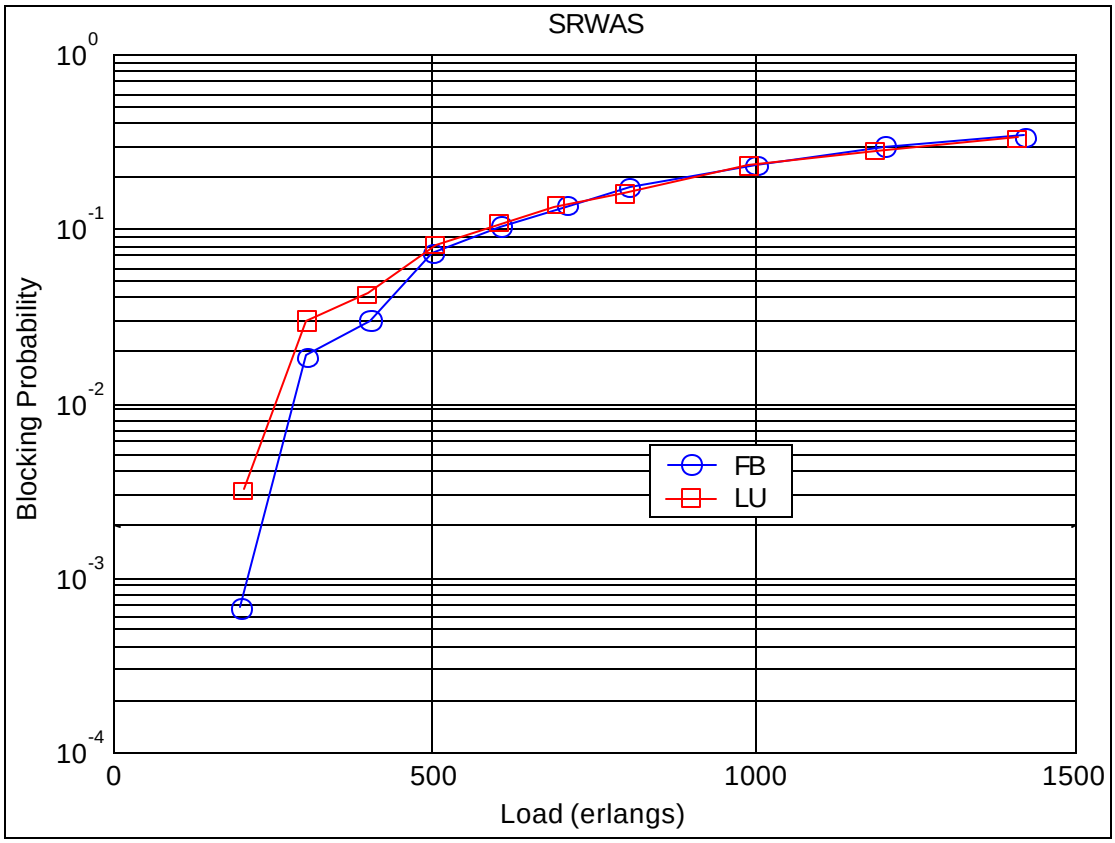

Figure 4.14: Comparison of Least-Used and Future-Based Integrated Routing Cost Assignments

4.4.3.1 Least Used (LU) Cost Function

With this cost, the least used ILs are favored. The costs are defined as:

$$C_{PL_i}^{LU} = \begin{cases} \dfrac{1 + f_{PL_i}^{used}}{f_{total}}, & \text{if } f_{PL_i}^{used} < f_{total} \\ \infty, & \text{otherwise} \end{cases}$$

$$C_{LL_i}^{LU} = \frac{N_{LL_i}^{PLs} \times (BW_{LL_i}^{used} - CR_{BW})}{BW_{total}}.$$

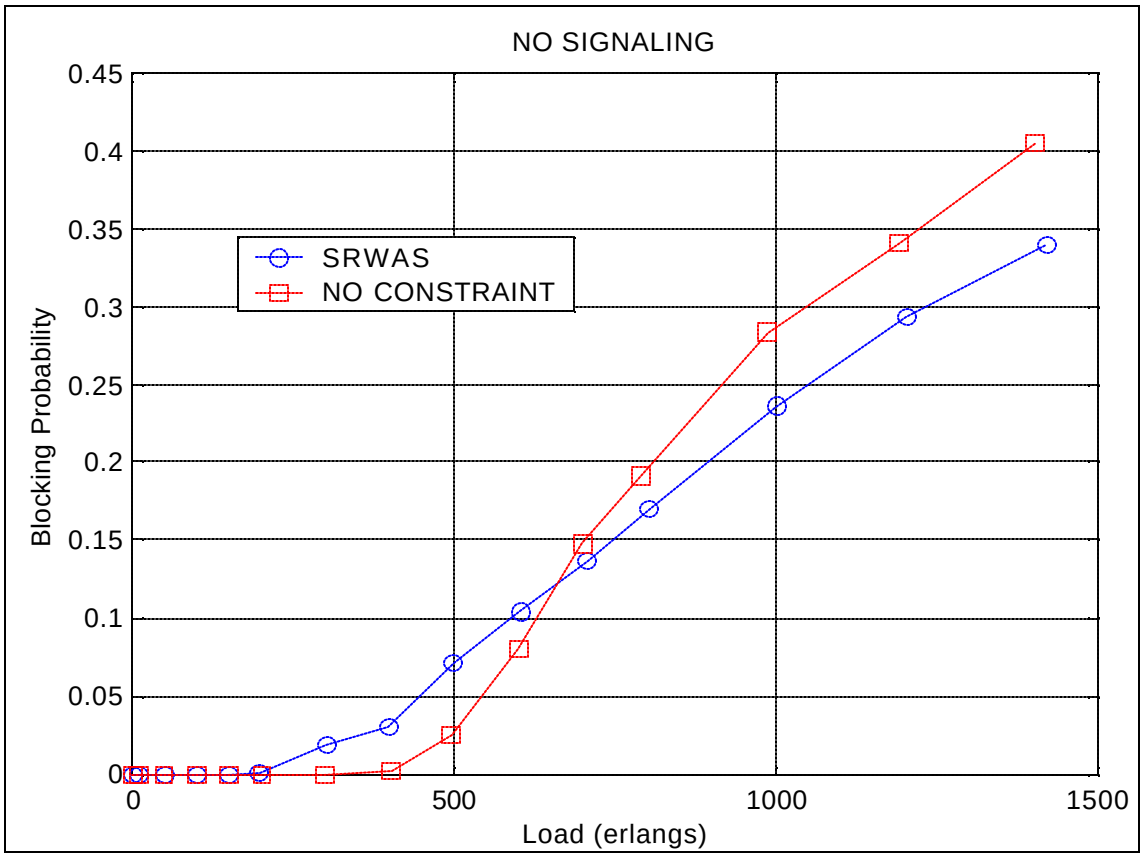

Figure 4.15: The effects of constraining the logical topology construction

4.4.3.2 Future-Based (FB) Cost Function

A Future-Based (FB) cost function is one that makes a link very expensive (it takes the cost of A_∞, where $1 << A_\infty < \infty$) when using it now renders it unavailable in the future. Then, for every call arrival a novel integrated link cost function that is based on the arrival of future calls is proposed as follows.

$$C_{PL_i}^{FB} = \begin{cases} \dfrac{1}{f_{PL_i}^{available} - 1}, & \text{if } f_{PL_i}^{available} > 1 \\ A_\infty, & \text{if } f_{PL_i}^{available} = 1 \\ \infty, & \text{if } f_{PL_i}^{available} = 0 \end{cases}$$

For the LL Cost, $C_{LL_i}^{FB}$, define the following:

T_{BW}^{mean} the traffic's mean (in number of BGs)

T_{BW}^{sdev} the traffic's standard deviation (in number of BGs)

Then,

$$C_{LL_i}^{FB} = \begin{cases} \dfrac{N_{PLs}^{LL_i}}{(BW_{LL_i}^{available} - CR_{BW}) - (T_{BW}^{mean} + T_{BW}^{sdev})}, & \text{if } BW_{LL_i}^{available} - CR_{BW} > T_{BW}^{mean} + T_{BW}^{sdev} \\ \dfrac{A_\infty \times N_{LL_i}^{PLs}}{BW_{LL_i}^{available} - CR_{BW} + 1}, & \text{if } BW_{LL_i}^{available} - CR_{BW} \le T_{BW}^{mean} + T_{BW}^{sdev} \end{cases}$$

If LP_{RWA}^{j} is the j_{th} lightpath that needs to be built for this call, then the total path cost is given by:

$$C_{path} = \sum_{LP_{RWA}^{j} \in path} \sum_{PL_i \in LP_{RWA}^{j}} C_{PL_i} + \sum_{LL_i \in path} C_{LL_i} .$$

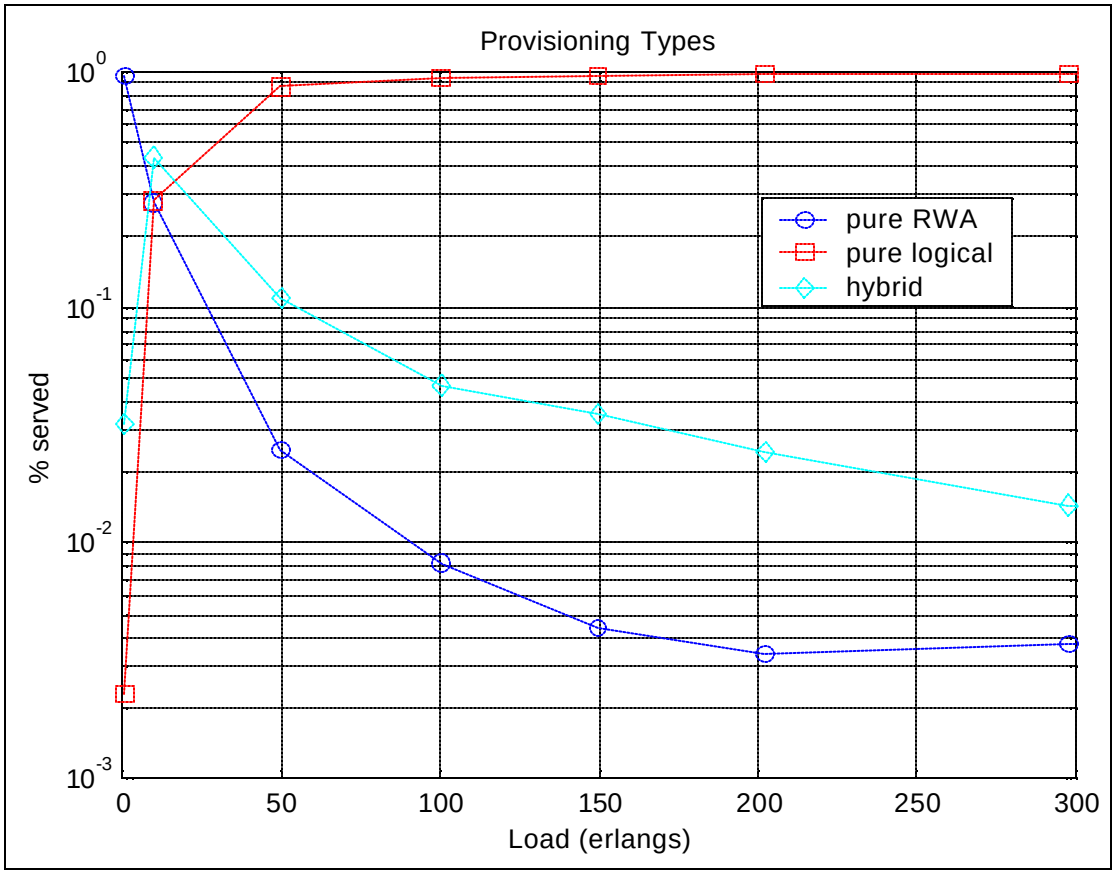

Figure 4.16: Types of successfully provisioned paths

For the evaluation of the integrated multi-fiber cost functions, the NSF16 network was used with 16 nodes and 50 unidirectional links and represents a backbone US network (Figure 4.13). Each unidirectional link is assumed to have 2 fibers. There are 4 wavelength channels on each link. The simulations used a dynamic traffic model in

which call requests arrive at each node according to a Poisson process (one every 5 minutes) and their durations follow an exponential distribution, which varies to reflect different network loads. The base granularity (BG) was assumed to be approximately STS-1 (50Mbps) and call demands were normally distributed around 400Mbps (8 BGs) with a standard deviation of 200Mbps (4 BGs).

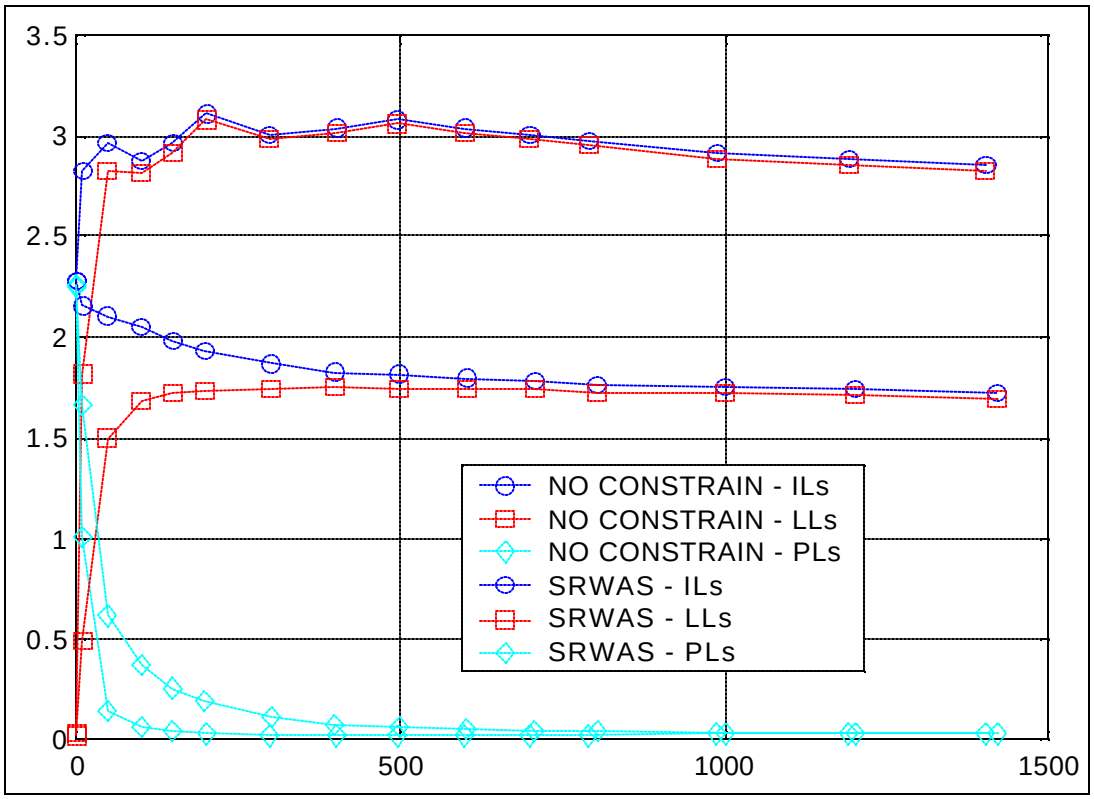

Figure 4.17: Average number of Links per successful call servicing vs. network load

Figure 4.14 presents the comparison of the two integrated cost functions (LU and FB), where network blocking is the performance metric and the signaling component is disabled. It is observed that at low loads, FB performs better but as loads increase the two cost functions have the same performance.

Figure 4.15 presents the effects in blocking probability of constraining the integrated graph construction using the SRWAS approach with the signaling component disabled (both under the FB cost assignment). It is clear that performance is heavily affected by the operation load regime. At low loads, NC perform noticeably better but this is reversed as the network is operated at higher loads. This is because using more segments to serve a connection at low loads does not tie up network resources for long times; however the burden in doing so becomes an issue as the holding times increase. The conclusion drawn is that longer paths degrade blocking probability performance only at high loads; in our experiments, degradation starts at 500 erlangs, and performance gain is reversed at around 700 erlangs.

In figure 4.17, the average number of ILs (PLs and LLs) per path are shown The results show that at 500 erlangs, the NC ILs per path adopt a negative slope which implies that reducing the path length is actually beneficial (consistent with the results of figure 4.15). From the same graph it is concluded that at higher loads a path is almost completely made up of LLs, as PLs are highly unlikely to be available to service a new connection. This is also confirmed in figure 4.16 where the different ways calls were provisioned are plotted against load under the NC approach. At extremely low loads ($\leq$10 erlangs) almost all calls are purely RWA provisioned because LLs are deleted from the integrated topology soon after their creation diminishing the probability of being reused. However, as loads increase the integrated topology is basically made up of only LLs (PL exhaustion takes place) and thus purely logically provisioning prevails.

Chapter 5

AN INTEGRATED SIGNALING COMPONENT

5.1 Introduction

Provisioning of connections requires algorithms for path selection, and signaling mechanisms to request and establish connectivity within the network along a chosen path. This chapter deals with the signaling component. Two types of distributed signaling protocols have been investigated in the literature, namely forward and backward reservation signaling protocols [5-7]. In the forward scheme, reservations are made in the forward direction where the source node sends a reservation message towards the destination along the selected path. In the backward scheme, the source node sends a probe message to the destination without reserving any resources. This message will collect information about resource availability along the path. Upon receiving the probe

message, the destination node then sends a reservation message back to the source node along the same path, where each intermediate node starts reserving the appropriate network resources. To conform to GMPLS-based signaling standards being proposed by IETF, where Resource ReSerVation Protocol with Traffic Engineering extensions (RSVP-TE) has been selected as the signaling protocol for optical networks, this work adopts the backward reservation scheme.

5.2 The Proposed Integrated Signaling Protocol

In this section, a fully distributed global information-based signaling protocol for real-time provisioning of diverse traffic granularity (on a per-call basis including both full-lambda and sub-lambda traffic flows) entirely on the optical layer's terms is devised. The main characteristic of the proposed signaling scheme are:

1) the integrated routing approach supports a single instance of integrated routing and signaling protocols running solely across the optical layer;
2) it is a fully adaptive scheme that uses the combined topology and resource usage information at the IP and optical layers to select an explicit path;
3) it follows a forward-probing, backward reserving scheme to conform with GMPLS-based signaling standards;
4) it is fully distributed without the need for synchronization;
5) to the best of our knowledge, this is the first integrated signaling scheme that can simultaneously setup and teardown both full lambda (lightpaths) and sub-lambda (LSPs) connection requests at the optical layer using a unified signaling scheme.

A Data Communications Network (DCN) is assumed to serve as the packet transport network for all the signaling and network state updates messages. The DCN is assumed to be an exact replicate of the physical topology and a control channel operating at a lower speed than data channels is reserved for signaling (i.e. a supervisory channel).

5.2.1. Protocol Description

In the proposed link-state approach, each optical node must maintain the complete network state information, including the combined topology and resource usage information at the IP and optical layers. Nodes exchange such information with each other through Link-State Advertisements (LSAs). All the nodes must be informed, when the state of the network changes.

As mentioned above, a forward-probing and backward-reserving scheme is used to setup a connection. Upon receiving a connection request, the source node performs explicit source routing, i.e. the source node selects the full route and assigns a wavelength (in the case of setting up new lightpath). In the case of multi-fiber environment, as it is assumed here, if the assigned wavelength is available on more than one outgoing fiber owned by an intermediate node along the path, the decision on which fiber to select is decided locally by that node.

The source node then sends a PROBE message, which contains the explicit path, towards the destination along the selected path. As the PROBE message travels downstream through the intermediate nodes along the path, it simply probes the involved nodes on the

availability of their resources that the path requires. No reservation is taking place during the probing. If the PROBE message is processed at all intermediate nodes successfully and reaches the destination, the latter reserves (allocates) its resources and sends an ACK message[3] upstream towards the source node. An ACK message is actually a reservation message, which means that if a node processing it finds its resources in question still available, it will allocate them to this request (rendering them unavailable for further PROBES or ACKs of other call requests). When an ACK reaches the source of the call, it too will check if its resources are still available and, if so, will admit the call in the network.

If a node processing a PROBE finds the resources in question unavailable, it will stop the signaling by sending a NACK in the upstream direction, to the source. If a node processing the ACK message finds the resources in question unavailable, it will send two messages: a NACK upstream to the source informing it of the failed signaling, and a downstream RELEASE message to the destination informing the downstream nodes, that have already reserved resources for this request, to release them.

While the implementation of the above generic signaling scheme is rather straightforward when independently provisioning lightpaths at the optical layer or LSPs at the logical layer (i.e. when a sequential provisioning approach is followed), it becomes more challenging and complicated when an integrated approach (i.e. signaling for both lightpaths and LSPs at the optical layer) is assumed. This is because an integrated routing might return a hybrid path, where a call might need to [possibly remotely] setup new

3 Only if the last link in the path is a PL. If the last link is an LL, the ACK generation occurs a node before the call destination, which is the source of the last LL (explained later, assumption 4 in 5.2.2).

lightpaths in addition to reserving bandwidth (LSP) on existing lightpaths, in which case the probed/reserved resources at the involved nodes are not always of the same type (i.e. logical or physical), but rather a function of the explicit path. One simple approach to resolve this problem is to use two independent generic PROBE and reservation signaling messages, one for setting up a new lightpath and the other one for setting up LSP (sub-lambda request). However, the choice to follow a truly integrated signaling and, thus, avoid the introduction of new signaling messages, while still retaining a single forward-probing backward reserving message, invoke the need to introduce further intelligence in the way signaling messages are formatted and processed.

5.2.2 Signaling under an Integrated Approach; Innovations and Challenges

First, it should be emphasized that most of the MPLS/GMPLS-based routing and signaling algorithms and protocols which have been reported by standards bodies as well as research communities, were developed to provision either full wavelength channels (lightpaths) at the optical layer only (find a route and assign a wavelength) or packets/LSPs at the IP/MPLS layer only (logical layer). Note that each layer supports its own suit of routing and signaling protocols and that each suite is totally independent of the other. While a number of works in the literature have dealt with the initiative of a unified control plane (peer interconnection model proposed by IETF) to integrate the optical and IP/MPLS layers into a single administrative domain that runs a single instance of routing and signaling protocols [1-4], the authors are not aware of any work

that has addressed the specifics of implementing integrated signaling protocols that can simultaneously provision both full lambda (lightpaths) and sub-lambda (LSPs) connection requests at a single domain in a unified manner.

To devise an integrated signaling protocol that can simultaneously provision both full lambda (new lightpath) and sub-lambda (LSP) connection requests at the optical layer using a single unified probe and reservation messages, the following innovations and assumptions are introduced:

1. In contrast to conventional provisioning of full wavelength channels at the optical layer where signaling messages are processed at each node along the entire path (on a hop-by-hop basis), signaling messages for sub-lambda connections are only processed at the source of the lightpath (LP).

2. In contrast to conventional provisioning of full wavelength channels at the optical layer where each node along the path owns its local resources including its outgoing physical links (PLs), the entire resources along the path is assumed here to be owned by the source that setup the LP. Thus, the source node of an existing LP owns all the logical link (LL) resources, e.g., all outgoing physical links, wavelength channel residual capacity, and ports comprising the entire lightpath. By owning, it is meant that the "owner node" allocates/de-allocates the call resources (a full lambda and/or sub-lambda), has the most updated status information about these links (physical and/or logical), and that any changes along the path can be advertised and updated only by the owner node. This implies that when a new LP is setup, the ownership of the entire LP resources,

(i.e. local physical resources of each intermediate node along the path), is transferred to the source of the new LP, and when a LP is torn down, the ownership of these physical resources is then transferred from the source of the torn-down LP to the intermediate nodes that originally owned them.

3. If a signaling message is to traverse a logical link (single LP), the message is assumed here to be forwarded along the static shortest path between the end points of the lightpath and need not follow the exact data path (out of band signaling). This means that intermediate nodes along the shortest path of an LP receiving a signaling message only *forward* it further without processing it since they are not involved in the signaling process. In addition, since NACK messages impose no intelligence in processing at intermediate nodes, they, too, are assumed to be forwarded instead of processed until they reach the source of the call.

4. If an LSP is to be provisioned over a path whose last link is an exiting LP, the PROBE message need not traverse the entire path; it is forwarded only up to the source of the last LP. This implies that a connection request provisioned over a single existing lightpath requires no signaling, as the source of the call owns the resources the call setup requires.

5. When a node receives a PROBE message, it needs to identify the probed resources. This is achieved by examining its previous (upstream) and next (downstream, if applicable) links in the path contained in the PROBE message.

6. Due to the added complexity and cost it would introduce, global synchronization among network nodes when exchanging LSAs is not assumed. In other words,

time-stamping the control messages is not necessary for deadlock-free protocol operation.

It is important to emphasize that all the assumptions and ideas presented in this work are the result of running initially extensive simulation experiments that revealed the challenges, deadlocks, bugs and complexity of the signaling protocol.

As an illustrative example, Figure 5.1 shows how signaling messages are processed for four different cases. Assume that *SP(a, b)* denotes the static shortest path (short, in terms of distance) between nodes *a* and *b*.

Figure 5.1 (a) shows that signaling messages are processed at each node (on a hop-by-hop basis) when a connection is being provisioned over a single LP to be setup (purely RWA) along path S-1-2-D. Figure 5.1 (b) shows that no signaling is required when a sub-lambda connection (LSP) is being provisioned over a single existing LP (purely logical). Figure 5.1 (c) shows that signaling messages are processed only at the sources of LPs when a sub-lambda connection (LSP) is purely logically provisioned over two existing LPs (LL_1 and LL_2). Note that the PROBE message reaches only up to the source of LL_2, where an ACK is generated, and that signaling over LL_1 is performed over *SP(S, 3)* (shown with one intermediate unmarked node), rather than the lightpath route. Finally, Figure 5.1 (d) considers the case of hybrid provisioning where a call is being serviced over three segments from S to D; the first segment is an LP to be setup with source S and destination 2 over PL_1; the second segment 2-3-4-5 (LL_1) is an existing LP with source 1 and destination 5; the third segment is another LP to be setup with source 5 and destination D, over PL_2, PL_3, and PL_4. Note that the PROBE message is processed only at

nodes S, 1, 5, 6, 7 and D, and that the LL_1 segment is signaled over *SP(1,4)* (shown with one intermediate unmarked node, which simply forwards, rather than processes, the messages).

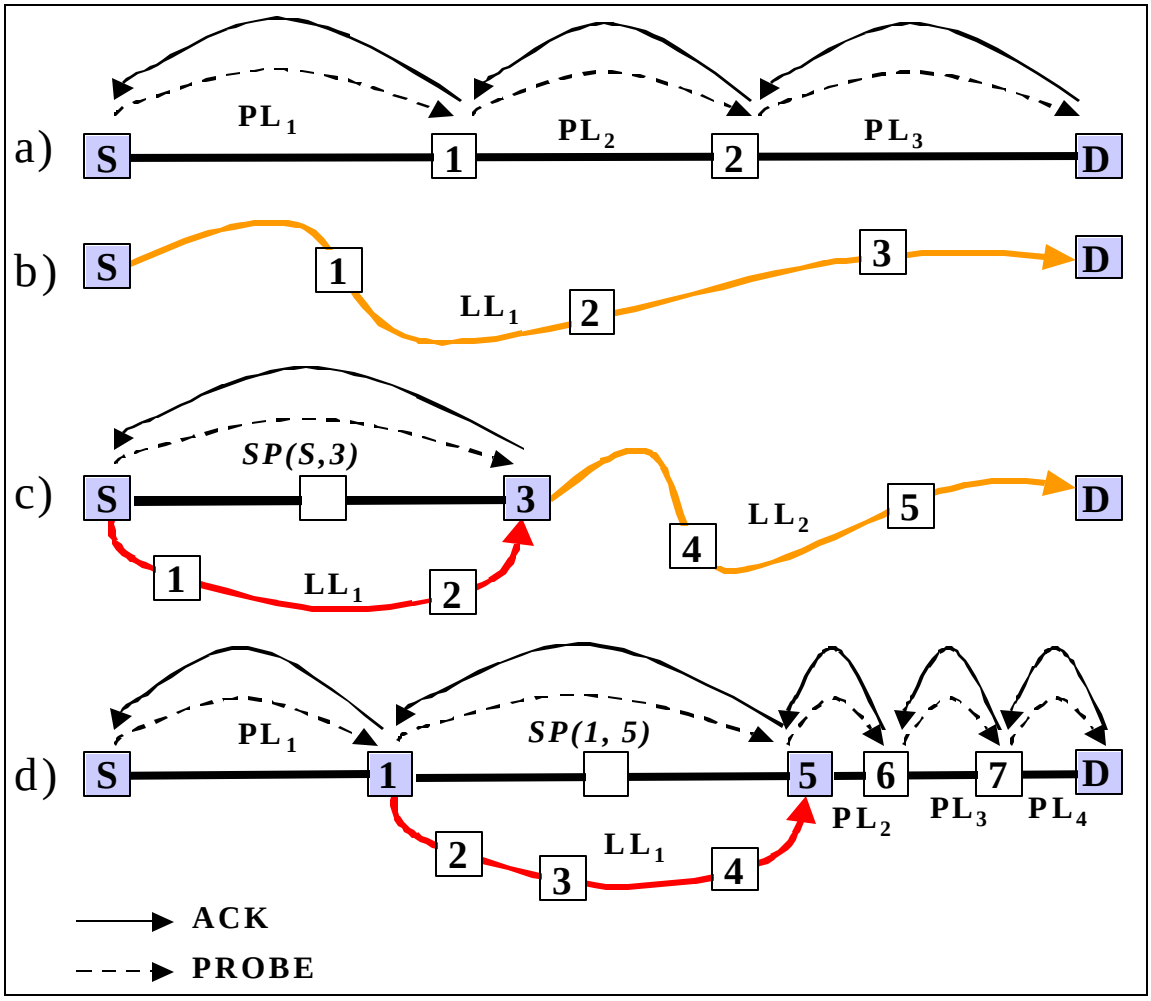

Figure 5.1: Provisioning Signaling under a) purely RWA, b) purely logical single hop, c) purely logical multi-hop, and d) hybrid scenarios

As mentioned above, when a node receives a PROBE message, it identifies the probed resources by examining the previous (upstream) and next (downstream) links. For instance, as shown in Figure 5.1 (d), when node 5 receives the PROBE message, since the path downstream link is a PL (PL_2) and the upstream link is a LL (LL_1), it checks on

the availability of both PL_2 and a transmitting port, as this node will be the source of a newly created lightpath to serve the call. On the other hand, when node 1 receives the PROBE message, since the path downstream link is a LL (LL_1) and the upstream link is a PL (PL_1), it will check on both the availability of LL_1 and a receiving port, as it will be the destination of a lightpath to be created to serve the call.

5.2.3 Connection Release

When a call is terminated, the network frees up the resources allocated to the call, i.e. a dynamic logical topology is considered. Since sub-lambda requests are assumed, the LPs are torn down when the last call using them departs the network. The signaling for a call release may consists of a single message, RELEASE_BW, or possibly two messages, RELEASE_BW and TEAR_DOWN. The RELEASE_BW message contains the call ID. A node processing (or generating, in the case it is the call's ingress) such a message de-allocates the resources (processing speed at the electronic switch and BW along the outgoing LP, etc.) the call was using and then propagates the message downstream to the next LP's source until it reaches the source of the last LP along the path. The source nodes then update the status of their lightpaths by increasing their residual bandwidth by an amount that corresponds the call's released bandwidth. RELEASE_BW message is transmitted (over the shortest path) to, and processed at all the sources of the involved LPs.

In addition, at each of these LP sources, the need for a TEAR_DOWN message transmission is also examined. Thus, the node always checks the LP's residual capacity (after releasing the call's bandwidth). If the residual capacity is equal to the full wavelength channel capacity (last call to use the LP), the node releases the hardware (e.g. optical switch ports) associated with the LP and then generates a TEAR_DOWN message that propagates downstream to all nodes comprising the LP to be torn down over the actual data path (LP) until the LP's egress is reached.

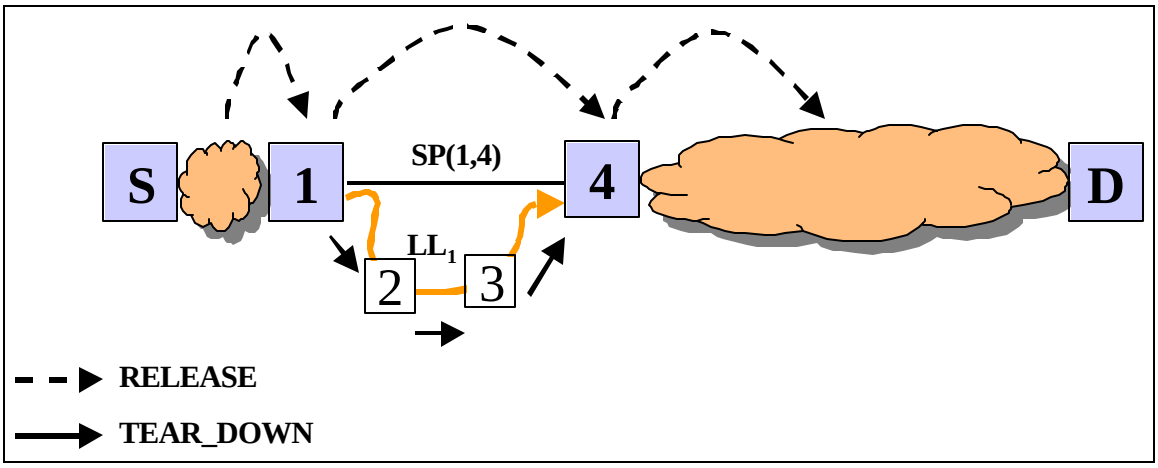

Figure 5.2: Call Release

As an illustrative example, assume that node 1 receives a RELEASE message from the upstream node, as shown in Figure 5.2. It first processes the message and then forwards it downstream (over *SP (1,4)*) and at the same time checks the LL_1 (existing lightpath 1) residual bandwidth. If that is equal to the total wavelength channel capacity, it means that no more calls are using LL_1 and so it can be torn down. Thus, node 1 sends a TEAR _ DOWN message over the data path (1-2-3-4). The nodes initiating the TEAR_DOWN signaling always release a transmitting port and the egress nodes of lightpaths receiving a TEAR_DOWN always release a receiving port.

5.3 Updating

5.3.1 Link-State Advertisements

In the proposed link-state approach, each optical node must maintain complete network state information, including combined topology and resource usage information at the IP and optical layers. Nodes exchange such information with each other in a distributed manner via LSAs messages. Whenever the state of the network changes, all the nodes must be informed. In other words, all nodes must maintain a synchronized and identical topology and link state information. Therefore, a change in the network status such as a call arrival/departure (setting-up/tearing-down one or more lightpaths/LSPs) may result in broadcasting of update messages to all nodes in the network. In this work an update can be triggered by a timer (timer-triggered) or by a change (change-triggered). Under the timer-triggered approach a node updates the network periodically whenever a local timer expires (unsynchronized with remote node timers). Under the change-triggered approach, the updates (changes created at the source node due to a change in the network status) are sent out to all other nodes whenever the number of changes reaches a predefined threshold.

To realize the updating part of the protocol, every node maintains a list of changes that occurred and involve its local resources (i.e. its outgoing physical and logical links). In this list, changes are added whenever they occur, provided that they don't cancel out or alter an old change already in the list. In other words, if a new change is to be added, it should not be relevant to an old change that exists in the list. If it is, as shown in Figure

5.3, one of two actions is taken: if the new change cancels out the old change, the new change is not added and the old change is removed (Figure 5.3, top); if the new change alters an existing old change, the new change is not added and the old change is replaced by one that reflects the final state after alteration (Figure 5.3, bottom).

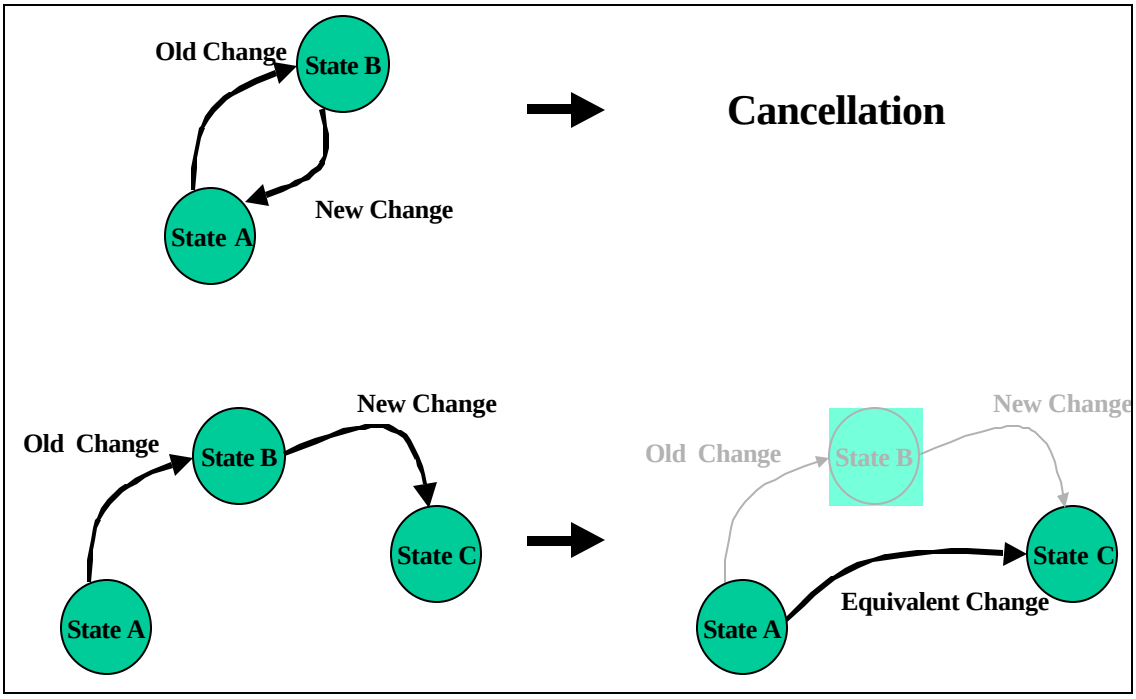

Figure 5.3: Adding a change about a lightpath for which there exists an unadvertised change

The LSA messages have a unique ID that is tied to the originator node (of local significance). When a node receives an LSA it first has to check if it already processed this LSA. This is done by comparing the received LSA's ID with a collection of received *<LSA_ID, originator>* combinations it keeps. If it is an already processed LSA, it is discarded. If, however, it is a fresh LSA, the node adds the *<LSA_ID, originator>* combination to the list of received LSAs and processes it.

5.3.2 Updating Challenges

The absence of global synchronization, the fully distributed scheme, and the two-level topological environment, introduce the challenge of how to properly update the network status information between the nodes. In fact, two major challenges face the implementation of an integrated updating strategy that is consistent with a fully integrated signaling protocol. From the outset, since dealing with sub-lambda connections at the optical layer, two-level topological updates information (physical an logical) is needed for exchanging such information among network nodes. However, if two independent updates are used (for LL, and PL availability) the updating scheme is not integrated, the protocol is prune to deadlocks, and extra redundancy is added.

To address the first challenge, a single-level updating scheme where the sole updating entity is the LL availability (i.e. lightpath status) is proposed. Information about the physical links can then be extracted from the lightpath state. The assumption is that physical resources in the network are not added or deleted as frequently as logical resources and can be thus considered constant. For instance, it is reasonably assumed that increases in the number of supported wavelengths on WDM links, or additions of extra fibers will occur very infrequently, allowing the network physical resources set to be considered fixed. What frequently changes in the status of an operational network is its logical resource availability, which is essentially a translation of resources from PLs to LLs and back. Thus, if **PRS(t)** and **LRS(t)** are the sets of Physical and Logical Resources as a function of time, the full view of the network status can be obtained by extracting the

PLs used by the LLs from the initial physical resource status (network status before any traffic, considered as fixed). Then:

$$PRS(t) = PRS(initial) - LRS(t)$$

The second challenge arises due to the absence of synchronization and message time-stamps. As mentioned above, when an existing lightpath is torn-down the ownership of its physical resources is transferred from its source to the intermediate nodes that originally owned them. Initial runs of the protocol, however, revealed an inherent deadlock under which different nodes updated the same resources differently to the rest of the network without any guarantee of the correct update to be sent last. Although infrequent, this took place when a lightpath was torn-down and before its deletion was advertised, another lightpath was setup originating from one of the torn-down lightpath's intermediate nodes using the same outgoing PL the torn down LP used. Then, the two sources (of the torn down LP and the newly setup LP) would update the intersection of their physical resources differently. We refer to this problem as the "*two ownership*" challenge.

To further illustrate this problem, consider the network in Figure 5.4 and assume that $t_0 < t_1 < t_2 < t_3 < t_4$. Suppose there is an existing LP_1 from 1 to 3 and that at t_0 it is starting to be torn down. At t_1, node 2 receives the TEAR_DOWN message and releases its OXC switch which makes its outgoing physical link (from 2 to 3, on the orange wavelength) available. Then, assume that node 2 receives a request to setup a second LP_2 to node 4 and at t_2 it completes the setup of a new LP_2 ($2 \rightarrow 3 \rightarrow 4$) on the orange wavelength (that is

available now) successfully. At t_3, node 2 advertises the existence of LP_2 and later, at t_4, node 1 advertises the tear down of LP_1. It is clear now that the rest of the nodes in the network will be misinformed about the status of link 2→3 (on the orange wavelength) as they will consider it available in their routing procedures.

To address the "two ownership" challenge, the following constraint needs to be imposed: unless a node is the original source of a torn-down lightpath (the node that initiates the tear-down action), it may not immediately use its released physical resources that become now available via a TEAR_DOWN message. Instead, it has to wait until the deletion update is received from the source of the torn-down lightpath (just like the rest of the network nodes). In the case that the torn lightpath was never advertised by its source to begin with, the TEAR_DOWN messages raise a special flag bit which entitles the intermediate nodes to consider their physical resources available immediately. Now, assumption 2 in section 5.2.2 needs to be modified to: "*…when a LP is torn down, the ownership of these physical resources is then transferred from the source of the torn-down LP to the intermediate nodes that originally owned them* ***only when these intermediate nodes have received: a) the tear-down update from the torn-down LP source, or b) a TEAR_DOWN message that explicitly allows the transfer***".

5.3.3 Contention and Cranckback

Since a source-based routing is assumed, a connection request can be blocked **f** the connection signaling fails en route. In this case, it is highly likely, especially at low

network loads, that the network has plenty of available resources that can accommodate the request, however not over the source-selected explicit route.

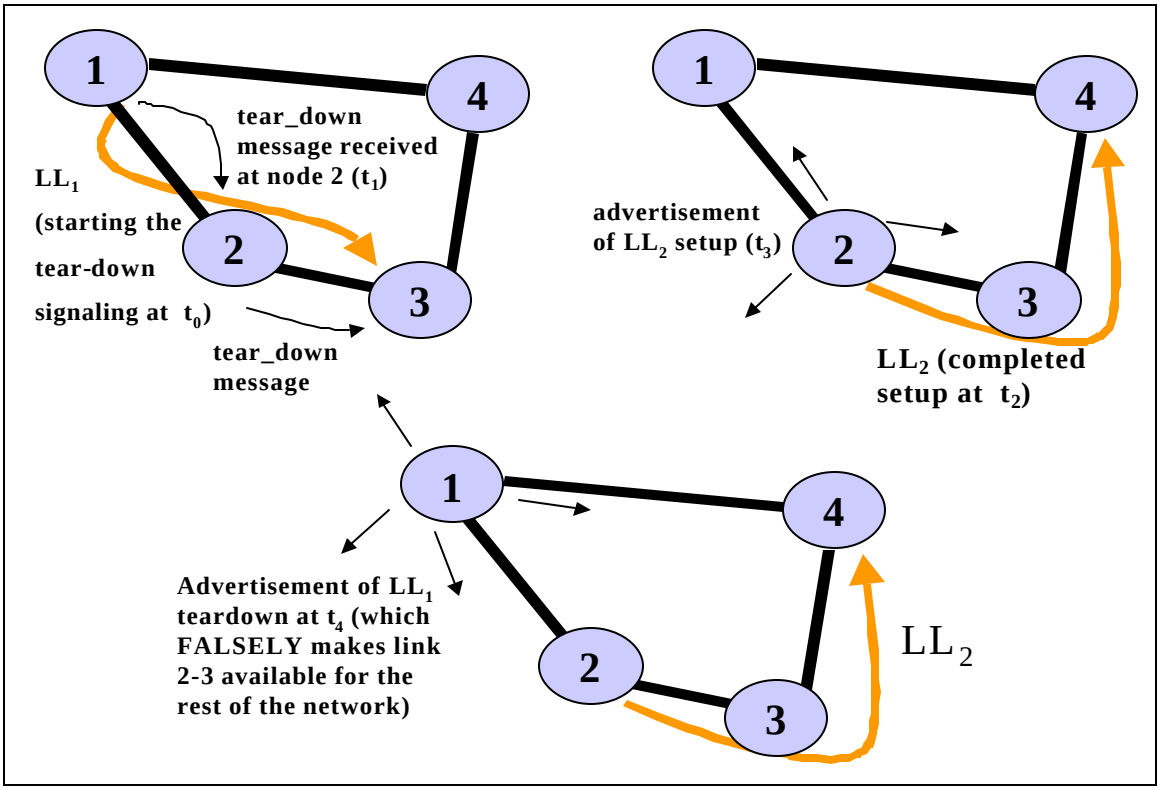

Figure 5.4: Updating deadlock

In the presence of an ideal centralized environment, all nodes are informed about the occurring changes instantly hence resource probing is not necessary. However, under a fully distributed and unsynchronized control plane, as this work assumes, the above no longer holds. Thus it is likely that different nodes will compete for the same remote network resources that they anticipate to be available based on their most recent views of the network status. This leads to contention as only the first node requesting the resources will finally have them allocated. The other nodes will simply experience provisioning signaling failures (e.g. through the reception of a NACK message), even if plenty of

network resources are available to serve the call over diverse paths. This is anticipated (also verified with simulation experiments) to be a frequent event at low loads, whenever the updates are less frequent than network changes. To address this issue, slight modifications are introduced to accommodate cranckback.

Provisioning signaling failures are realized when the source receives a NACK message or when the source processing the last ACK finds that it has allocated its resources to another request. In either case, the source identifies the node where signaling failed[4]. The node of failure together with the path, provide the source with the necessary information to identify the unavailable resources. With this information, the source reconstructs its integrated topology with the failed resources unavailable[5] and if a new path is returned, it commences signaling over the new path. The number of retries a source attempts for a specific call is a network variable (each retry will mark one extra path IL unavailable).

5.4 Connection Setup Delay

In this work the connection setup delay is measured between the time a call request arrives at the network and the time the network decides to accept it. It is considering the effects of connection request processing (including signaling initialization), round-trip

[4] When a node processing a PROBE or an ACK finds its resources in question unavailable, it is required to insert its node ID in a special err_node entry in the NACK message.

[5] We refer to this process as *pseudo-updating* as the unavailability information obtained from the failed signaling at the source of the call will only be used for routing this call and will not be inputted in the source's local network status tables. This is because the unavailable resources might not be advertised as unavailable at all, which means they will never be advertised as available (when they become) either.

propagation delay, signaling message transmission times, signaling message processing times, cross-connect (XC) execution times, and message/XC queuing delays (if applicable).

The local cross-connections execution time depends on the XC architecture which can be classified broadly into two classes [8-10]: (1) sequential cross-connects where a XC operation cannot begin until after the previous one has completed; (2) parallel cross-connects where each XC operation is executed immediately without any waiting. In this work sequential XC architecture (commercially available) is considered, where a XC operation is performed one at a time. This means that signaling message processing as well as XC execution is performed one at a time. Thus, signaling message buffering delays, as well as XC execution buffering delays must be taken into account.

The XC node architecture assumed here consists of an IP/MPLS-based intelligent controller and an all-optical XC fabric. The controller selects channels and initiates cross-connection requests that are executed by the XC fabric. Since a sequential XC architecture is assumed, the time involved in completing the request, t_{OXC}, consists of two components [8]:

1. the time the request waits before it gets serviced (assuming the XC fabric is busy executing other cross-connections, so the current request has to be buffered waiting for these cross-connections to complete), and

2. the time to perform a physical cross-connection (which depends on the technology used for the switch fabric).

5.4.1 Pipelining the Cross-Connect Operation

After receiving the setup message, the destination node first reserves local resources and then initiates the local cross-connect action. However, the destination does not wait for the local cross connect to complete; instead, it sends an ACK message (upstream towards the source) to the next node in the path. All intermediate nodes that receive the message, first check whether the appropriate cross connect is possible. If it is, similar to the destination node, they initiate the local cross-connection and at the same time forward the ACK message further upstream without waiting for the cross-connection to complete.

To completely pipeline the cross-connect operation, a node inserts an entry called *ready_time* in every ACK message, which is the time delay the node needs before it can satisfy the request. Figure 5.5, depicts the various delays for an ACK requesting an OXC operation at a node assuming that the ACK does not fail. Note that whenever a request is sent to the OXC switch, the latter responds with the time it will have the connection ready. The node, then, will not wait for the switch operation to complete before propagating the ACK further upstream. Instead, it will insert the next ACK *ready_time* to:

$$\max\{(prev\_ready\_time - ready\_time\_difference), local\_ready\_time\} - data\_prop,$$

where (refer to Figure 5.5) *prev_ready_time* is the *ready_time* entry of the received ACK, *ready_time_difference* is the time difference between two consecutive *ready_time* generations and *data_prop* is the propagation delay between the source of the call and

this node. With respect to Figure 5.5, note that previous ACK propagation and transmission times, as well as *data_prop* time are fixed, enabling this node to include them in calculating its own *ready_time* even in the absence of synchronization.

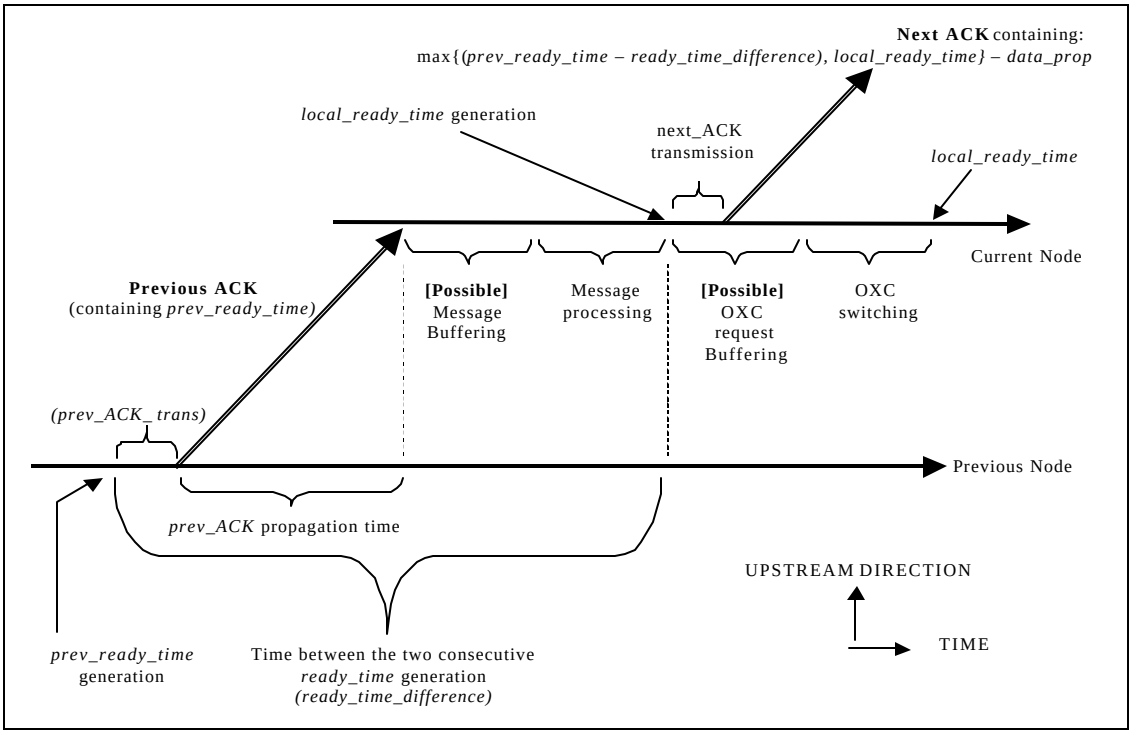

Figure 5.5: *ready_time* generation when successfully processing an ACK involving OXC switch operation

As an example, consider that the previous ACK contains a *prev_ready_time* of 100 (this means it will be ready to receive data in 100 arbitrary units time). Assume that ACK message transmission is 1 unit time, propagation from the previous node is 5 units time, and that previous ACK was buffered for 10 units and processed for 7 units time at this node. If this node is generated a *local_ready_time* of 10, and data need 25 units time to reach this node, then this node will insert a *ready_time* of $\max\{100-10-1-5-7,10\}-25=53$. In this case the previous node is delaying the

connection *ready_time.* As an alternative scenario, consider all the parameters are the same except *prev_ready_time* is now 7 units time. Then, this node will generate the new *ready_time* using $\max\{7 - 10 - 1 - 5 - 7, 10\} - 25 = \max\{-16, 10\}25 = -15$. In this case neither the previous nodes, nor this node delay the transmission because whatever OXC delay exists, it will be consumed by the time data will reach this node.

5.4.2 Incorporating all Delays

Suppose that a call experiences cranckback k times before it is served, or blocked at the source (due to unavailable resources, or because k is the maximum number of allowed cranckbacks). Then, the total setup delay for a signaled connection ($T_{connection}$) is given by:

$$T_{connection} = \sum_{i=1}^{k} (T^{i}_{forward} + T^{i}_{backward}),$$

where $T^{i}_{forward}$ and $T^{i}_{backward}$ are the downstream and upstream imposed delays on the i_{th} path. $T^{i}_{forward}$ is given by:

$$T^{i}_{forward} = T^{i}_{conn\_request\_proc} + t^{i}_{forward,RWA} + t^{i}_{forward,LOG},$$

where $T^{i}_{conn\_request\_proc}$ is the time the source of the call needs to generate the i_{th} path. Assume that the i_{th} path has K^{i}_{RWA} segments to build and uses K^{i}_{LOG} existing segments

(existing lightpaths)[6]. Furthermore, assume that N_j^{RWA} is the number of nodes in the j_{th} segment to be setup and N_j^{LOG} is the number of nodes in the shortest path between the j_{th} existing lightpath's source and destination. In both cases consider node 1 to be the [existing or to be setup] lightpath source. For the following equations, also consider t_a^b to be the time it takes node *b* to perform action *a* on a signaling message, and $t_{propagation}^{x,y}$ to be the propagation delay between nodes *x* and *y*. Then:

$$t_{forward,RWA}^{i} = \sum_{j=1}^{K_{RWA}^{i}} \sum_{k=1}^{N_j^{RWA}-1} (t_{propagation}^{k,k+1} + t_{processing}^{k+1} + t_{buffering}^{k+1} + t_{transmission}^{k}), \text{ and}$$

$$t_{forward,LOG}^{i} = \sum_{j=1}^{K_{LOG}^{i}} (t_{processing}^{N_j^{LOG}} + \sum_{k=1}^{N_j^{LOG}-1} (t_{propagation}^{k,k+1} + t_{buffering}^{k+1} + t_{transmission}^{k}) + \sum_{k=2}^{N_j^{LOG}-1} t_{forwarding}^{k}).$$

$T_{backward}^{i}$ is given by:

$$T_{backward}^{i} = t_{backward,RWA}^{i} + t_{backward,LOG}^{i} + T_{OXC}^{MAX},$$

where T_{OXC}^{MAX} is the *ready_time* the source of the lightpath to be setup segment generates (as per the aforementioned discussion),

$$t_{backward,RWA}^{i} = \sum_{j=K_{RWA}^{i}}^{1} \sum_{k=N_j^{RWA}}^{2} (t_{propagation}^{k,k-1} + t_{processing}^{k-1} + t_{buffering}^{k-1} + t_{transmission}^{k}), \text{ and}$$

[6] If the chosen path's last link is a LL, K_{LOG}^{i} excludes it, as signaling over it will not be performed.

$$t^{i}_{backward,LOG} = \sum_{j=K^{i}_{LOG}}^{1} (t^{N^{LOG}_{j}}_{processing} + \sum_{k=N^{LOG}_{j}}^{2} (t^{k,k-1}_{propagation} + t^{k-1}_{buffering} + t^{k}_{transmission}) + \sum_{k=N^{LOG}_{j}-1}^{2} t^{k}_{forwarding})$$

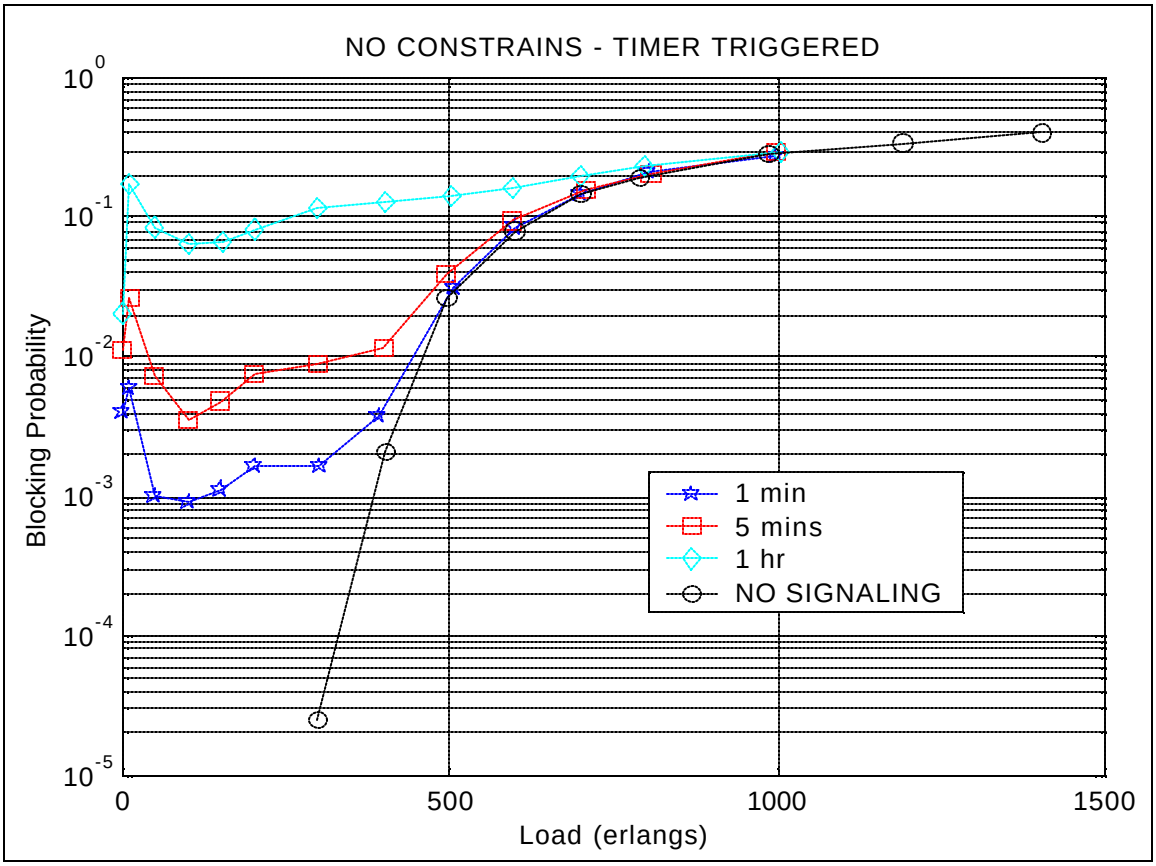

Figure 5.6: Blocking under timer-triggered updating scheme

5.5 Performance Evaluation

A custom-built C++ discrete-event simulator was constructed to conduct simulation experiments. The performances of the proposed integrated routing and signaling schemes were evaluated through the simulation of several network topologies that demonstrated similar conclusions. We present results for the 16-node, and 25 bi- [50 uni-] directional link NSF16 network representing a backbone US network (Figure 4.13). The core LSRs

are assumed to have enough interfaces and process all the traffic that can potentially pass through them. This assumption can be relaxed to account for the cases where routers have limited processing capabilities. All WDM links were accommodating 4 wavelength channels and each multi-fiber unidirectional physical link was assumed to contain two fibers.

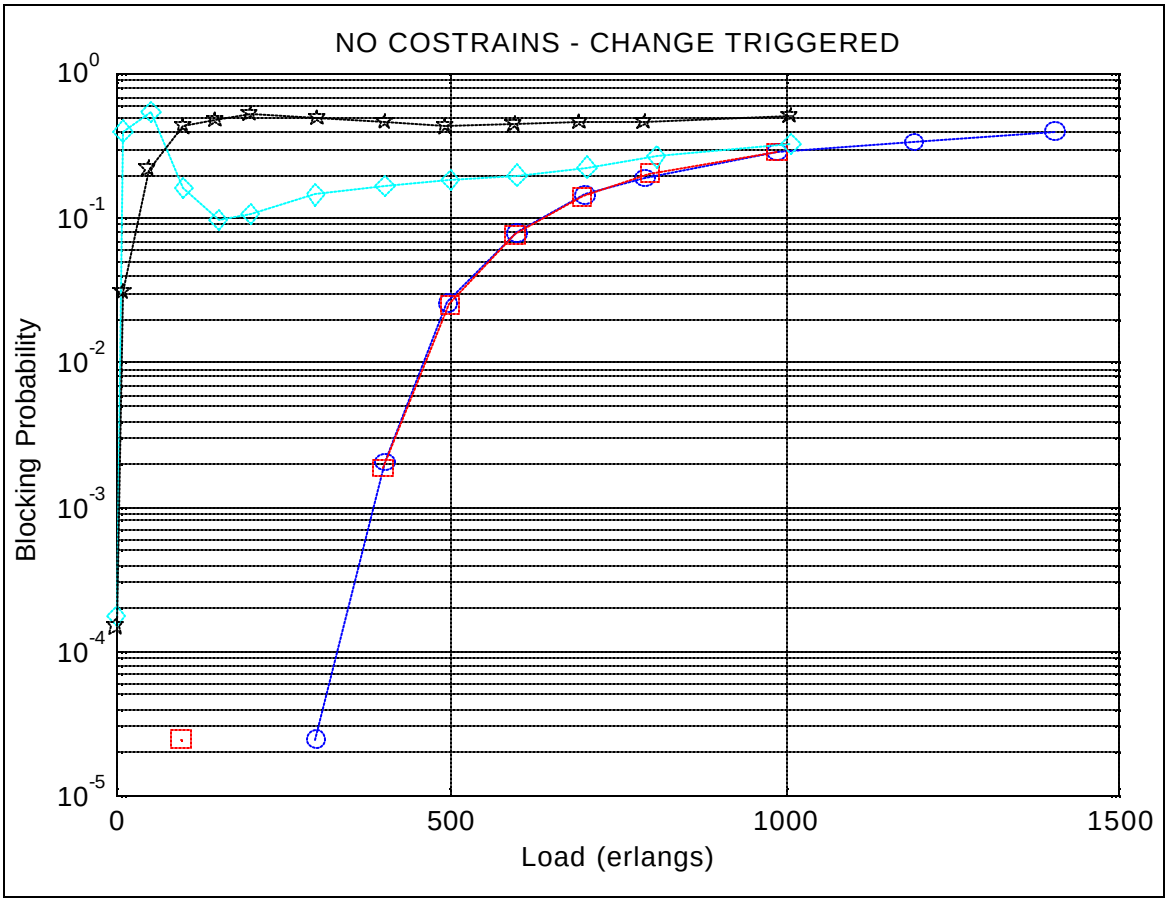

Figure 5.7: Blocking under the change-triggered updating scheme.

Control messages were assumed to have a fixed length of 100 bytes, and the out-of-band channel (DCN network) to have a speed line of 100Mbps. The wavelength channel capacity was fixed at 2.5 Gbps. The simulations used a dynamic traffic model in which call requests arrived at each node according to a Poisson process (one every 5 minutes)

and their durations followed an exponential distribution, which varied to reflect different network loads. The base granularity (BG) was assumed to be approximately STS-1 (50Mbps) and call demands were normally distributed around 400Mbps (8 BGs) with a standard deviation of 200Mbps (4 BGs). The control message processing time was fixed at 1ms, the message forwarding time fixed at 100μs, the request processing time at a source fixed at 5ms, and the CX cross-connection operation was assumed to take a fixed 15 ms. Results were obtained with the signaling component enabled as well as disabled. The latter implies that no signaling effects were taken into account as message-related time delays were all set to zero. This was performed in order to evaluate the cost functions, the integrated graph construction approaches and to present a clear indication of the signaling effects.

Figures 5.6 and 5.7 consider the different updating approaches under the NC approach. Figure 5.6 is presenting the blocking probability of the timer-triggered scheme when nodes update periodically every 1 minute, 5 minutes (equal to the arrival rate) and 1 hour,

and compares them with the no signaling case. The results present the effect of contention, which is most severe at low loads and when updates are less frequent. Figure 5.7 repeats the same comparison when the network updates are change-triggered, using thresholds at 1 change (squares), 4 changes (diamonds) and 10 changes (pentagrams). Again, contention is severely degrading network performance at low loads, whenever updating is not done frequently enough for nodes to obtain an accurate network status. The worst performance of figures 5.6 and 5.7 is exhibited by the case when updates are triggered when 10 changes are gathered in the list. This is because when setting the

updating threshold at 10 changes, there is hardly any updates at all (especially at lower lows), as new changes are most likely to cancel previous changes and thus reduce (instead of increase) the changes list size. This is in contrast with the best performance achieved when the nodes update for every single change. In fact, updating every single change performs as good as when the signaling component is disabled because nodes are always informed about the most accurate network status.

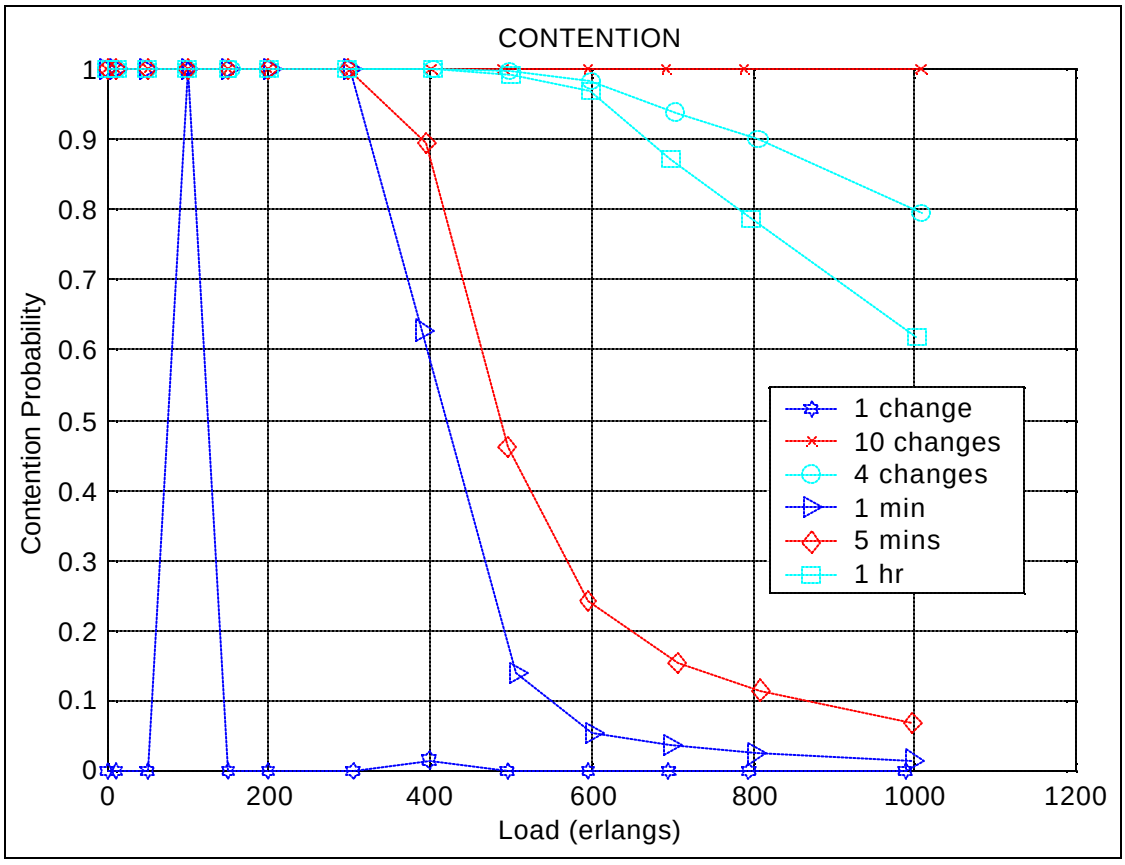

Figure 5.8: Contention Probabilities

To verify the contention observation, figure 5.8 presents the contention probability of a blocked call (if no blocking exists, the value is zero). It is clearly seen that blocking up to 300 erlangs is solely due to contention under any scheme.

The penalty of improved blocking performance in the previous Figures is depicted in Figure 5.9 where the number of exchanged messages to implement the various updating schemes is presented. It is clearly seen that whenever blocking probability is improved, it is because more messages were exchanged and the most recent network status has been conveyed to the network nodes via LSAs.

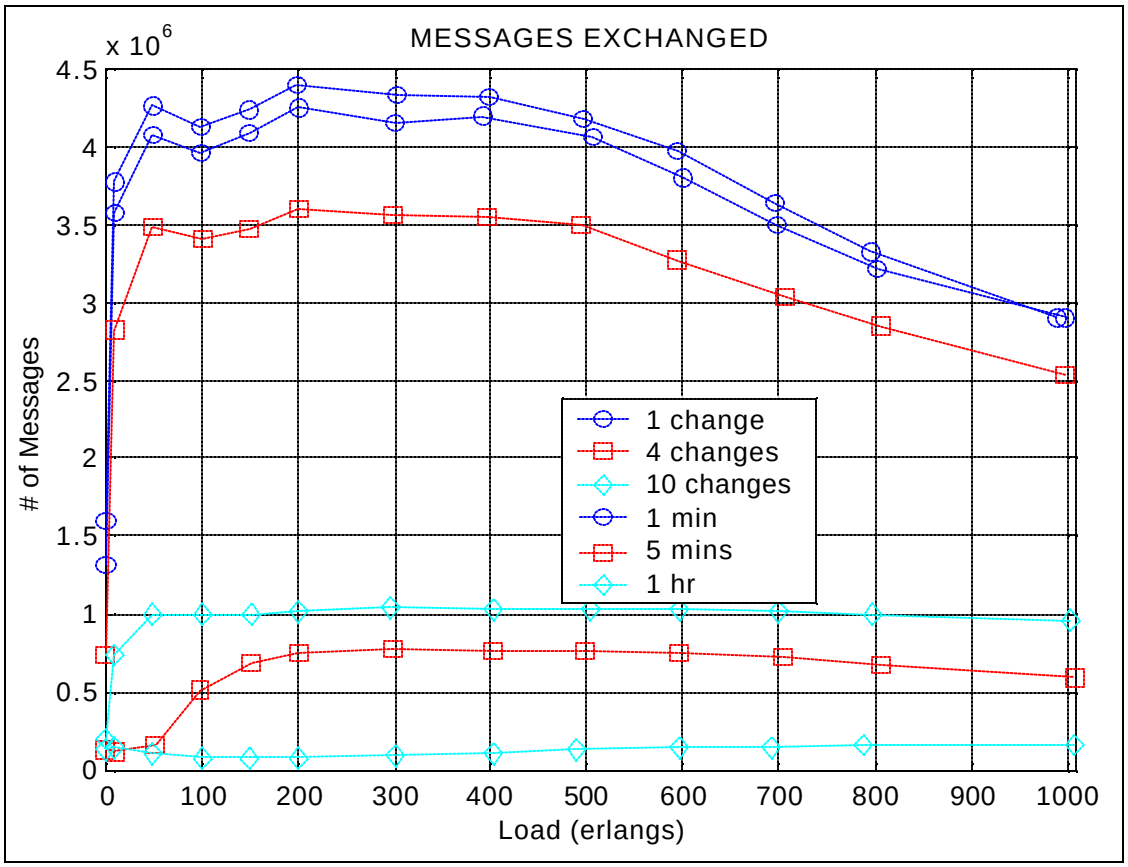

Figure 5.9: Control messages exchanged

Figure 5.10 presents the connection setup times as a function of the network load under the NC approach. Again an increased performance in blocking reflects to longer connection setup times. However, in contrast with the previous discussion, this penalty comes from provisioning longer paths, rather than updating more frequently. For

instance, under the change-triggered updating scheme with 10 changes, the successful paths are most likely to be comprised by a single IL because the view of remote network resources at any given node is severely outdated, which in fact leads to the lowest connection setup times. The connection setup times presented here span some tens of milliseconds and are heavily affected only by propagation delay (consider that the round trip optical signal propagation delay in the fiber between Princeton, NJ (node 14 in Figure 4.13) and Palo Alto, CA (node 1 in Figure 4.13) on NSF16 measures close to 50 ms). This might make one think that the complex discussion on connection setup times is rather obsolete. However, the presented delay analysis and equations are general, and can be used for network restoration, in which case all delay components will be critical (subject of future work).

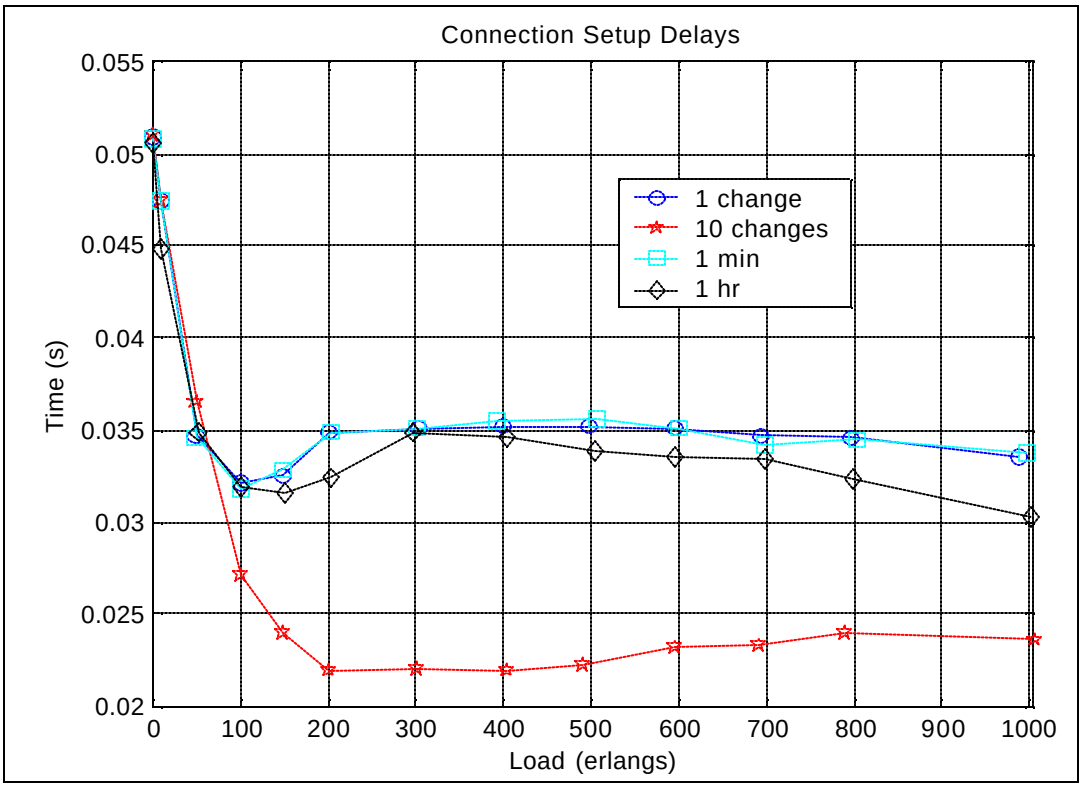

Figure 5.10: Connection setup times

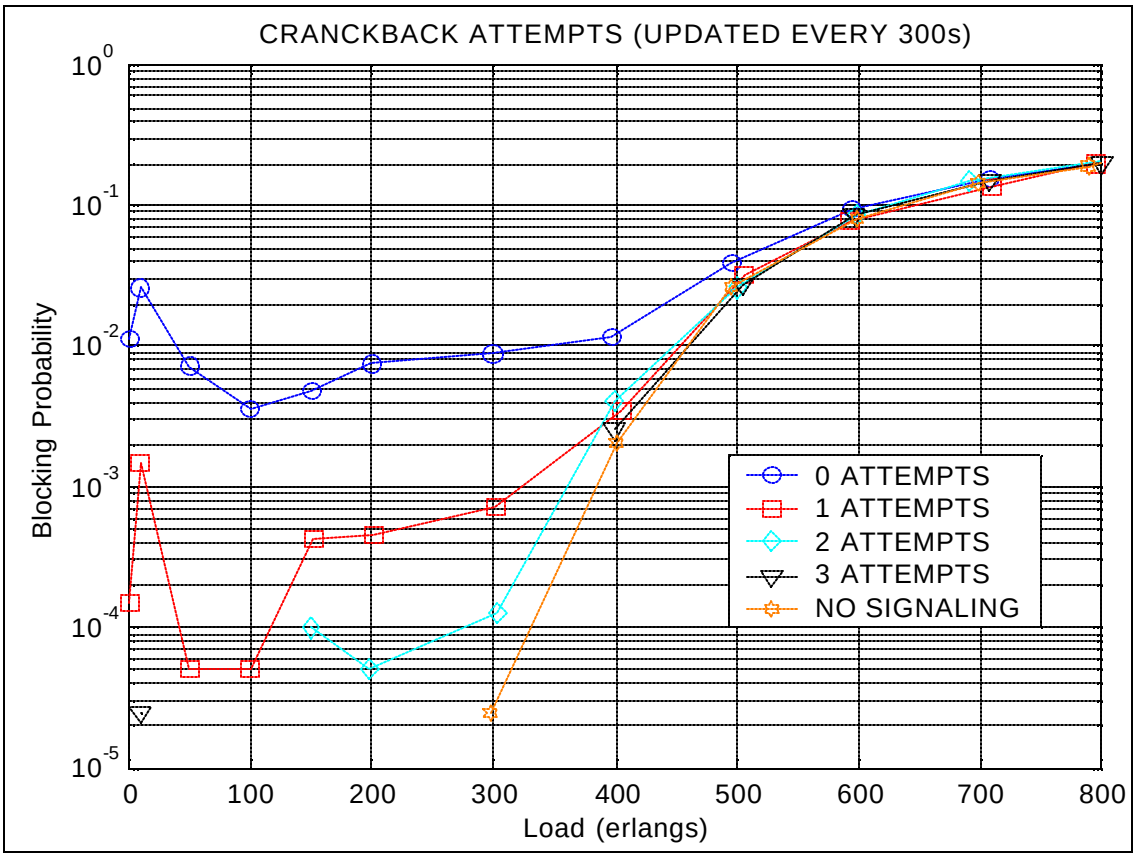

Figure 5.11: Improvement introduced with CranckBack (CB) attempts

Finally, Figure 5.11 presents the improvement that cranckback (CB) attempts introduce in terms of contention. The case of timer-triggered (every 5 mins) was chosen as it seems a realistic updating scheme in addition to having been heavily impacted by contention. Clearly, a single CB attempt, although improving contention, does not alleviate the problem completely. To achieve so, up to 3 CB attempts where required. The effects of CB attempts in average connection setup times and number of messages exchanged were not noticeable (graph omitted due to space limitations) as introducing cranckback only attacks the contented calls (a very small fraction of the serviced calls) and does not affect the updating scheme, which is the main contributor in the number of exchanged messages.

Chapter 6

PRACTICAL IMPLICATIONS OF THE PROPOSED MODEL

The proposed model has many practical implications many of which look well into the future and prove advantageous both in terms of efficiency and simplicity. In this chapter we present a brief list of them and stimulate preliminary research for future work. We start with the issue of restoration, where optics have always been favored due to rapid response (IP convergence times after failures can take minutes), but IP was always chosen due to efficiency. Then we briefly revisit the problem of grooming and show that the proposed model provides the means for using a literature-proposed complete and generic graph to attack this problem. Then we look well into the future and explore far-reaching solutions for both inter-domain routing and signaling, and an envisioned

solution for transporting Native Ethernet frames between end-users through an optical WDM-based core.

6.1 Selective Restoration on a Per-Call Basis

Using the proposed model, sub-wavelength traffic streams can now be restored at the optical layer [1]. The feasibility of an optical layer capable of provisioning/restoring on a per-call basis introduces the capability of selective restoration, which in turn allows for different levels of restoration (differentiated resilience) for different classes of service.

6.1.1 Edge Router Failure

When considering the overlay model where logical and physical layers are segregated whenever an edge client router fails IP would be alarmed to reroute around it. However, IP restoration times that are dictate by OSPF convergence, are dramatically high (in the order of minutes). In the proposed approach, the affected traffic *traversing* the failed router can be restored by rerouting the individual affected calls under fast optical restoration. In addition, all lightpaths originating or terminating at the failed router can be immediately released so that the resources can be made available for future connections.

As an illustrative example (Figure 6.1), if router E fails, the lightpath from A to E and the lightpath from E to F are released. If there were calls utilizing these two lightpaths sequentially (i.e. traffic from A to F, through E), they can be restored on a per-call basis.

For example, these calls can be serviced by setting-up a new lightpath directly from A to F (lightpath on the path A-C-E-F, not shown).

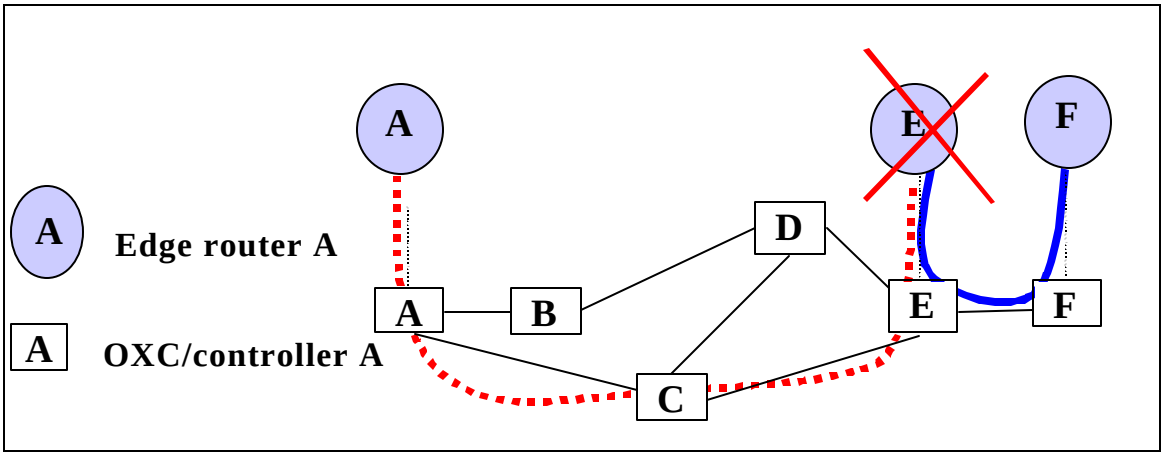

Figure 6.1: Edge Router Failure

6.1.2 Physical Link Failure (Trunk cut)

In the case of a trunk cut, all affected calls appear to the network as new calls; information about how these calls were previously serviced is erased and only the call characteristics are kept (i.e. the source, the destination, the bandwidth requirement etc). In this way, the affected calls can be individually re-provisioned logically (over existing lightpaths), or physically (setting up a new lightpath from the source (s) to the destination (s) of the call), or by using a hybrid (physical and logical) approach. Below we present a way to achieve this and compare the optical restorability of the proposed model (on a per-call basis) with the conventional one)

6.1.2.1 Conventional Lightpath Restoration

The affected lightpaths are stored in order of descending bandwidth allocated to them prior to the failure. The optical layer then tries to restore the lightpaths one-by-one, starting with the lightpath with the highest bandwidth allocation. In the case of a call that had multi-lightpath servicing, all lightpaths must be restored for the call to be restored. Failure to restore a lightpath is, consequently, a failure to restore all the calls previously traversing it.

6.1.2.2 Sub-Lambda Restoration

Each OXC controller is maintaining a list of all the calls and their characteristics associated with every trunk. After a trunk failure, the affected calls are grouped based on their corresponding (*S*-*D*) pairs. The different (*S*-*D*) pairs are stored in order of descending bandwidth. The algorithm then tries to restore the calls one-by-one starting from the ones that belong to the (*S*-*D*) pair with the highest bandwidth using the servicing order logical-hybrid-physical.

6.1.3 Performance Evaluation

The performance of the proposed approach is evaluated by simulating the NSF network consisting of 14 nodes and 21 bi-directional links (Figure 4.5). Adjacent nodes are connected through bi-directional physical links that consist of 4 fibers (two in each direction), where each fiber is assumed to have 4 wavelengths. The wavelength channel

capacity is assumed to be OC-48. The sub-lambda requests have bps demands that are normally distributed around 400 Mbps with a standard deviation of 200 Mbps, in multiples of 50Mbps. The edge routers are assumed to have enough interfaces and process all the traffic that can potentially pass through them. We further assume a dynamic logical topology; i.e. lightpaths are torn-down whenever the last call utilizing them departs the network.

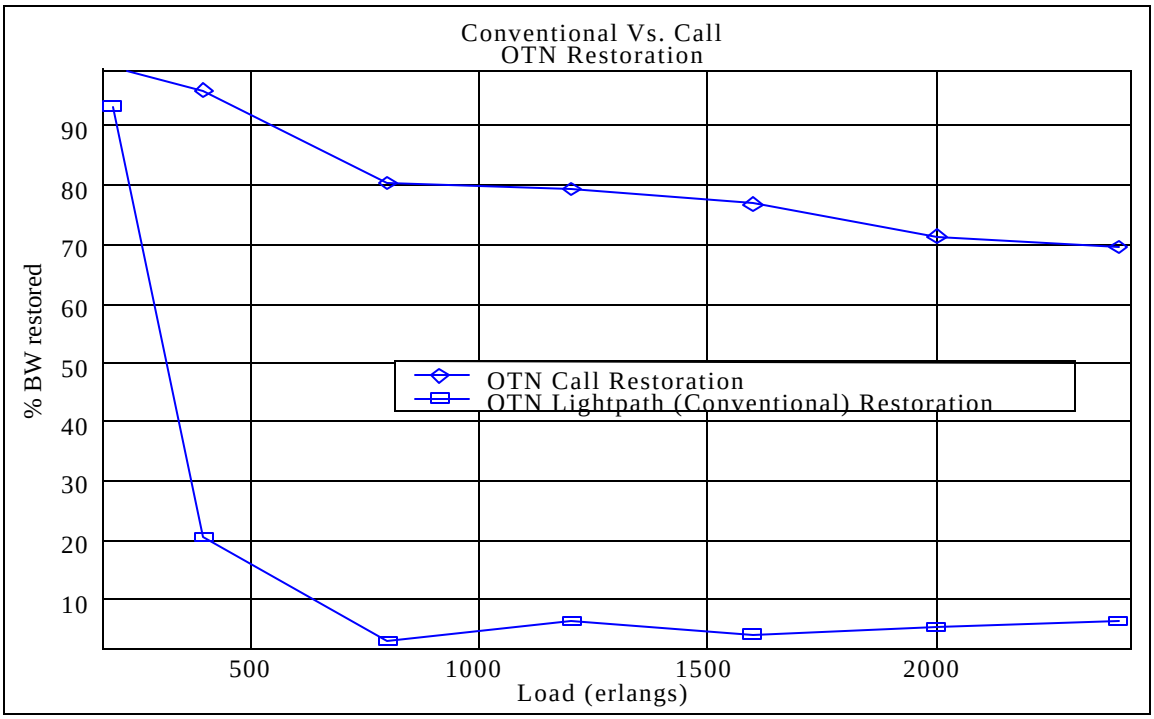

Figure 6.2: Conventional versus Per-Call Optical Restoration

Figure 6.2 compares the new optical layer-based sub-lambda restoration capability (in terms of the percentage of restored bandwidth) with that of the conventional path-based restoration (restoring full-lambdas). The fault is a trunk cut that fails all four fibers in the trunk. The fault is simulated after steady state network operation has been achieved. The results show that when the granularity of the restoration is on a per-call basis, much more

percentage of the bandwidth affected can be brought back into the network utilizing the remaining resources.

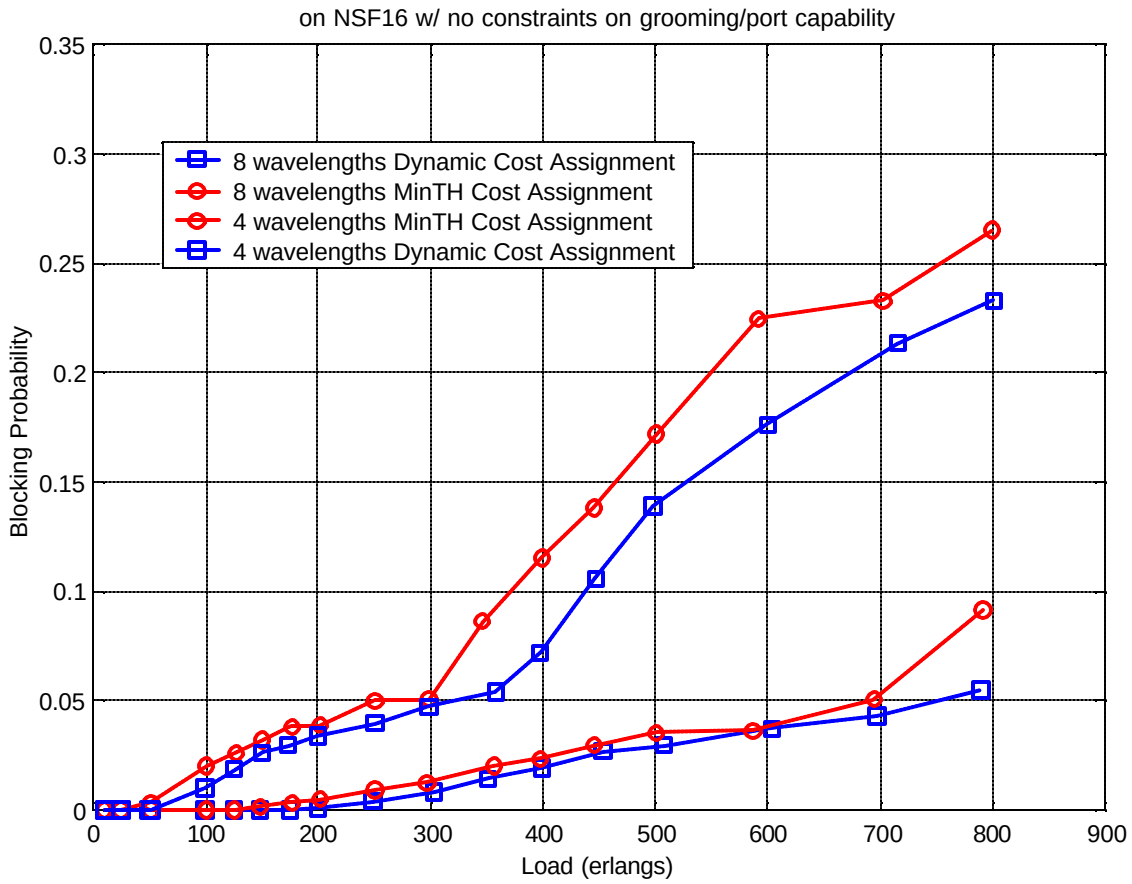

Figure 6.3: Using the Generic Graph to solve the traffic grooming in a really integrated fashion: Static versus Dynamic cost assignments

6.2 An Integrated Traffic Grooming Policy

As stated earlier, the grooming problem has received plenty of attention in the research community. Many of the approaches have dealt with a unified control plane where lightpaths and LSPs are treated as one [3-6]. However, there is no indication how to achieve such a scenario from an administrative point of view. The proposed model solves

this problem and provides the architecture to realize this scenario. In [2] the authors provided a generic graph which depicts all kinds of data streams as virtual links. For instance, wavelength translation, optical switching, electronic switching, opto-electronic and electro-optic conversions, grooming capabilities etc. can be "seen" as links with weighted costs. However, in [2] there have been static cost assignments, which, sadly, reduce the powerful graph to a sequential routing approach. Having in mind the model, we use this graph to assign dynamic costs. In Figure 6.3 we present the results obtained under the Minimum Traffic Hop (MinTH) cost assignment of [2], versus a dynamic approach where links are assigned costs based on their availability. The results were obtained on NSF16 with 4 and 8 wavelengths in place and a single fiber for every unidirectional link. No constraints on grooming or port capabilities were imposed.

6.3 End-to-End Online Inter-Domain Routing and Signaling

The strength of the proposed approach lays in its potential of opening up new avenues for implementing several significant novel applications that can drastically enhance and transform the vision of the next generation Internet. Two of the most important applications that can only be envisioned through the adaptation of the proposed interconnection model are:

1) The vision of implementing a scalable optical network - a network capable of being managed end-to-end (from the access network through the metro and core networks to another access network), across multiple Autonomous Systems (Ass).

2) The vision of an optical layer capable of restoring all disrupted traffic (both full-lambda and/or sub-lambda) independently in the case of a link/OXC and router failure. Thus, the current notion that the IP layer holds the upper hand in the debate about which layer should restore IP services may now be revisited and possibly altered.

For the most part, the Internet today is a set of autonomous, interconnected and inter-operating networks that ride on top of a physical telecommunications infrastructure that was largely designed to carry voice telephony. The Internet is a hierarchical structure with backbone providers at the highest level, interconnecting with other backbone providers private peering and public Network Access Points (NAPs). Connected below the backbone providers are enterprise networks and regional Internet Service Providers (ISPs). Local ISPs, that ultimately provide access to the end-user, connect to the regional ISPs. This interconnection of TCP/IP sub-networks is the Internet, an open "network of networks". This hierarchical structure leads to congestion on the Internet at points where the sub-networks must interconnect. The decentralized nature of the Internet means that there is only limited coordination between network providers, a fact that exacerbates the problems of network performance.

These problems are further manifested by the so called "hot potato routing", where backbone providers place traffic destined for another backbone as soon as possible at the nearest traffic exchange points, usually a public NAP [7]. This creates traffic flows that are asymmetrical and causes congestion at the public NAPs, resulting in performance degradation. This practice also limits the level of control a provider has over end-to-end

service quality. In response to this, ISPs have established private NAPs (P-NAPs) to ensure that traffic is handed off to other carrier's networks with fewer problems. P-NAPs allow the provider to establish control over paths. Note, however, that P-NAPs have their own limitations and can't provide real end-to-end quality of service particularly for real-time streaming applications that require bounded delay and jitter and consequently real-time connection oriented dynamic provisioning.

Compounding the problem, is the fact that most of the GMPLS-based routing and signaling algorithms cited above [8-11], which are currently being developed to address the problem of real-time provisioning of IP-over-optical networks, were developed to provision connection requests at the full wavelength capacity and run only within the boundaries of a single AS "the OTN". Furthermore, recent work carried out by the Optical Domain Service Interconnect (ODSI) task force in order to provision an end-to-end optical path across multiple sub-networks where communication between these sub-networks is accomplished via the introduction of a network-to-network interface (NNI) standard, has not materialized.

Consequently, the notion of supporting end-to-end QoS (provisioning end-to-end connection requests) is little more than cleverly disguised marketing. What is needed is to develop a new strategy for the migration from the conventional multi-sub-networks, to a scalable optical network - a network capable of being managed end-to-end (from the access network through the metro and core networks to another access network), across multiple ASs.

Generally speaking, faster and more robust delivery of content over the Internet can occur if the path is under the control of a single provider, implying an end-to-end managed network capable of routing packets optimally. So, by controlling as much of the network as possible and/or minimizing the number of "different hands" that "touch" the packets, service providers can minimize delay and avoid congestion points. To implement this vision, one needs to develop a scalable optical network - a network capable of being managed end-to-end, from the access network through the metro and core networks to another access network, across multiple Optical Domains (ODs).

The potential of implementing such a vision is made possible only if the cloud of traffic-bearing edge routers attached to OD_1 can see other clouds of routers attached to OD_2, OD_3, ..., OD_n. Practically, this is only possible if the edge routers attached to a given OD are not burdened by the significant amount of state and control information that they have to exchange. This is precisely the potential that our proposed interconnection model offers; that is to shift most of the intelligence and burden from the IP layer (traffic-bearing edge routers) to the optical layer (IP/MPLS-aware non-traffic bearing OXC controllers). Rather than burdening their database with the significant optical network semantic that they are attached to, the edge routers' databases can now be used to store some limited information about those clouds of edge routers that are attached to different ODs.

However, the proposed model provides the means to achieve end-to-end connectivity across multiple ODs that belong to different administrative domains. Provisioning an end-to-end optical path across multiple ODs involves the establishment of path segments

in each OD sequentially. To illustrate the concept of this architecture, we use the arbitrary network topology shown in Figure 6.4. Here, we have 3 optical domains that belong to different administrative domains (e.g. Sprint, AT&T and UUNet). Each OD consists of multiple OXCs interconnected via WDM links in a general mesh topology. The intelligence of the OXCs lies in their controllers that are capable of performing routing and wavelength assignment (RWA), as well as maintaining a database with already setup lightpaths (intra-domain logical connectivity). Attached to the ODs are high performance backbone IP routers from and to which traffic requests are generated and terminated respectively. Routers interconnecting multiple optical cores are defined as Gateway Routers (GRs). For example, in Figure 6.4, routers R_4, R_5, R_6, and R_7 are GRs. Two steps are required in order to achieve end-to-end connectivity across these ODs, namely, the *initialization phase*, and the *provisioning phase*.

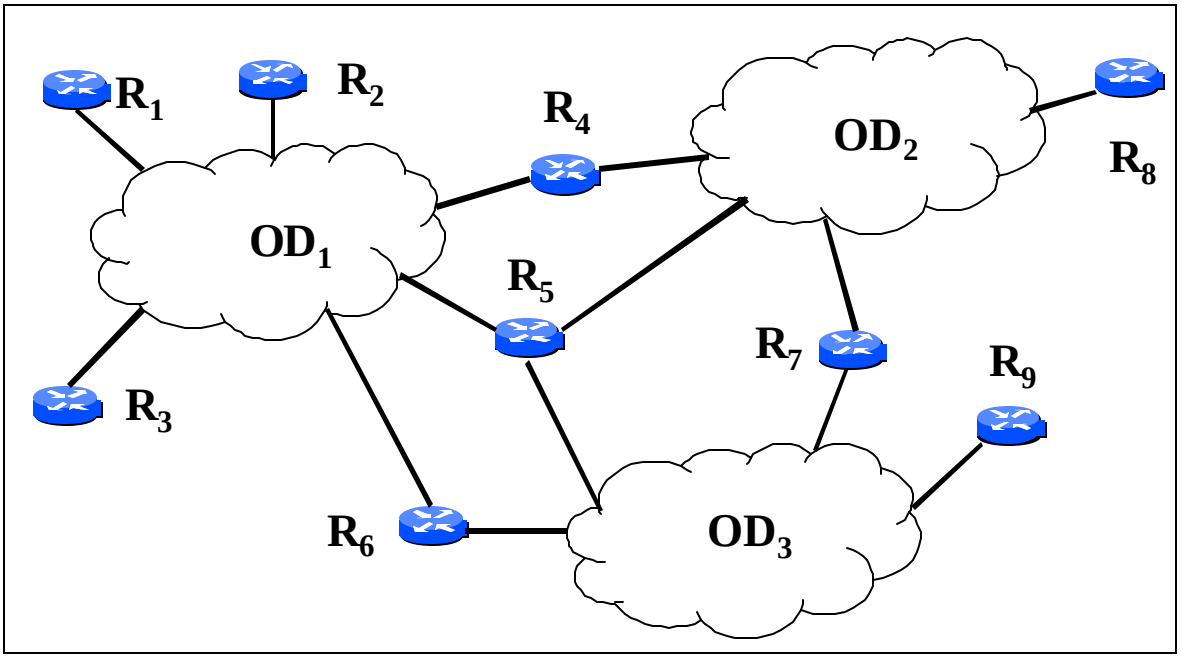

Figure 6.4: Arbitrary Multi-Domain Topology

6.3.1 Initialization Phase

During the initialization phase, the OXCs disseminate information about the presence of edge routers attached to them. It is thus possible for routers to identify reachable routers attached to the same OD through an intra-domain connectivity table. In this case, each edge router runs a limited reachability protocol with the corresponding edge OXC and obtains the address of every other edge router belonging to the same OD. Using this information, an initial set of IP routing adjacencies are established between edge routers. The edge routers then run an IP routing protocol amongst themselves to determine all the IP destinations reachable over the OD that they are attached to.

It is important to note that GRs are "seen" by all ODs they are connected to. In Figure 6.4, R_4 is "seen" by both OD_1 and OD_2, and R_5 is "seen" by all three ODs. The next step in the initialization phase is for the GRs to advertise their intra-domain connectivity tables from one OD to all other ODs they are connected to, declaring themselves as the gateway points. It is possible then to define a Global Topology (GT), which considers the ODs as black-box switches interconnecting all routers. With respect to Figure 6.4, after the initialization phase, the routers keep the following global topology database:

ODs are servicing …:	***GRs connecting …:***
OD_1: R_1, R_2, R_3, R_4, R_5, R_6.	**R_4: OD_1, OD_2.**
OD_2: R_4, R_5, R_7, R_8.	**R_5: OD_1, OD_2, OD_3.**
OD_3: R_5, R_6, R_7, R_9.	**R_6: OD_1, OD_3.**
	R_7: OD_2, OD_3.

Table 1: Global Topology Database for Example in Figure 6.4

6.3.2 Provisioning Phase

When a request arrives at a router (the request may originate at this router, or at another router within the global topology; in the latter case, the router is a GR), the router checks whether the destination belongs to its own OD:

- If yes, then it requests service from its OD via a defined UNI. The OD will then provision the request by setting up a lightpath, using already established lightpath(s), or by combining the two. In the case that no optical resources can be allocated, the call is blocked.
- If no, then the router consults its global topology table and finds out all gateways that can connect it to the destination OD, and simultaneously requests (using different instances of signaling) service to them from its OD. These gateway routers then acts as the source for the path and the above steps are repeated.

All ODs traversed have to return additive costs associated with the paths they computed (as opposed to returning the paths). Note that information about the so-far traversed ODs is kept in the signaling message, so that an OD can only be traversed once. When an instance of the signaling reaches the destination router, the latter sends it back to the source router (using the instance-specific route the signaling information carries). The source router, then, after collecting all the instances of its signaling, chooses the lowest-cost path, and initiates a reservation protocol to reserve the resources along the global path.

6.3.3 An Illustrative Example

If R_1 gets a request to R_9, it will send three parallel signaling instances requesting service from OD_1 to R_4, R_5 and R_6, respectively. OD_1 will decide the best routes to these destinations and will propagate the three signaling instances (note that whenever an OD performs route calculations from a source to a destination, it attaches the path-cost and its ID to the signaling information) to them:

- When R_4 receives the signaling information it will request paths to R_5 and R_7 from OD_2, which will then request service to the destination from OD_3.
- When R_5 receives the signaling information it will request a path to the destination from OD_3.
- When R_6 receives the signaling information it will request a path to the destination from OD_3.

In this specific example, a total of four possible paths will be examined:

R_1 - R_4 - R_5 - R_9,
R_1 - R_4 - R_7 - R_9,
R_1 - R_5 - R_9,
R_1- R_6 - R_9.

Global logical links in the paths above (i.e. from R_x to R_y) will have additive costs associated with them (which the corresponding OD that created them assigned). The least-cost path will then be chosen by the source router R_1, which will then use a reservation protocol to reserve the resources.

Note that the main strength and novelty of this architecture lies in its capability of provisioning end-to-end full optical channel capacity as well as sub-lambda connection requests across multiple ODs. We will develop, simulate and test several fully distributed GMPLS-based signaling protocols and algorithms that can support both full lambda and sub-lambda connection requests. One critical issue that we will thoroughly investigate is the viability of routing and signaling scalability across the multiple ODs, to determine the maximum number of ODs that can be traversed.

6.4 End-to-End Native Ethernet Transport

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

investigations on both the first mile (i.e. how to get the customer native Ethernet traffic on to the access network) and the core (how to interface fast GigE with the WDM-based OTN).

6.4.1 Ethernet in the First Mile

6.4.1.1 Ethernet Passive Optical Network (EPON)

Considerable amount of work has recently been on the issue of the first mile optical communications and the ways fiber can reach the end user premises in an affordable way. Ethernet-based Passive Optical Network (EPON) technology is emerging as a viable choice for the next-generation broadband access network [12-21]. A PON is a point-to-multipoint fiber optical network with no active elements in the signal's path. It consists of a single, shared optical fiber connecting a service provider's central office (head end) to a passive star coupler (S/C), which is located near residential customers. The S/C is intentionally positioned a substantial distance away from the central office, but close enough to the customers in order to save fiber. Customers receive a dedicated short optical fiber (that connects them to the S/C), but share the long distribution trunk fiber. All transmissions in a PON are performed between an Optical Line Terminal (OLT) and Optical Network Units (ONUs). Traffic from an OLT to an ONU is called 'downstream' (point-to-multipoint), and traffic from an ONU to the OLT is called 'upstream' (multipoint-to-point). Two wavelengths are used: typically 1310 nm (λ_{up}) for the upstream transmission and 1550 nm (λ_{d}) for the downstream transmission.

In the downstream direction, an EPON operates as a broadcast and select network. The OLT has the entire bandwidth of the channel to broadcast standard formatted 802.3 Ethernet frames to all ONUs. Each ONU extracts those packets that contain the ONUs unique Media Access Control (MAC) address. In the upstream direction, multiple ONUs share the transmission channel. Thus, the ONUs need to employ some arbitration mechanism to avoid collisions. In general, the OLT arbitrates the upstream transmissions by allocating an appropriate timeslot, or Transmission Window (TW) to each ONU. An ONU is only allowed to transmit during the TW allocated to it by the OLT. Within each cycle, in order to inform the OLT about its bandwidth requirements, ONUs use REPORT Messages that are also transmitted along with the data in the TW. Upon receiving a REPORT, the OLT passes the message to a Dynamic Bandwidth Allocation (DBA) module responsible for bandwidth allocation decision. The OLT assigns the TWs via GATE messages.

Several bandwidth allocation schemes have recently been reported in the literature ranging from a static allocation to a dynamically adapting scheme based on instantaneous queue size in every ONU [16-19]. The simplest is the static TDMA scheme in which every ONU gets a fixed timeslot [16]. While this scheme is very simple, it results in inefficient upstream channel utilization since statistical multiplexing between the ONUs is not possible. A DBA scheme called Interleaved Polling with Adaptive Cycle Time (IPACT) based on Grant and Request messages has been presented in [21]. This scheme uses an interleaved polling approach where the next ONU is polled before the transmission from the previous one has arrived. This scheme provides statistical multiplexing for ONUs and results in efficient upstream channel utilization.

6.4.1.2 A Decentralized EPON Approach

To date, the mainstream of these EPON bandwidth allocation schemes as well as the new IEEE 802.3ah EFM Task Force specifications [13] have been centralized–relying on a component in the central office (OLT) to provision upstream traffic. Hence, the OLT is the only device that can arbitrate time-division access to the shared channel. Since the OLT has global knowledge about the state of the entire network, this is a centralized control plane in which the OLT has a centralized intelligence. One of the major problems associated with a centralized architecture is the "single-point of failure problem"; that is the failure of the OLT will bring down the whole access network. To this end, we have proposed a decentralized EPON architecture and in the process prove that, in addition to the added flexibility and reliability, the performance of the proposed decentralized EPON architecture and the associated bandwidth allocation algorithms are at least as efficient as their centralized counterparts [21-22].

Figure 6.5 shows the general architecture of this approach. As can be seen, a portion of the optical signal power transmitted by an upstream transmitter (λ_{up}) toward the OLT will be redirected back and broadcasted to all ONUs. This can be achieved by connecting two ports of a 3 x N Star Coupler (S/C) with each other through an optical isolator as shown in Figure 6.5 [21]. Note that in addition to the conventional transceiver maintained at each ONU (a λ_{up} transmitter and a λ_{down} receiver), this approach requires an extra receiver tuned at λ_{up}. A baseband direct detection circuit is needed to detect the redirected control channel (λ_{up}) in order to recover the control update information.

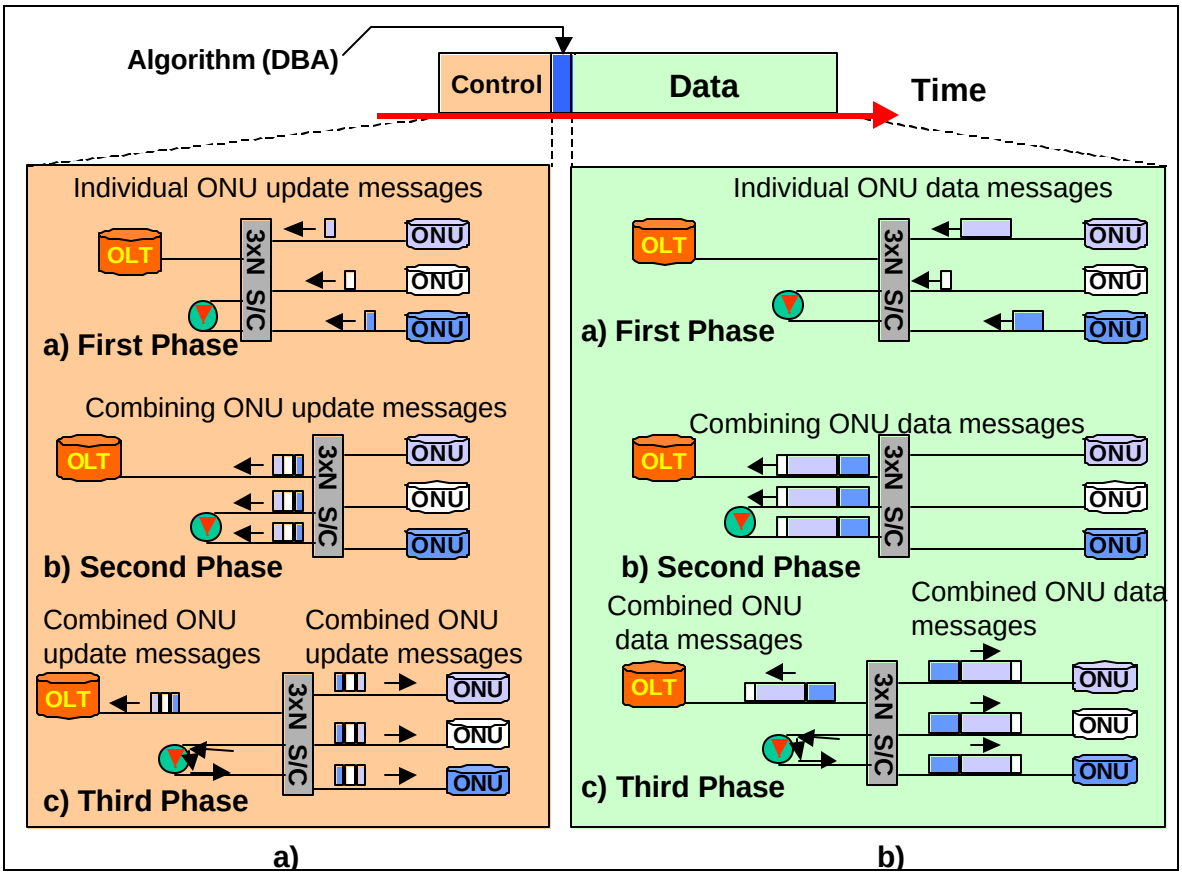

Figure 6.5: (a) Cycle updating process, (b) Transmission process.

This architecture assumes a cycle-based upstream link; a cycle is defined as the time that elapses between two executions of the scheduling algorithm. The cycle size can either have fixed, or variable length (confined within a certain upper bound) to accommodate the dynamic upstream traffic conditions. The cycle is divided into three periods; a static update period (control plane), a fixed waiting period (processing control messages and running the algorithm) and a dynamic transmission period (data plane).

The proposed cycle, along with the details of how the control plane performs the updating process is shown in Figure 6.5 (a) in three phases. Each ONU transmits its

update control message in its own assigned fixed time update slot (first phase). These messages are then combined at the S/C and a multiplexed update message is created (second phase). In the third phase, a fraction of the multiplexed control signal is transmitted through the first output port of the S/C and propagates to the OLT (which could discard it, make use of it as a synchronization message, and/or process the control information). Another fraction of the multiplexed control signal is redirected back and broadcasted to all ONUs (through the isolator). A baseband direct detection circuit located at each ONU is then used to detect the redirected control channel (λ_{update}). The detected signal is then processed in order to recover the control data information belonging to each of the other (N-1) ONUs. Since there are only two operating communication wavelengths (λ_{up} and λ_{d}), signaling and upstream transmission take place on the same communication channel (λup) and the periods will appear sequentially as on the top of Figure 6.5.

6.4.1.2.1 The First Period (Control Plane):

The update period is divided into N equal fixed time slots where N is the number of the ONU stations in the network. The update period is used for the ONUs to communicate their status and to exchange signaling and control message information with one another. Each ONU uses its own fixed time slot within the update period to transmit its control message. For simplicity, and to avoid collisions, the assignment of these N timeslots follows a fixed TDMA assignment since control messages are fixed in size. Note that the control slots in the proposed distributed scheme are all transmitted sequentially in one period (update period). This in contrast to the centralized schemes reported above [14-

21], where the control slot (REPORT Message) of each ONU is transmitted along with the data in the TW allocated to it by the OLT. All control update messages are transmitted as Ethernet frames. Because the signaling information is segregated from the upstream traffic, signaling information can be timelier and complete thus increasing the efficiency of the Dynamic Bandwidth Allocation algorithm. These enhanced DBA algorithms would have the ability to support better QoS characteristics because transmission of the signaling information is not constrained by the shared data/control upstream channel associated with the centralized schemes.

6.4.1.2.2 The Second Period (Algorithm Execution):

The second period of fixed length, is a waiting period (no upstream transmissions are allowed during this period) and is used for allowing the ONU's to process the information gathered from the multiplexed control message. Each ONU maintains a table with information about the state of the queues at each other ONU. This information is updated each cycle whenever the ONU receives a new multiplexed control message from all other ONUs. The DBA algorithm module uses the table maintained at each ONU. Note that instances of the same DBA algorithm are executed simultaneously and independently at each ONU. An execution of the algorithm yields a unique set of ONU assignments (wi) identically produced in each ONU (wi is the amount of bytes that an ONU is allowed to transmit in its TW). In other words, the algorithm should not incorporate any assumptions or randomness to handle exceptions. This is because several instances of it will run locally and independently at each ONU.

6.4.1.2.3 The Third Period (Data Plane):

The third period or (transmission period) is essentially a giant slot used for actual upstream data transmission. During the transmission period, the ONU's follow exactly the allocation scheme the algorithm produced (i.e. their transmissions start at specific times and last for specific bytes) as shown in Figure 6.5 (b). Note that the order of ONU's transmission may be different in each cycle and need not be fixed; but rather is a function of the ONU's traffic demand. This is a major advantage compared to the fixed transmission order proposed in [21].

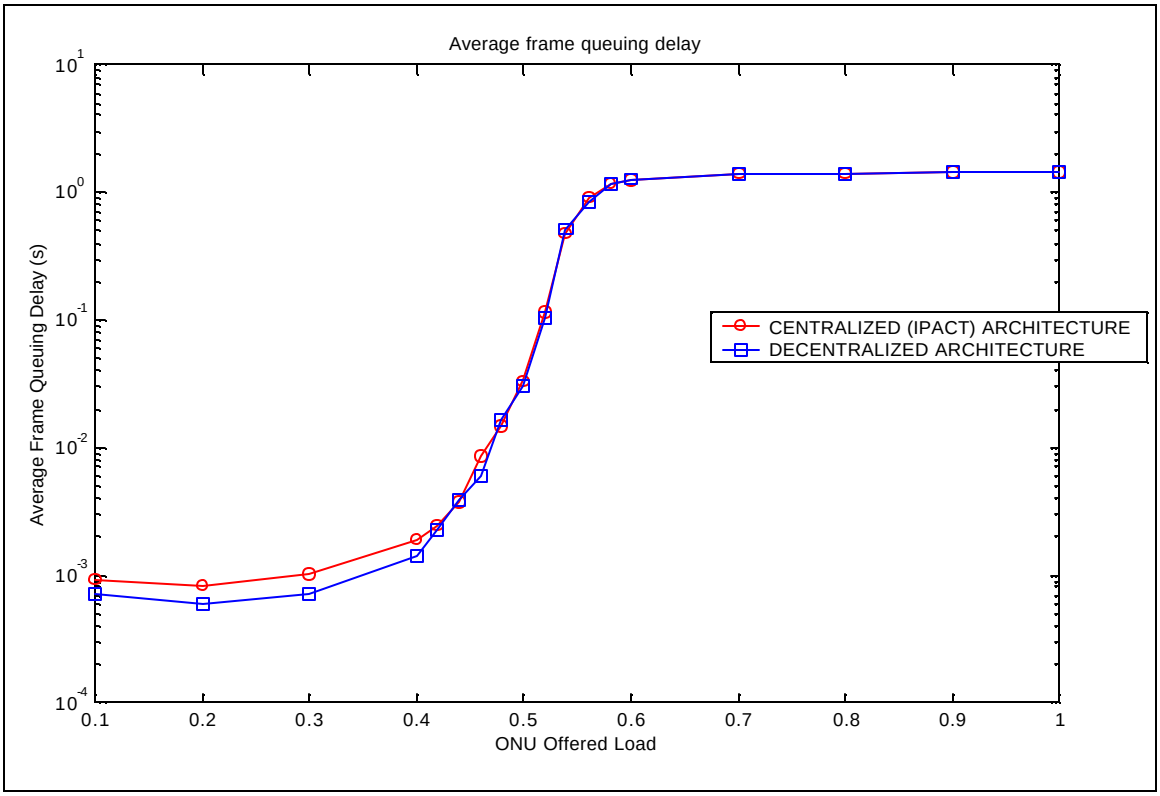

Figure 6.6: Average Frame Queuing Delay for Centralized and Decentralized architectures

6.4.1.3 Decentralized Performance Evaluation

The traffic model used here is the same as that reported in [21] where each ONU is modeled to be fed by a number of ON/OFF sources, each with a Pareto distribution governing the lengths of the ON/OFF periods, to capture the self-similar nature of Ethernet traffic. To compare the performance results of the proposed decentralized model with that of the centralized scheme (IPACT) of [21], we use the same system parameters used therein; a system with 16 ONUs, access link data rate from users to an ONU of 100 Mb/s, and a 1 Gb/s upstream link data rate (from an ONU to the OLT). Several bandwidth allocation algorithms were studied in [20-21], namely: fixed, limited, gated, constant credit, and linear credit. Amongst these algorithms, the limited (where the OLT grants the requested number of bytes, but no more than a given predetermined maximum), was shown to exhibits the best performance. Due to the space limitations, we use the simple limited DBA algorithm used in [21] for comparing our distributed architecture versus that of the centralized scheme reported therein.

Figure 6.6 presents the mean frame queuing delay, for both the centralized and distributed architectures using the Limited DBA algorithm, as a function of an ONU's offered load. In the case of the proposed decentralized approach, the order of the transmitting ONUs in a given cycle is not fixed (as in IPACT), but rather ordered based on the allocated TW determined by the DBA algorithm (the highest allocation transmits first; ties are broken by the ONU ID). From the results, it is observed that the decentralized approach improves IPACT in terms of the average frame delay at low loads. This is because by interchanging the order of transmissions, a given ONU's update

message is closer in time to its corresponding transmission. Thus, a more current depiction of its buffer status is governing the transmission. As the load increases more ONU's request more than the maximum allowed window, and thus more get the same allocation (maximum window). This, in turn, makes the advantage of the interchanged order of transmission to vanish.

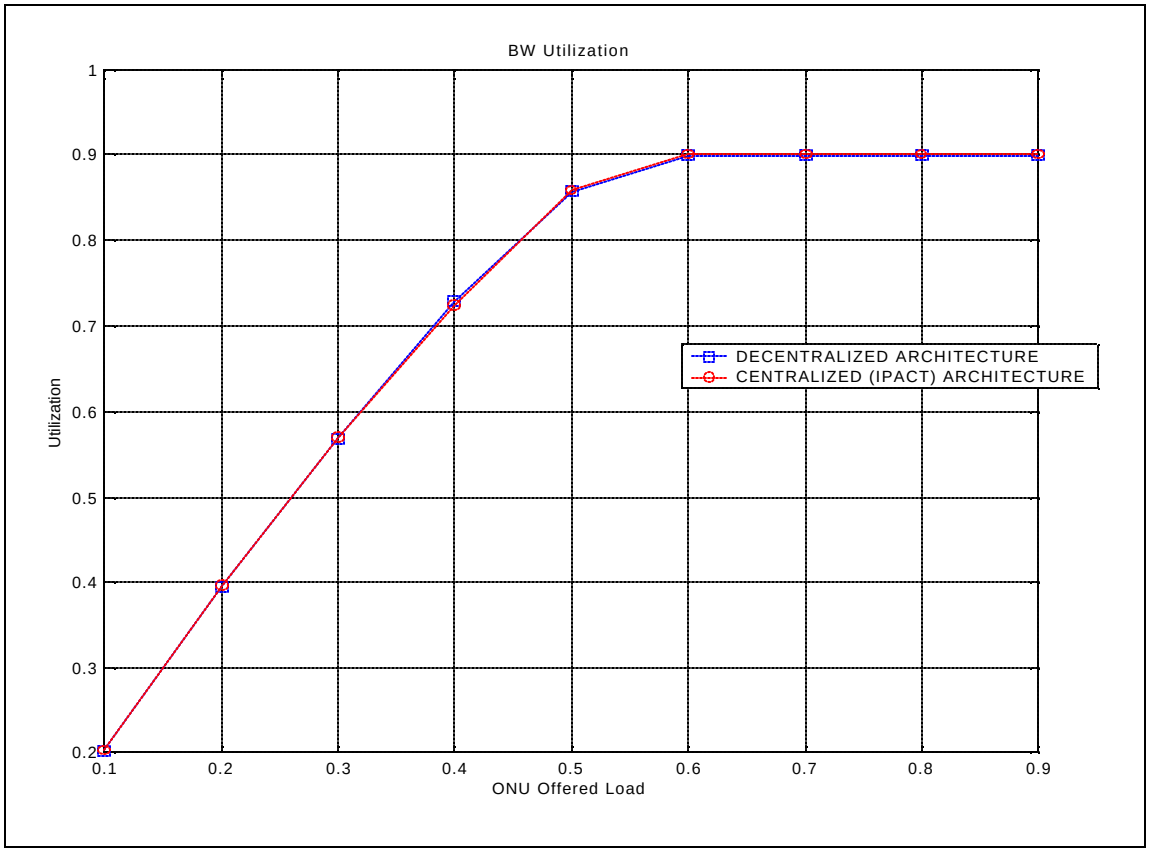

Figure 6.7: Bandwidth Utilization for Centralized and Decentralized architectures

Figure 6.7 shows the channel bandwidth utilization for both the centralized and distributed architectures using the Limited DBA algorithm, as a function of an ONU's offered load. As can be seen from the Figure, the performances of the two architectures are almost identical, with the centralized approach exhibiting a slight advantage (less than 1%).

Finally, it is important to emphasize that, in general, distributed architecture-based DBA algorithms (future work) would outperform those of the centralized architecture-based DBA algorithms reported in [20-21]. This is because a distributed DBA algorithm takes into account all other ONU requests when allocating a TW to a given ONU. This in contrast to the centralized architecture reported in [21], where all the proposed DBA algorithms take into account only that particular individual ONU request when allocating a TW to it.

6.4.2 GigE-over-WDM

6.4.2.1 Motivation

Scaling metro Ethernet networks into a global multi-services infrastructure is a direct implication of the proposed model of this work. Specifically, a truly native end-to-end layer-2 MAC frame-based Optical Ethernet infrastructure seamlessly stretching from enterprise LAN to Metro to Global. By combining the simplicity and cost effectiveness of Ethernet technology with the ultimate intelligence of WDM-based optical transport layer, Optical Ethernet (direct Ethernet-over-WDM) could evolve as a next generation networking paradigm providing a seamless global transport infrastructure for end-to-end transmission of native Ethernet frames.

Unlike today's notion of supporting "IP directly over WDM" (IP/MPLS-over-WDM interconnection models), which is little more than cleverly disguised marketing; "IP-over-

WDM" for example, is almost invariably IP packets mapped into SONET/SDH, coupled with SONET/SDH-based point-to-point DWDM systems [3-5]. The proposed "Ethernet-over-WDM" model is truly a two-layer model where native Ethernet frames are mapped directly over WDM. It offers advantages over existing Layer-2 and MPLS solutions in that it divorces the Ethernet from legacy transport mechanisms like SONET/SDH and other layer-2 protocols. The key for realizing such a very high-risk initiative rests on devising innovative networking solutions to replace the legacy layer-3 switching (routing) and hierarchal IP addressing scheme with layer-2 switching and flat (non-hierarchal) MAC/VLAN-based addressing scheme?

To implement such an ambitious vision of a global multi-services Ethernet infrastructure, several key critical issues need to be thoroughly examined and addressed including:

1. How to totally eliminate the reliance on Spanning Tree (ST) routing and redundancy functionality?
2. How to reliably transport native Ethernet frames that have no overhead capability to perform network Operations, Administrations, Management and Protection (OAM&P) across the WAN?
3. How to integrate layer-2 (L-2) Ethernet control plane functionality with that of the optical transport layer (layer-1)?
4. How to devise a novel global layer-2 MAC and/or VLAN ID-address structure and space that is unique, hierarchal, and scalable with a source and destination addresses.

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

- Conventional Ethernet MAC frames and/or jumbo Ethernet frames must be transported natively (translation into some other protocol is not allowed) end to

end from the access network through the metro and core networks to another access network.

- Only pure layer-2, switching at the packet/frame granularity, is allowed throughout the entire network including access, MAN and WAN.
- Unlike layer-2 MPLS, VPNs (point-to-point and multipoint Virtual Private LAN Services (VPLS)), where a full mesh of static label switched paths (LSPs) must be set up between all L2 VPN sites, the proposed optical Ethernet is dynamically reconfigurable network that supports real-time additions/deletions of all customer connections (EVCs). It can also support a fully automated optical networking service (layer 1) at any bandwidth granularity.
- Supports an IP/GMPLS-based unified control plane that offers a tighter integration between layer-1 (optical transport layer) and layer-2 (Ethernet layer), leading to the collapse of the two layers into a single integrated layer managed and traffic engineered in a unified manner. The unified control plane supports real-time provisioning and restoration of both full lambda and EVCs by running a single instance of an integrated routing and signaling protocols (use of ST, RST, and MST routing are totally eliminated).
- Native Ethernet frames are routed across the MAN/WAN using only layer-2 addressing scheme (MAC and/or VLAN ID).

6.4.2.2 Implementation Strategy

It is important to emphasize from the outset that there are strong analogies between the IP-over-WDM interconnection models and the proposed Ethernet-over-WDM model. Anyone who has followed the development of IP-over-WDM interconnection models (the overlay and peer models) throughout 1990s can easily observe that most of the initial problems encountered in the development process were mainly due the optical network

and IP gap caused by the optical networking and the IP/MPLS research communities. It took the industry, the standards bodies, and the two research communities nearly ten years of extensive continuous hard work and collaboration to narrow this gap. Likewise, we strongly believe that the main problem that will hinder the viability of implementing the vision of a global optical Ethernet infrastructure is the wide gap that exists between the optical networking research communities and the Ethernet communities. It is our expectations that the proposed research program would be an important starting point to bring the two communities together, leading eventually to bridging this gap and the realization of the proposed vision.

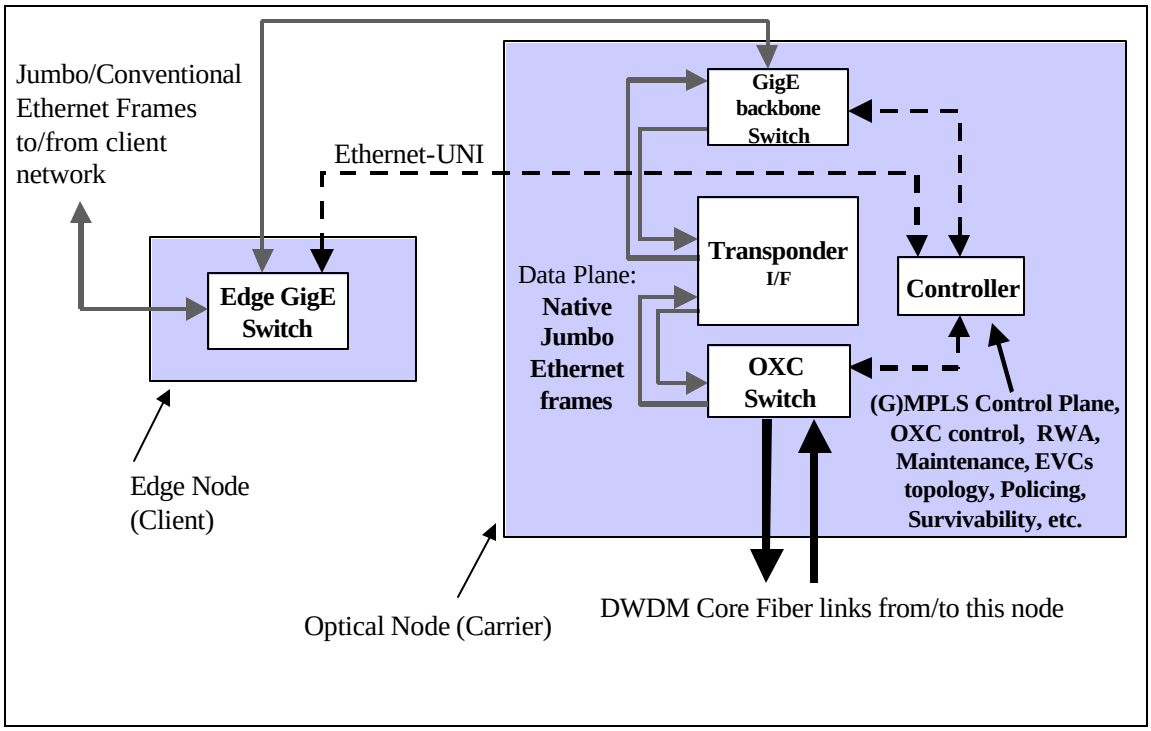

Figure 6.8: Modification to the proposed node architecture for implementation of GigE/WDM architecture

Now there is an opportunity to reapply a lot of this technology, suitably modified to Ethernet technology by taking full advantage of the knowledge and developments gained during the past ten-year course of developing the IP-over-WDM interconnection models.

The key to a successful strategy rests on taking the best features from both the overlay and peer models while avoiding their limitations. Specifically, using the proposed model but replacing the EXC with a core (backbone) GigE switch as shown in Figure 6.8.

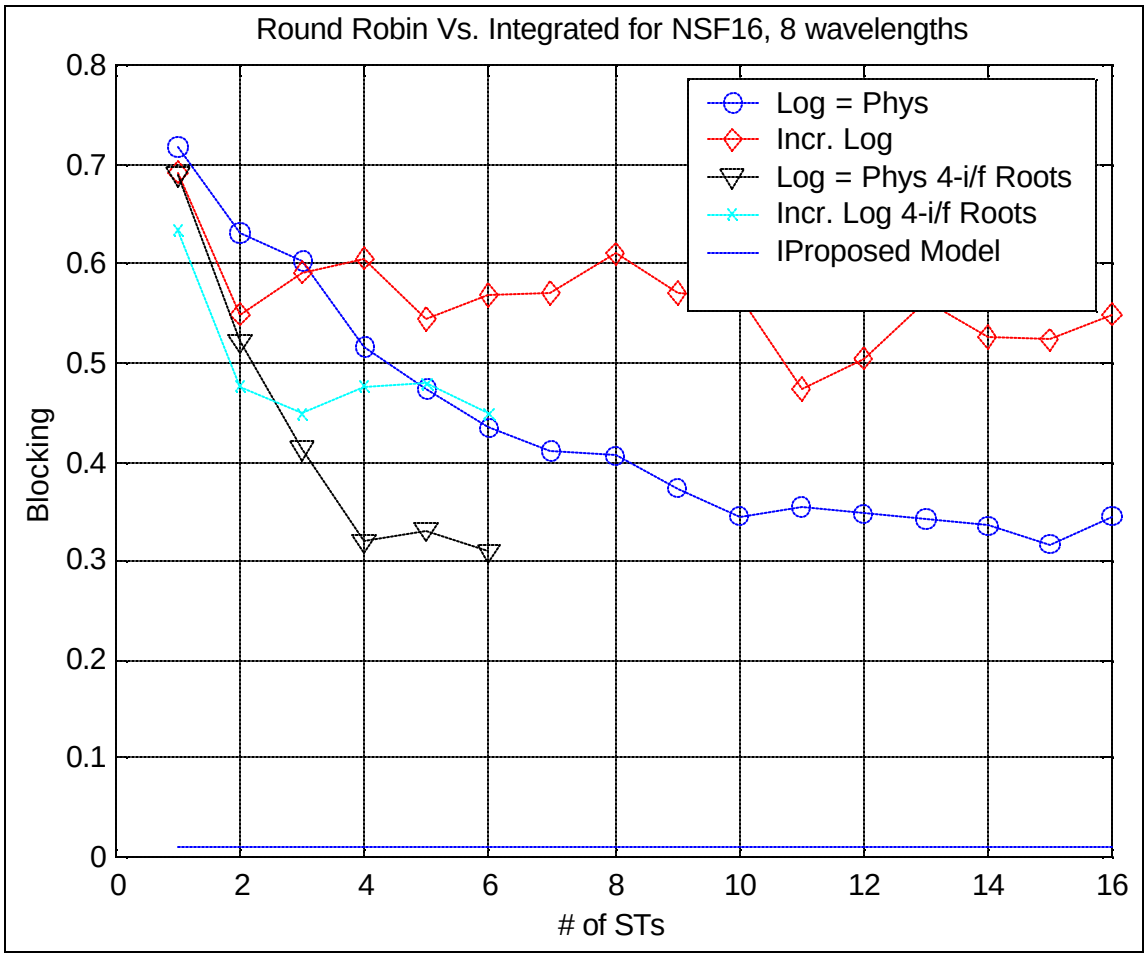

Figure 6.9: Comparing current Spanning Tree with the Integrated Routing fro GigE/WDM Networks

Initial results show the impediments of Single Spanning Tree (SST), or Multiple Spanning Trees (MSTs), compared to the case that both logical and physical topologies are maintained by the optical core. Figure shows 6.9 how bad blocking results are even for MSTs when compared to the case where the optical layer performs all routing decisions based on its updated view of both layers' resources. The results investigate

increasing the number of STs in the case that the Logical Topology is equal to the Physical Topology (i.e. Lightpaths span exactly one physical link) and in the case that the traffic builds the Logical Topology incrementally. The results that stop at 6 STs are the ones that only choose ST Roots nodes that have rich connectivity (for NSF16, these are the 4-interface nodes). The admission of the EVC requests follows a Round Robin fashion; i.e. calls are pre assigned a ST (perhaps based on their IDs) and they are tried only on that ST. The straight line shows the blocking of the same traffic on the same network but using our proposed model where integrated routing can be applied.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

The purpose of this thesis has been to propose and investigate the potentials of a novel client-over-carrier interconnection model architecture where all of the OAM&P functionalities are shifted down into the carrier layer. These functionalities include building the logical topology and maintaining and updating forwarding tables that keep track of established sub-lambda connections between edge client devices in addition to the conventional, physical layer database maintenance. Thus, under this model both the logical and physical topologies belong to a single administrative domain, leading to the creation of a unified control plane that can efficiently support both full-lambda and sub-lambda routing, signaling and survivability functionalities in the optical domain. The

issue of integrated dynamic routing algorithms was investigated and a novel unified signaling model appropriate for such a model was devised and evaluated via thorough simulation. The proposed optical node architecture is composed of three components. A backbone IP/MPLS router used for Electronic Cross-Connections (EXC), an OXC switch, and an IP/MPLS-aware, non-traffic bearing OXC controller module. The backbone IP/MPLS router belongs to the optical layer and is a high-speed router capable of statistically multiplexing data streams up to the capacity that can be supported by the OXC (several full wavelength channel speeds).

The presented implications of the model include, not only the *real* integrated traffic grooming solution and the selective, per-call, restoration, but also the realizing novel network architecture of overlaying GigE directly over the Optical WDM core for native Ethernet transport between end-users. The latter is made possible through simple modifications due to this models' flexibility.

However, there are new avenues opened up by the model and this thesis has not investigated. Below we summarize some of them.

7.2 Future Work

7.2.1 Signaling for Selective Restoration

Although this work presented a novel signaling for the integration of both full- and sub-wavelength connection provisioning, the problem of signaling in the presence of a fault

needs deeper investigation. For instance, current restoration signaling approaches should be thoroughly examined and, whenever possible, be modified and used to this aim.

Another issue that requires deeper investigation is quality of service as it could be provided under the proposed unified control plane. If the optical layer knows both-layer connectivity matrices, how could the carrier network administrators make advantageous decisions that use of this information in order to provide clients with a long array of differentiated service provisioning?

7.2.2 Native Ethernet Transport Issues

The solution for true seamless integration of telecom data networks could very well lay in the Ethernet protocol. This is made possible by recent advances in fast (Gigabit) Ethernet technologies and the introduction of support for Virtual LANs (VLANs). As this work argued, the novel architectural model presented here could very well be the answer to directly overlaying the GigE with the optical transport network. However, many of the problems remain unsolved when it comes to such a solution realization. The list below presents some of them but is by no means exhaustive:

1. if the assumption that OAM&P functionalities are settled when it comes to the edge points, where the GiGE and WDM layers interface, more investigation should be performed to how these will be performed elsewhere, given that GigE is not supporting them.

2. end-to-end differentiated quality of service that spans from the source-user across different network administrative domains to the destination-user.
3. is there such a thing as a survivable EPON if the star/coupler architecture is adopted? Could wireless control-plane communication between ONUs be a solution that further enhances performance and/or survivability?
4. the addressing scheme of Ethernet is based on the learning process of VLANs and layer-2 MAC addresses. There should be work performed on how the destination addresses will be reachable from within the optical core and how these will be registered in a scalable manner.
5. ?

7.2.3 New Optical Telecommunication Avenues

It is eminent that my future work will not be limited to the above mentioned issues but will instead by broadened to even completely new networking aspects such as:

- **The security problem in all-optical networks using quantum encryption:**

  The usage of such networks is rapidly growing in fields such as e-commerce and banking, which makes the issue of secure networks extremely timely and critical. This problem is especially important in WDM-based networks, where the high data bit rates purport that even a short attacks will cause the compromise of a large amount of data.

- **Wireless Optical Networks:**

An interesting research topic which gains ground nowadays is wireless optics whereby optical network speeds are reaching end-users via photodiodes installed at high-rise rooftops.

BIBLIOGRAPHY

CHAPTER 1: INTRODUCTION

[1] K. Zhu and B. Mukherjee, "A Review of Traffic Grooming in WDM Optical Networks: Architectures and Challenges", SPIE Optical Networks Magazine, Vol. 4, March - April 2003, pp. 55-64.

[2] K. Zhu, B. Mukherjee, "Traffic Grooming in an Optical WDM Mesh Network", IEEE ICC'01, 11-14 June 2001, Volume 3, pp.721 - 725.

[3] K. Zhu, H. Zhu, and B. Mukherjee, "Traffic Engineering in Multigranularity Heterogeneous Optical WDM Mesh Networks through Dynamic Traffic Grooming", IEEE Network Magazine, PP. 8-16, March/April 2003.

[4] K. Zhu and B. Mukherjee, "Traffic Grooming in an Optical WDM Mesh Network", IEEE Journal on Selected Areas in Communications, vol. 20, no. 1, January 2002, pp. 122-133.

[5] C. Assi, Y. Ye, A. Shami, S. Dixit, and M. Ali, "Integrated Routing Algorithms for Provisioning 'Sub-Wavelength' Connections in IP-Over-WDM Networks", Journal of Photonic Network Communications, vol. 4, 2002, pp. 377-390.

[6] Y. Sun, J. Gu, "Traffic Grooming in All-Optical Networks", IEEE ICC'01, June 2001.

[7] C. Xin, C. Qiao, and S. Dixit, "Traffic Grooming in Mesh WDM Optical Networks—Performance analysis", IEEE Journal on Selected Areas in Communications, vol. 22, no. 9, November 2004, pp. 1658-1669.

[8] E. Modiano and P. J. Lin, "Traffic Grooming in WDM Networks", IEEE Communications Magazine, vol. 39, no. 7, pp. 124-129, July 2001.

[9] H. Zhu, H. Zang, K. Zhu, and B. Mukherjee, "A Novel, Generic Graph Model for Traffic Grooming in Heterogeneous WDM Mesh Networks", IEEE/ACM Transactions on Networking, vol.11, no. 2, April 2003, pp.285-299.

[10] C. Ou, K. Zhu, H. Zang, L. Sahasrabuddhe, and B. Mukherjee, "Traffic Grooming for Survivable WDM Networks--- Shared Protection", IEEE Journal on Selected Areas in Communications, vol. 21, no. 9, November 2003, pp. 1367-11383.

[11] D. Awduche et al, "Multiprotocol Lambda Switching", Internet Draft, work in progress, November 1999.

[12] “Generalized Multi-Protocol Label Switching (GMPLS) Architecture”, draft-ietf-ccamp-gmpls-architecture-03.txt, IETF Draft, August 2002.

[13] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[14] “Generalized MPLS, Signaling Functional Description”, draft-ietf-mpls-generalized-signaling-09.txt, IETF Draft, August 2002.

[15] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects”, IEEE Communications Magazine, pp. 94-102, September 2000.

[16] N. Ghani et al., “on IP-over-WDM Integration”, IEEE Communications magazine, March 2000.

[17] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[18] E. Mannie et al., “Generalized Multi-Protocol Label Switching (GMPLS) architecture”, IETF Internet draft, Mar. 2002.

[19] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, “Optical layer-based unified control plane for emerging IP/MPLS over WDM networking architecture”, in Proc. 29th European Conf. on Optical Communication, Rimini, Italy, September 2003, pp. 836-837.

[20] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, “A Novel IP-Over-Optical Network Interconnection Model for the Next-Generation Optical Internet”, in Proc. IEEE Globecom, San Francisco, December 2003, pp. 3984-3989.

CHAPTER 2: LESSONS LEARNED FROM IP-OVER-WDM INTERCONNECTION MODELS

[1] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects”, IEEE Communications Magazine, pp. 94-102, September 2000.

[2] M. A. Ali, A. Shami, C. Assi, Y. Ye, R Kurtz, “Architecture Options for Next-Generation Networking Paradigm: Is Optical Internet the Answer”, Journal of Photonic Network Communications, vol. 3, no. 1/2, Jan-Jun 2001.

[3] N. Ghani et al., “On IP-over-WDM Integration”, IEEE Communications magazine, March 2000.

[4] D. Awduche et al, “Multiprotocol Lambda Switching”, Internet Draft, work in progress, November 1999.

[5] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[6] E. Mannie et al., “Generalized Multi-Protocol Label Switching (GMPLS) architecture”, IETF Internet draft, Mar. 2002.

[7] D. Awduche et al, “Multiprotocol Lambda Switching”, Internet Draft, work in progress, November 1999.

[8] “Generalized Multi-Protocol Label Switching (GMPLS) Architecture”, draft-ietf-ccamp-gmpls-architecture-03.txt, IETF Draft, August 2002.

[9] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[10] “Generalized MPLS, Signaling Functional Description”, draft-ietf-mpls-generalized-signaling-09.txt, IETF Draft, August 2002.

[11] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects”, IEEE Communications Magazine, pp. 94-102, September 2000.

[12] N. Ghani et al., “on IP-over-WDM Integration”, IEEE Communications magazine, March 2000.

[13] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[14] E. Mannie et al., “Generalized Multi-Protocol Label Switching (GMPLS) architecture”, IETF Internet draft, Mar. 2002.

CHAPTER 3: THE PROPOSED NETWORK MODEL ARCHITECTURE

[1] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects”, IEEE Communications Magazine, pp. 94-102, September 2000.

[2] M.A. Ali, A. Shami, C. Assi, Y. Ye, R Kurtz, “Architecture Options for Next-Generation Networking Paradigm: Is Optical Internet the Answer”, Journal of Photonic Network Communications, vol. 3, no. 1/2, Jan-Jun 2001.

[3] N. Ghani et al., “on IP-over-WDM Integration”, IEEE Communications magazine, March 2000.

[4] D. Awduche et al, “Multiprotocol Lambda Switching”, Internet Draft, work in progress, November 1999.

[5] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[6] E. Mannie et al., “Generalized Multi-Protocol Label Switching (GMPLS) architecture”, IETF Internet draft, Mar. 2002.

[7] K. Zhu and B. Mukherjee, “Traffic Grooming in an Optical WDM Mesh Network”, IEEE Journal on Selected Areas in Communications, Special Issue on "WDM-Based Network Architectures", vol. 20, no. 1, pp. 122-133, Jan. 2002.

[8] E. Modiano and P. J. Lin, “Traffic Grooming in WDM Networks”, IEEE Communications Magazine, vol. 39, no. 7, pp. 124-129, July 2001.

[9] K. Zhu, H. Zhu, and B. Mukherjee, “Traffic Engineering in Multigranularity Heterogeneous Optical WDM Mesh Networks through Dynamic Traffic Grooming”, IEEE Network Magazine, PP. 8-16, March/April 2003.

[10] M. A. Ali, A. Hadjiantonis, Keren Bergman, and G. Ellinas, “Transportation & Switching of native Ethernet frames across MPLS/GMPLS Managed and Controlled Optical data networks“, (INVITED), Proceedings of the 17th IEEE/LEOS Annual meeting on Optical Networks and Systems, Puerto Rico, Nov 7-11 2004.

[11] A. Hadjiantonis, A. Khalil, G. Ellinas, and M. A. Ali, “A novel optical-layer based Resilience Scheme for the Next-Generation data networks”, (INVITED), Proceedings of the 16th IEEE/LEOS Annual meeting on Optical Networks and Systems, Arizona, Oct. 2003.

[12] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, “A Novel IP-Over-Optical Network Interconnection Model for the Next-Generation Optical Internet”, Proceeding of IEEE GLOBECOM, San Francisco, USA, Dec. 2003.

[13] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, “Optical layer-based unified control plane for emerging IP/MPLS over WDM networking architecture”, 29th European Conference on Optical Communication (ECOC), Rimini, Italy, pp 836-837, Sept. 2003.

CHAPTER 4: THE ROUTING MECHANISM

[1] C. Assi, A. Shami, R. Kurtz, and M. Ali, "Optical networking and real-time provisioning; An integrated vision for the next-generation Internet", IEEE Network Magazine, July/August 2001, Vol. 15 No. 4, PP. 36-45.

[2] M. Kodialam, and T. V. Lakshman, "Integrated dynamic IP and wavelength routing in IP over WDM networks", Proc. IEEE Infocom 2001, pp. 358-366, Anchorage, Alaska, April 2001.

[3] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, "A Novel IP-Over-Optical Network Interconnection Model for the Next-Generation Optical Internet", in Proc. IEEE Globecom, San Francisco, December 2003, pp. 3984-3989.

[4] C. Chen and S. Banerjee, "A new model for optimal routing and wavelength assignment in wavelength division multiplexed optical networks", IEEE INFOCOM '96, pp. 148–155, 1996.

[5] "IP over Optical Networks: A Framework", draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[6] E. Mannie et al., "Generalized Multi-Protocol Label Switching (GMPLS) architecture", IETF Internet draft, Mar. 2002.

[7] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, "Optical layer-based unified control plane for emerging IP/MPLS over WDM networking architecture", in Proc. 29th European Conf. on Optical Communication, Rimini, Italy, September 2003, pp. 836-837.

[8] C. Assi, Y. Ye, A. Shami, S. Dixit, and M. Ali, "Integrated Routing Algorithms for Provisioning "Sub-Wavelength" Connections in IP-Over-WDM Networks", Journal of Photonic Network Communications, vol. 4, 2002, pp. 377-390.

CHAPTER 5: AN INTEGRATED SIGNALING COMPONENT

[1] A. Khalil, A. Hadjiantonis, G. Ellinas, and M. A. Ali, "A Novel IP-Over-Optical Network Interconnection Model for the Next-Generation Optical Internet", in Proc. IEEE Globecom, San Francisco, December 2003, pp. 3984-3989.

[2] M. Kodialam, and T. V. Lakshman, "Integrated dynamic IP and wavelength routing in IP over WDM networks", Proc. IEEE Infocom 2001, pp. 358-366, Anchorage, Alaska, April 2001.

[3] C. Chen and S. Banerjee, “A new model for optimal routing and wavelength assignment in wavelength division multiplexed optical networks”, IEEE INFOCOM ’96, pp. 148–155, 1996.

[4] C. Assi, A. Shami, and M. A. Ali, “Optical networking and real-time provisioning; an integrated vision for the next-generation Internet”, IEEE Network Magazine, vol. 15, no. 4, pp. 36-45, July/August 2001.

[5] A. Shami, C. Assi, and M. Ali “Performance Evaluation of Two GMPLS-Based Distributed Control and Management Protocols for Dynamic Lightpath Provisioning in Future IP Networks”, Proc. IEEE ICC’02, NY, 2002.

[6] R. Ramaswami and A. Segall, “Distributed Network Control for Wavelength Routed Optical Networks”, Proceedings, IEEE INFOCOM ’96, San Francisco, CA, pp. 138-147, Mar. 1996.

[7] C. Qiao, and D. Xu, “Distributed Partial Information Management (DPIM) Schemes for Survivable Networks-Part I”, IEEE INFOCOM 2002, NY.

[8] M. Goyal, G. Li, and J. Yates, “Shared Mesh Restoration: A Simulation Study”, OFC’02, March 2002.

[9] M. Goyal, J. Yates, G. Li, and W. Feng, “Benefits of Restoration Signaling Message Aggregation”, Proc. OFC’03, March 2003.

[10] G. Li, J. Yates, R. Doverspike, and D. Wang, “Experiments in Fast Restoration Using GMPLS in Optical/Electronic Mesh Networks”, Proc. OFC’01, March 2001.

CHAPTER 6: PRACTICAL IMPLICATIONS OF THE PROPOSED MODEL

[1] A. Hadjiantonis, A. Khalil, , G. Ellinas, and M. A. Ali, “A novel optical-layer based Resilience Scheme for the Next-Generation data networks”, (INVITED), Proceedings of the 16th IEEE/LEOS Annual meeting on Optical Networks and Systems, Arizona, Oct. 2003.

[2] H. Zhu, H. Zang, K. Zhu, and B. Mukherjee, “A Novel, Generic Graph Model for Traffic Grooming in Heterogeneous WDM Mesh Networks”, IEEE/ACM Transactions on Networking, vol.11, no. 2, April 2003, pp.285-299.

[3] “IP over Optical Networks: A Framework”, draft-ietf-ipo-framework-03.txt, IETF Draft, January 2003.

[4] E. Mannie et al., “Generalized Multi-Protocol Label Switching (GMPLS) architecture”, IETF Internet draft, Mar. 2002.

[5] D. Awduche et al, “Multiprotocol Lambda Switching”, Internet Draft, work in progress, November 1999.

[6] “Generalized Multi-Protocol Label Switching (GMPLS) Architecture”, draft-ietf-ccamp-gmpls-architecture-03.txt, IETF Draft, August 2002.

[7] Jeff Camp, A. Shah, and R .T, Lee, “Internet Infrastructure Services”, Industry report, Morgan Stanley Dean Witter, June 2001.

[8] P. Ashwood-Smith et al, “Generalized MPLS, Signaling Functional Description”, Internet Draft, work in progress, November 2000.

[9] B. Rajagopalan et al, “IP over Optical Networks: Architecture Aspects”, IEEE Communication Magazine, PP.94-102, September 2000.

[10] M. Kodialam, and T. V. Lakshman, “Integrtaed dynamic IP and wavelength routing in IP over WDM networks”,IEEE INFOCOM 2001, PP 358-366.

[11] C. Xin, Y. Ye, S. Dixit, and C. Qiao, “An integrated lightpath provisioning approach in messh optical networks”, OFC 2002.

[12] Full Services Access Networks, http://www.fsanet.net/.

[13] IEEE 802.3 Ethernet in the First Mile Study Group,

http://www.ieee802.org/3/efm/public/index.html

[14] B. Lung, “PON Architecture ‘Future proofs’ FTTH”, J. Lightwave, vol.16, no.10, pp.104–107, Sept. 1999.

[15] D. Mynbaev, “From Core to Metro to Access Networks – The Need for Passive Optical Networks”.

[16] G. Kramer and G. Pesavento, “Ethernet Passive Optical Network (EPON): Building a Next Generation Optical Access Network”, IEEE Com. Mag., pp. 66-73, Feb. 2002.

[17] Alloptic, “Ethernet Passive Optical Networks”, The International engineering Consortium, http://www.iec.org.

[18] Kramer, et al. “Ethernet PON (EPON): Design and Analysis of an Optical Access Network”, Photonic Network Communications Journal, vol. 3, No. 3, July 2001.

[19] G. Kramer, B. Mukherjee, and G. Pesavento, “IPACT: A dynamic Protocol for an Ethernet PON (EPON)”, IEEE Comm. Mag., pp. 74-80, Feb. 2002.

[20] Chae Chang-Joon, Elaine Wong, Rodney S. Tucher, “Optical CSMA/CD Media Access Scheme for Ethernet Over Passive Optical Network”, IEEE Photonics Technology Letters, vol. 14, No.5, March 2002.

[21] A. Hadjiantonis, S. R. Sherif, G. Ellinas, C. Assi and M. A. Ali, “A Novel Decentralized Ethernet-Based PON Architecture”, Proceedings of IEEE ICC2004, Paris, France, 20-24 June 2004, paper ON06-8.

[22] S. R. Sherif, A. Hadjiantonis, G. Ellinas, C. Assi and M. A. Ali, “A novel decentralized Ethernet-Based PON access architecture for provisioning differentiated QoS”, Journal of Lightwave Technology, November 2003, Volume 22, pp. 2483- 2497.